Encyclopedia of

Health Economics

Anthony J. Culyer

ENCYCLOPEDIA OF
HEALTH ECONOMICS

How to go to your page

This eBook set contains 3 volumes.

The chapter numbers are contiguous between the first two volumes, but Volume 3 begins anew. To search for pages in Volume 3 use the example below:

To go to page 18 of Volume 3, type "Vol 3:18" in the "page #" box at the top of the screen and click "Go."

To go to page "306" of Volume 3, type "Vol 3: 306"… and so forth.

Please refer to the eTOC for further clarification.

ENCYCLOPEDIA OF
HEALTH ECONOMICS

EDITOR-IN-CHIEF
Anthony J Culyer
University of Toronto, Toronto, Canada
University of York, Heslington, York, UK

ELSEVIER

Amsterdam • Boston • Heildelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

First edition 2014

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the Library of Congress.

ISBN 978-0-12-375678-7

Printed and bound in the United States of America

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

Project Manager: Gemma Taft
Associate Project Manager: Joanne Williams

EDITORIAL BOARD

CONTRIBUTORS TO VOLUME 1

AK Acharya
OP Jindal Global University, Sonipat, India, and London School of Hygiene and Tropical Medicine, London, UK

D Almond
Columbia University and NBER, New York, NY, USA

R Ara
University of Sheffield, Sheffield, UK

MC Auld
University of Victoria, Victoria, BC, Canada

A Basu
University of Washington, Seattle, WA, USA

GJ van den Berg
University of Mannheim, Mannheim, Germany; IFAU Uppsala; VU University Amsterdam, and IZA

PM Bernet
Florida Atlantic University, Boca Raton, FL, USA

L Bojke
University of York, York, UK

J Brazier
University of Sheffield, Sheffield, UK

BW Bresnahan
University of Washington, Seattle, WA, USA

S Bryan
University of British Columbia, Vancouver, BC, Canada; Vancouver Coastal Health Research Institute, Vancouver, BC, Canada, and University of Aberdeen, Aberdeen, UK

K Carey
Boston University School of Public Health, Boston, MA, USA

C Carpenter
Vanderbilt University, Nashville, TN, USA

M Chalkley
University of York, Heslington, York, UK

P Chatterji
University at Albany and NBER, Albany, NY, USA

T Chen
Boston University, Boston, MA, USA

RA Cookson
University of York, York, UK

Z Cooper
Yale University, New Haven, CT, USA

JM Currie
Princeton University, Princeton, NJ, USA

D Cutler
Harvard University and NBER, Cambridge, MA, USA

PM Danzon
University of Pennsylvania, Philadelphia, PA, USA

DM Dave
Bentley University, Waltham, MA, USA

G David
University of Pennsylvania, Philadelphia, PA, USA

A Dor
George Washington University, Washington, DC, USA

B Dormont
PSL, Université Paris Dauphine, Paris, France

DM Dror
Micro Insurance Academy, New Delhi, India, and Erasmus University Rotterdam, Rotterdam, The Netherlands

MF Drummond
University of York, York, UK

A Ebenstein
Hebrew University of Jerusalem, Jerusalem, Israel

RP Ellis
Boston University, Boston, MA, USA

MA Espinoza
Pontificia Universidad Católica de Chile, Santiago, Chile, and Institute of Public Health of Chile, Santiago, Chile

E Fenwick
University of Glasgow, Glasgow, Scotland, UK

E Fichera
University of Manchester, Manchester, UK

LP Garrison Jr.
University of Washington, Seattle, WA, USA

D Gilleskie
University of North Carolina, Chapel Hill, NC, USA

J Glazer
Boston University, Boston, MA, USA, and Tel Aviv University, Tel Aviv, Israel

H Grabowski
Duke University, Durham, NC, USA

M Grignon
McMaster University, Hamilton, ON, Canada

G Gumus
Florida Atlantic University, Boca Raton, FL, USA

D Gyrd-Hansen
University of Southern Denmark, Odense, Denmark

M Haacker
London School of Hygiene and Tropical Medicine, London, England, UK

A Harmer
University of Edinburgh, Edinburgh, UK

DL Heymann
Centre on Global Health Security, Chatham House, UK, and Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, UK

B Hollingsworth
Lancaster University, Lancaster, UK

J Hsu
London School of Hygiene and Tropical Medicine, London, UK

S Jha
University of Pennsylvania, Philadelphia, PA, USA

T Joyce
City University of New York, New York, NY, USA

JP Kelleher
University of Wisconsin–Madison, Madison, WI, USA

IR Kelly
Queens College of the City University of New York, Flushing, NY, USA

M Lindeboom
VU University Amsterdam, HV Amsterdam, The Netherlands

A Lleras-Muney
UCLA, Los Angeles, CA, USA

G Long
Analysis Group, Inc., Boston, MA, USA

CE Luscombe
Boston University, Boston, MA, USA

D Madden
University College, Dublin, Ireland

A Manca
University of York, York, UK

JA Matheson
University of Leicester, Leicester, England, UK

J Mauskopf
RTI International, NC, USA

A McGuire
LSE Health, London, UK

TG McGuire
Harvard Medical School, Boston, MA, USA

K Meckel
Columbia University, New York, NY, USA

NR Mehta
Riddle Hospital, Media, PA, USA, and University of Pennsylvania, Philadelphia, PA, USA

EM Melhado
University of Illinois at Urbana–Champaign, Urbana, IL, USA

A Mills
London School of Hygiene and Tropical Medicine, London, UK

MA Morrisey
University of Alabama at Birmingham, Birmingham, AL, USA

R Mortimer
Analysis Group, Inc., Boston, MA, USA

J Mullahy
University of Wisconsin-Madison, Madison, USA

JE Murray
Rhodes College, Memphis, TN, USA

NY Ng
Yale School of Public Health, New Haven, CT, USA

S Nikolova
University of Manchester, Manchester, UK

E Nord
Norwegian Institute of Public Health, Oslo, Norway, and The University of Oslo, Oslo, Norway

JA Nyman
University of Minnesota, Minneapolis, MN, USA

MV Pauly
University of Pennsylvania, Philadelphia, PA, USA

D Polsky
University of Pennsylvania, Philadelphia, PA, USA

K Reinhardt
Centre on Global Health Security, UK

TJ Rephann
Charlottesville, VA, USA

P Rosa Dias
University of Sussex, Brighton, UK

JP Ruger
Yale Schools of Medicine, Public Health, and Law, New Haven, CT, USA

JA Salomon
Harvard School of Public Health, Boston, MA, USA

I Sanchez
University of York, Heslington, York, UK

RE Santerre
University of Connecticut, Storrs, CT, USA

MJ Sculpher
University of York, York, UK

L Siciliani
University of York, Heslington, York, UK

R Smith
London School of Hygiene and Tropical Medicine, London, UK

M Soares
University of York, York, UK

RR Soares
São Paulo School of Economics, FGV-SP, São Paulo, SP, Brazil

N Spicer
London School of Hygiene and Tropical Medicine, London, UK

T Stoltzfus Jost
Washington and Lee University, Harrisonburg, VA, USA

OR Straume
University of Minho, Braga, Portugal

M Sutton
University of Manchester, Manchester, UK

E Umapathi
George Washington University, Washington, DC, USA

TS Vogl
Princeton University, Princeton, NJ, USA, and The National Bureau of Economic Research, Cambridge, MA, USA

M Vujicic
Health Policy Resources Center, Chicago, IL, USA

D de Walque
The World Bank, Washington, DC, USA

TN Wanchek
Charlottesville, VA, USA

HLA Weatherly
University of York, York, UK

G Wester
McGill University, Montréal, QC, Canada

Elizabeth T Wilde
Columbia University, New York, NY, USA

I Williams
University of Birmingham, Birmingham, UK

AS Wilmot
University of Pennsylvania, Philadelphia, PA, USA

J Wolff
University College London, London, UK

SH Zuvekas
Agency for Healthcare Research and Quality, Rockville, MD, USA

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you're interested in: an alphabetical contents list, cross-references to other relevant articles within each article, and a full subject index.

1 Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

2 Cross-References

Most of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an entry as a See also list, serve four different functions:

i. To draw the reader's attention to related material in other entries.
ii. To indicate material that broadens and extends the scope of the article.
iii. To indicate material that covers a topic in more depth.
iv. To direct readers to other articles by the same author(s).

Example

The following list of cross-references appears at the end of the entry Abortion.

> *See also*: Education and Health in Developing Economies. Fertility and Population in Developing Countries. Global Public Goods and Health. Infectious Disease Externalities. Nutrition, Health, and Economic Performance. Water Supply and Sanitation

3 Index

The index includes page numbers for quick reference to the information you're looking for. The index entries differentiate between references to a whole entry, a part of an entry, and a table or figure.

4 Contributors

At the start of each volume there is list of the authors who contributed to that volume.

SUBJECT CLASSIFICATION

Demand for Health and Health Care

Collective Purchasing of Health Care
Demand Cross Elasticities and 'Offset Effects'
Demand for Insurance That Nudges Demand
Education and Health: Disentangling Causal Relationships from Associations
Health Care Demand, Empirical Determinants of
Medical Decision Making and Demand
Peer Effects, Social Networks, and Healthcare Demand
Physician-Induced Demand
Physician Management of Demand at the Point of Care
Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment
Quality Reporting and Demand
Rationing of Demand

Determinants of Health and Ill-Health

Abortion
Addiction
Advertising as a Determinant of Health in the USA
Aging: Health at Advanced Ages
Alcohol
Education and Health
Illegal Drug Use, Health Effects of
Intergenerational Effects on Health – *In Utero* and Early Life
Macroeconomy and Health
Mental Health, Determinants of
Nutrition, Economics of
Peer Effects in Health Behaviors
Pollution and Health
Sex Work and Risky Sex in Developing Countries
Smoking, Economics of

Economic Evaluation

Adoption of New Technologies, Using Economic Evaluation
Analysing Heterogeneity to Support Decision Making
Budget-Impact Analysis
Cost-Effectiveness Modeling Using Health State Utility Values
Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties
Economic Evaluation, Uncertainty in
Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis
Infectious Disease Modeling
Information Analysis, Value of
Observational Studies in Economic Evaluation
Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes
Problem Structuring for Health Economic Model Development
Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation
Searching and Reviewing Nonclinical Evidence for Economic Evaluation
Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies
Statistical Issues in Economic Evaluations
Synthesizing Clinical Evidence for Economic Evaluation
Value of Information Methods to Prioritize Research
Valuing Informal Care for Economic Evaluation

Efficiency and Equity

Efficiency and Equity in Health: Philosophical Considerations
Efficiency in Health Care, Concepts of
Equality of Opportunity in Health
Evaluating Efficiency of a Health Care System in the Developed World
Health and Health Care, Need for
Impact of Income Inequality on Health
Measuring Equality and Equity in Health and Health Care
Measuring Health Inequalities Using the Concentration Index Approach
Measuring Vertical Inequity in the Delivery of Healthcare
Resource Allocation Funding Formulae, Efficiency of
Theory of System Level Efficiency in Health Care
Welfarism and Extra-Welfarism

Global Health

Education and Health in Developing Economies
Fertility and Population in Developing Countries

Health and Its Value

Health and the Macroeconomy

Health Econometrics

Health Insurance

Human Resources

Markets in Health Care

Pharmaceutical and Medical Equipment Industries

Public Health

Supply of Health Services

Ambulance and Patient Transport Services
Cost Function Estimates
Healthcare Safety Net in the US
Home Health Services, Economics of
Long-Term Care
Production Functions for Medical Services
Understanding Medical Tourism

PREFACE

What Do Health Economists Do?

This encyclopedia gives the reader ample opportunity to read about what it is that health economists do and the ways in which they set about doing it. One may suppose that health economics consist of no more than the application of the discipline of economics (that is, economic theory and economic ways of doing empirical work) to the two topics of health and healthcare. However, although that would usefully uncouple 'economics' from an exclusive association with 'the (monetized) economy,' markets, and prices, it would miss out a great deal of what it is that health economists actually do, irrespective of whether they are being descriptive, theoretical, or applied. One distinctive characteristic of health economics is the way in which there has been a process of absorption into it (and, undoubtedly, from it too); in particular, the absorption of ideas and ways of working from biostatistics, clinical subjects, cognitive psychology, decision theory, demography, epidemiology, ethics, political science, public administration, and other disciplines already associated with 'health services research' (HSR) and, although more narrowly, 'health technology assessment' (HTA). But to identify health economics with HSR or HTA would also miss much else that health economists do.

… And How Do They Do It?

As for the ways in which they do it, in practice, the overwhelming majority of health economists use the familiar theoretical tools of neoclassical economics, although by no means all (possibly not even a majority) are committed to the welfarist (specifically the Paretian) approach usually adopted by mainstream economists when addressing normative issues, which actually turns out to have been a territory in which some of the most innovative ideas of health economics have been generated. Health economists are also more guarded than most other economists in their use of the postulates of soi-disant 'rationality' and in their beliefs about what unregulated markets can achieve. To study healthcare markets is emphatically not, of course, necessarily to advocate their use.

A Schematic of Health Economics

To think of health economics merely in these various restricted ways would be indeed to miss a great deal. The broader span of subject matter may be seen from the plumbing diagram, in which I have attempted to illustrate the entire range of topics in health economics. A version of the current schematic first appeared in Williams (1997, p. 46). The content of the encyclopedia follows, broadly, this same structure. The arrows in the diagram indicate a natural logical and empirical order, beginning with **Box A** (Health and its value) (**Figure 1**).

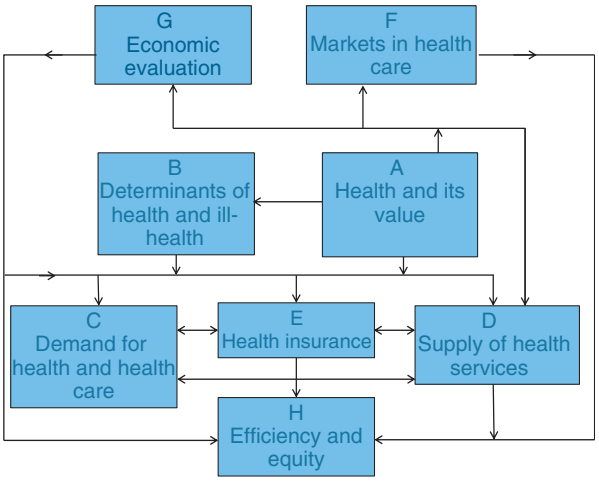


Figure 1 A schematic of health economics.

Box A, in the center-right of the schematic, contains fundamental concepts and measures of population health and health outcomes, along with the normative methods of welfarism and extra-welfarism; measures of utility and health outcomes, including their uses and limitations; and methods of health outcome valuation, such as willingness to pay and experimental methods for revealing such values, and their uses and limitations. It includes macro health economic topics like the global burden of disease, international trade, public and private healthcare expenditures, Gross Domestic Product (GDP) and healthcare expenditure, technological change, and economic growth. Some of the material here is common to epidemiology and bioethics.

Box A Health and its value

Concepts and measures of population health and health outcomes.
Ethical approaches (e.g., welfarism and extrawelfarism).
Measures of utility and the principal health outcome measures, their uses, and limitations.
Health outcome valuation methods, willingness to pay, their uses, and limitations.
Macro health economics: global burdens of disease, international trade, healthcare expenditures, GDP, technological change, and economic growth.

Box B (Determinants of health and ill health) builds on these basics in various 'big-picture' topics, such as the population health perspective for analysis and the determinants of lifetime health, such as genetics, early parenting, and schooling; it embraces occupational health and safety, addiction (especially tobacco, alcohol, and drugs), inequality as a determinant of ill health, poverty and the global burden of disease in low- and middle-income countries, epidemics, prevention, and public health technologies. Here too, much is

Box B Determinants of health and ill health

The population health perspective.
Early determinants of lifetime health (e.g., genetics, parenting, and schooling).
Occupational health and safety.
Addiction: tobacco, alcohol, and drugs.
Inequality as a determinant of ill health.
Poverty and global health (in LMICs).
Epidemics.
Prevention.
Public health technologies.

shared, both empirically and conceptually, with other disciplines.

From this it is a relatively short step into **Box C** (Demand for health and healthcare): here we are concerned with the difference between demand and need; the demand for health as 'human capital'; the demand for healthcare (as compared with health) and its mediation by 'agents' like doctors on behalf of 'principals'; income and price elasticities; information asymmetries (as in the different types of knowledge and understandings by patients and healthcare professionals, respectively) and agency relationships (when one, such as a health professional, acts on behalf of another, such as a patient); externalities or spillovers (when one person's health or behavior directly affects that of another) and publicness (the quality which means that goods or services provided for one are also necessarily provided for others, like proximity to a hospital); and supplier-induced demand (as when a professional recommends and supplies care driven by other interests than the patient's).

Box C Demand for health and healthcare

Demand and need.
The demand for health as human capital.
The demand for healthcare.
Agency relationships in healthcare.
Income and price elasticities.
Information asymmetries and agency relationships.
Externalities and publicness.
Supplier-induced demand.

Then comes **Box D** (Supply of healthcare) covering human resources; the remuneration and behavior of professionals; investment and training of professionals in healthcare; monopoly and competition in healthcare supply; for-profit and nonprofit models of healthcare institutions like hospitals and clinics; health production functions; healthcare cost and production functions that explore the links between 'what goes in' and 'what comes out;' economies of scale and scope; quality of care and service; and the safety of interventions and modes of delivery. It includes the estimation of cost functions and the economics of the pharmaceutical and medical equipment industries. A distinctive difference in this territory from many other areas of application is the need to drop the assumption

Box D Supply of health services

Human resources, remuneration, and the behavior of professionals.
Investment and training of professionals in healthcare.
Monopoly and competition in healthcare supply.
Models of healthcare institutions (for-profit and nonprofit).
Health production functions.
Healthcare cost and production functions.
Economies of scale and scope.
Quality and safety.
The pharmaceutical and medical equipment industries.

of profit-maximizing as a common approach to institutional behavior and to incorporate the idea of 'professionalism' when explaining or predicting the responses of healthcare professionals to changes in their environment.

Supply and demand are mediated (at least in the high-income world) by insurance: the major topic of **Box E** and a large part of health economics as practiced in the US. This covers the demand for insurance; the supply of insurance services and the motivations and regulations of insurance as an industry; moral hazard (the effect of insurance on utilization); adverse selection (the effect of insurance on who is insured); equity and health insurance; private and public systems of insurance; the welfare effects of soi-disant 'excess' insurance; effects of insurance on healthcare providers; and various specific issues in coverage, such as services to be covered in an insured bundle and individual eligibility to receive care. Although the health insurance industry occupies a smaller place in most countries outside the US, the issues invariably crop up in a different guise and require different regulatory and other responses.

Box E Health insurance

The demand for insurance.
The supply of insurance services.
Moral hazard.
Adverse selection.
Equity and health insurance.
Private and public systems.
Welfare effects of 'excess' insurance.
Effects of insurance on healthcare providers.
Issues in coverage: services covered and individual eligibility.
Coverage in LMICs.

Then, in **Box F**, comes a major area of applied health economics: markets in healthcare and the balance between private and public provision, the roles of regulation and subsidy, and the mostly highly politicized topics in health policy. This box includes information and how its absence or distortion corrupts markets; other forms of market failure due to externalities; monopolies and a catalog of practical difficulties both for the market and for more centrally planned systems; labor markets in healthcare (physicians, nurses, managers, and allied professions), internal markets (as when the public sector of healthcare is divided into agencies that commission care on behalf of populations and those that

Box F Markets in healthcare

Information and markets and market failure.
Labor markets in healthcare: physicians, nurses, managers, and allied professions.
Internal markets in the healthcare sector.
Rationing and prioritization.
Welfare economics and system evaluation.
Comparative systems.
Waiting times and lists.
Discrimination.
Public goods and externalities.
Regulation and subsidy.

provide it); rationing and the various forms it can take; welfare economics and system evaluation; waiting times and lists; and discrimination. It is here that many of the features that make healthcare 'different' from other goods and services become prominent.

Box G is about evaluation and healthcare investment, a field in which the applied literature is huge. It includes cost-benefit analysis, cost-utility analysis, cost-effectiveness analysis, and cost-consequences analysis; their application in rich and poor countries; the use of economics in medical decision making (such as the creation of clinical guidelines); discounting and interest rates; sensitivity analysis as a means of testing how dependent one's results are on assumptions; the use of evidence, efficacy, and effectiveness; HTA, study design, and decision process design in agencies with formulary-type decisions to make; the treatment of risk and uncertainty; modeling made necessary by the absence of data generated in trials; and systematic reviews and meta-analyses of existing literature. This territory has burgeoned especially, thanks to the rise of 'evidence-based' decision making and the demand from regulators for decision rules in determining the composition of insured bundles and the setting of pharmaceutical prices.

Box G Economic evaluation

Decision rules in healthcare investment.
Techniques of cost-benefit analysis in health and healthcare.
Techniques of cost-utility analysis and cost-effectiveness analysis in health and healthcare in rich and poor countries.
Techniques of cost-consequences analysis.
Decision theoretical approaches.
Outcome measures and their interpretation.
Discounting.
Sensitivity analysis.
Evidence, efficacy, and effectiveness.
Economics and health technology assessment.
Study design.
Risk and uncertainty.
Modeling.
Systematic reviews and meta-analyses.

The final Box, H, draws on all the preceding theoretical and empirical work: concepts of efficiency, equity, and possible conflicts between them; inequality and the socioeconomic 'gradient;' techniques for measuring equity and inequity; evaluating efficiency at the system level; evaluating equity at system level: financing arrangements; evaluating equity at system level: service access and delivery; institutional arrangements for efficiency and equity; policies against global poverty and for health; universality and comprehensiveness as global objectives of healthcare; and healthcare financing and delivery systems in low- and middle-income countries (LMICs). This is the most overtly 'political' and policy-oriented territory.

Box H Efficiency and equity

Concepts of efficiency, equity, and possible conflicts.
Inequality and the socioeconomic 'gradient.'
Evaluating efficiency: international comparisons.
Techniques for measuring equity and inequity.
Evaluating equity at system level: financing arrangements.
Evaluating equity at system level: service access and delivery.
Institutional arrangements for efficiency and equity.
Global poverty and health.
Universality and comprehensiveness.
Healthcare financing and delivery systems in LMICs.

A Word on Textbooks

The scope of a subject is often revealed by the contents of its textbooks. There are now many textbooks in health economics, having various degrees of sophistication, breadth of coverage, balance of description, theory and application, and political sympathies. They are not reviewed here but I have tried to make the (English language) list in the Further Reading as complete as possible. Because the assumptions that textbook writers make about the preexisting experience of readers and about their professional backgrounds vary, not every text listed here will suit every potential reader. Moreover, a few have the breadth of coverage indicated in the schematic here. Those interested in learning more about the subject to supplement what is to be gleaned from the pages of this encyclopedia are, therefore, urged to sample what is on offer before purchase.

Acknowledgments

My debts of gratitude are owed to many people. I must particularly thank Richard Berryman (Senior Project Manager), at Elsevier, who oversaw the inception of the project, and Gemma Taft (Project Manager) and Joanne Williams (Associate Project Manager), who gave me the most marvelous advice and support throughout. The editorial heavy lifting was done by Billy Jack and Karen Grépin (Global Health); Aki Tsuchiya and John Wildman (Efficiency and Equity); John Cawley and Kosali Simon (Determinants of Health and Ill health); Richard Cookson and Mark Suhrcke (Public Health); Erik Nord (Health and its Value); Richard Smith (Health and the

Macroeconomy); John Mullahy and Anirban Basu (Health Econometrics); Tom McGuire (Demand for Health and Healthcare); John Nyman (Health Insurance); Jim Burgess (Supply of Health Services); Martin Gaynor and Sean Nicholson (Human Resources); Patricia Danzon (Pharmaceutical and Medical Equipment Industries); Pau Olivella and Pedro Pita Barros (Markets in Healthcare); and John Brazier, Mark Sculpher, and Anirban Basu (Economic Evaluation). Finally, my thanks to the Advisory Board: Ron Akehurst, Andy Briggs, Martin Buxton, May Cheng, Mike Drummond, Tom Getzen, Jane Hall, Andrew Jones, Bengt Jonsson, Di McIntyre, David Madden, Jo Mauskopf, Alan Maynard, Anne Mills, the late Gavin Mooney, Jo Newhouse, Carol Propper, Ravindra Rannan-Eliya, Jeff Richardson, Lise Rochaix, Louise Russell, Peter Smith, Adrian Towse, Wynand Van de Ven, Bobbi Wolfe, and Peter Zweifel. Although the Board was not called on for frequent help, their strategic advice and willingness to be available when I needed them was a great comfort.

Anthony J Culyer
Universities of Toronto (Canada) and York (England)

Further Reading

Cullis, J. G. and West, P. A. (1979). *The economics of health: An introduction*. Oxford: Martin Robertson.

Donaldson, C., Gerard, K., Mitton, C., Jan, S. and Wiseman, V. (2005). *Economics of health care financing: The visible hand*. London: Palgrave Macmillan.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. and Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*, 3rd ed. oxford: Oxford University Press.

Evans, R. G. (1984). *Strained mercy: The economics of Canadian health care*. Markham, ON: Butterworths.

Feldstein, P. J. (2005). *Health care economics*, 6th ed. Florence, KY: Delmar Learning.

Folland, S., Goodman, A. C. and Stano, M. (2010). *The economics of health and health care*, 6th ed. Upper Saddle River: Prentice Hall.

Getzen, T. E. (2006). *Health economics: Fundamentals and flow of funds*, 3rd ed. Hoboken, NJ: Wiley.

Getzen, T. E. and Allen, B. H. (2007). *Health care economics*. Chichester: Wiley.

Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (eds.) (1996). *Cost-effectiveness in health and medicine*. New York and Oxford: Oxford University Press.

Henderson, J. W. (2004). *Health economics and policy with economic applications*, 3rd ed. Cincinnati: South-Western Publishers.

Hurley, J. E. (2010). *Health economics*. Toronto: McGraw-Hill Ryerson.

Jack, W. (1999). *Principles of health economics for developing countries*. Washington, DC: World Bank.

Jacobs, P. and Rapoport, J. (2004). *The economics of health and medical care*, 5th ed. Sudbury, MA: Jones & Bartlett.

Johnson-Lans, S. (2006). *A health economics primer*. Boston: Addison Wesley/ Pearson.

McGuire, A., Henderson, J. and Mooney, G. (1992). *The economics of health care*. Abingdon: Routledge.

McPake, B., Normand, C. and Smith, S. (2013). *Health economics: An international perspective*, 3rd ed. Abingdon: Routledge.

Mooney, G. H. (2003). *Economics, medicine, and health care*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

Morris, S., Devlin, N. and Parkin, D. (2007). *Economic analysis in health care*. Chichester: Wiley.

Palmer, G. and Ho, M. T. (2008). *Health economics: A critical and global analysis*. Basingstoke: Palgrave Macmillan.

Phelps, C. E. (2012). *Health economics*, 5th (international) ed. Boston: Pearson Education.

Phillips, C. J. (2005). *Health economics: An introduction for health professionals*. Chichester: Wiley (BMJ Books).

Rice, T. H. and Unruh, L. (2009). *The economics of health reconsidered*, 3rd ed. Chicago: Health Administration Press.

Santerre, R. and Neun, S. P. (2007). *Health economics: Theories, insights and industry*, 4th ed. Cincinnati: South-Western Publishing Company.

Sorkin, A. L. (1992). *Health economics – An introduction*. New York: Lexington Books.

Walley, T., Haycox, A. and Boland, A. (2004). *Pharmacoeconomics*. London: Elsevier.

Williams, A. (1997). Being reasonable about the economics of health: Selected essays by Alan Williams (edited by Culyer, A. J. and Maynard, A.). Cheltenham: Edward Elgar.

Witter, S. and Ensor, T. (eds.) (1997). *An introduction to health economics for eastern Europe and the Former Soviet Union*. Chichester: Wiley.

Witter, S., Ensor, T., Jowett, M. and Thompson, R. (2000). *Health economics for developing countries. A practical guide*. London: Macmillan Education.

Wonderling, D., Gruen, R. and Black, N. (2005). *Introduction to health economics*. Maidenhead: Open University Press.

Zweifel, P., Breyer, F. H. J. and Kifmann, M. (2009). *Health economics*, 2nd ed. Oxford: Oxford University Press.

CONTENTS OF ALL VOLUMES

VOLUME 1

VOLUME 2

VOLUME 3

Abortion

T Joyce, City University of New York, New York, NY, USA

Glossary

Difference-in-differences It is subtracting the change in an outcome for a 'control' group from the change in the same outcome among the 'treated' group.

Household production It is the use of time and goods to create commodities such as health.

Identification strategies These are research designs that uncover parameters or associations of interest.

Induced abortion An intentional stoppage of a pregnancy by medication or by surgery.

Infant mortality These are the deaths within the first year of life.

Instrumental variables A statistical method in which a variable is used to isolate variation in a regressor variable that is orthogonal to unobserved components of the outcome of interest.

Pregnancy intention It is the status of a pregnancy as planned, mistimed, or unwanted.

Spontaneous abortion It is the natural termination of a pregnancy.

Introduction

Induced abortion is not an obvious topic in a volume on health economics. Although being a common procedure, abortion does not contribute to rising medical expenditures or inflation. There were 1.1 million surgical abortions in the US in 2008, but the number of abortions has fallen overtime, although the inflation-adjusted cost of a first trimester abortion has remained remarkably stable at approximately $450. Nor have there been dramatic technological breakthroughs in the delivery of abortion. The most significant innovation is RU-486, more commonly referred to as the 'abortion pill.' However, its impact on the demand for and availability of abortion services has been modest at best. Finally, abortions are extremely safe with only 0.7 deaths per year per 100 000 procedures between 1988 and 1997. In contrast, the maternal mortality rate in the US is 15 times greater.

So why include an article on abortion? Two reasons. First, induced abortion, a medical procedure performed only by physicians, is one of the most contentious and divisive issues in the politics of many countries today. In the US, clinicians who perform abortions and staff workers who assist them have been murdered and their clinics vandalized. Politicians are defined by their stance on abortion and Supreme Court nominees must tread carefully when discussing the precedent set by the Court's decision in Roe versus Wade. Academic research on abortion has not been protected from this scrutiny. Donohue and Levitt's (2001) study linking the legalization of abortion to the decrease in homicide rates 20 years later was extremely controversial, received widespread exposure in the popular press, and became a central chapter in the hugely successful book Freakonomics.

The second reason to include a review of abortion is because the indirect effect of abortion on health is potentially large but empirically challenging to document. Induced abortion, the focus of this article, also represents a conscious decision to end a pregnancy, unlike spontaneous abortion which is an involuntary and largely random termination of pregnancy. Arguably the most notable link between abortion and health or well-being is the hypothesized relationship between abortion and crime (Donohue and Levitt, 2001). If Donohue and Levitt are correct, then the legalization of abortion averted 15 000 homicides over a 10-year period (Joyce, 2009). But homicide is but one measure of well-being. If abortion has a profound effect on crime, then it likely affected other measures of well-being such as marriage, schooling, drug use, and sexually transmitted diseases to name but a few. And yet the empirical challenge of isolating a cohort effect from constantly evolving period effects may be insurmountable given the data and methods available to researchers.

In this article the focus is on the link between induced abortion and health. Health is broadly viewed to include measures of well-being such as crime and drug use in addition to the more commonly associated measures of health such as infant mortality. Given space limitations, the author concentrates primarily on the US experience with legalized abortion from roughly 1970 to present. The history of abortion in the US is available from a number of sources (Garrow, 1998). The author concentrates instead on two empirical challenges for researchers that have tried to uncover a link between abortion and health. The first is identification. How does one measure the impact of a pregnancy that is never carried to term? The second is data. Unlike births, induced abortions are not part of a national vital registration system. Moreover, abortions are poorly reported in surveys as women are reluctant to admit to them. Finally, the review is selective. The author discusses in detail papers believed to be the most important because of the quality of the research design and their impact on subsequent research. There is more to be learned by careful study of the best papers than a quick pass through the entire literature.

The article is organized as follows. The author first discusses the conceptual mechanisms by which abortion is linked to health. This is followed by a description of data on abortion and the demographics of abortion. The next few sections discuss empirical work supporting possible links. The literature is broadly divided between studies on the determinants of abortion and its impact on fertility and those that estimate either the structural or reduced-form association between abortion and health. There has been relatively little work on the supply side of abortion markets.

Conceptual Link between Abortion and Health

How does one study the health of a fetus that has never been born? The simple answer is that you cannot, which necessitates indirect approaches. Demographers, for example, consider abortion as an expression of an unwanted pregnancy. They assume that the wantedness of a pregnancy varies along a continuum from those that are aborted to pregnancies that are mistimed but carried to term. Thus, even pregnancies that result in live births may be characterized as unwanted and contrasted with the outcomes of births described as wanted. Data on wantedness in the US come from surveys of new mothers in which they are asked about their pregnancy intention when they first discovered that they were pregnant. Births are classified as wanted, mistimed, or unwanted on the basis of a series of responses by the mother. Mothers whose pregnancies are unwanted at conception are hypothesized to smoke more or receive less prenatal care than mothers whose pregnancies were planned. As a result, births from pregnancies that are unwanted are expected to be less healthy than births from pregnancies that are wanted. Neglect is hypothesized to continue after birth. Children who were unwanted at conception may receive less nurturing than those who are wanted. The result would be lower academic achievement, behavioral problems, and possible delinquency as adolescents (Brown and Eisenberg, 1995).

It is unclear whether unwanted pregnancies based on post hoc surveys of women who gave birth provide insights as to the outcomes of pregnancies that are aborted had they instead been carried to term. Early studies of wantedness in Europe tried to estimate the impact of the latter by analyzing outcomes of women who were denied abortion. The most famous sample is the Prague Cohort of 1961–63. A total of 220 children whose mothers were twice denied an abortion for the same pregnancy were matched to children whose pregnancies had been wanted and followed for 30 years. There were few differences between the unwanted cohort and their wanted controls at birth, but by the age of 20 years, there was evidence of less personal satisfaction and psychological instability.

Economic models that linked abortion and health were first discussed by Grossman and Jacobowitz (1981). The authors argued that abortion as a method of fertility control helps parents to achieve a desired family size. Using models of the family and household production pioneered by Becker and Lewis (1973); Grossman and Jacobowitz (1981) incorporated abortion reform into a model of infant mortality. Parents maximized a utility function that depended on consumption goods, the number of births, and the survival probability of each. Both the number of children and their survival probability were choice variables. The survival probability depended on a set of endogenous inputs. Thus, parents affected the health of an infant by their choice of goods (e.g., cigarettes) and medical care during pregnancy. The model generated a structural and reduced-form production function of child survival. Grossman and Jacobowitz (1981) argued that subsidized family planning services and legalized abortion decreased the cost of fertility control which lowered the optimal number of births but raised the survival probability of each.

This quantity–quality framework became the explicit model in many of the empirical analyses that followed. Lowering the price of an abortion allowed women and parents greater control over the timing and number of children. This gave parents more control over the quality of each child as parents used time and market goods to enhance a child's health and human capital. Thus, pregnant teens could delay birth until they were more financially and emotionally prepared for parenthood. Older women could terminate unwanted fetuses that could divert resources from their current children or abort fetuses that were at a greater risk of poor health. With the advent of genetic testing and advanced sonography, abortion as a fetal selection mechanism became even more explicit.

Refinements of the selection mechanism followed. Abortion was characterized as one decision along a sequence that included the decision to get pregnant, the decision to abort or give birth, and the decision to marry or remain single (Grossman and Joyce, 1990; Lundberg and Plotnick, 1990). Increases in the cost of abortion impacted these other decisions. For instance, some women use pregnancy as a way to assess the suitability of a potential father. Increasing the cost of an abortion raises the price of this 'option,' resulting in fewer abortions but fewer pregnancies as well.

Abortion as a sorting mechanism is not the only pathway through which women and their potential offspring were affected. Akerlof *et al.* (1996) developed a model in which women's bargaining position with men was weakened by the availability of safe, legal abortion. Before abortion, sex was more closely linked to commitment. If an unmarried woman became pregnant, there was pressure on the man to 'to do the right thing' by marrying her. Abortion altered that expectation. Women willing to abort could have sex without an implied commitment of marriage in the case of pregnancy. Men could insist that a pregnancy be terminated instead of marriage. This put women opposed to abortion at a disadvantage in attracting men. To compete for men they had to be more willing to have sex without a commitment of marriage. The model predicts that the legalization of abortion will result in a decrease in 'shotgun' marriages and an increase in out-of-wedlock childbearing. Both predictions are consistent with the stylized facts in the 1970s. The link to health comes through the immiseration of women and children as the number of female-headed households rise. Economists used the model by Akerlof *et al.* (1996) to argue that the legalization of abortion could be associated with the rise in crime, in direct contradiction to Donohue and Levitt (2001).

The Akerlof *et al.* (1996) framework has not been used in the empirical literature on abortion and health. The quantity–quality framework has been the mainstay in the literature. By enabling parents to achieve an optimal number of births, abortion enhances the resources devoted to the children who are born. Thus, any empirical association between abortion and health rests importantly on the association between abortion and fertility. This may seem obvious because an aborted pregnancy is an averted birth. However, other methods of fertility control are substitutes for abortion which implies that a rise in the abortion rates need not be associated with a fall in birth rates. Couples that may have used condoms before the legalization of abortion may be less vigilant about contraception after legalization. A pregnancy that occurs under a regime on legalized abortion may not have occurred

under a regime in which abortion is prohibited. Without demonstrating that a change in the birth rate is associated with a decrease in the price of an abortion, it is difficult to establish that parents are trading off quantity for quality.

Abortion: Data and Demographics

Data

One of the biggest challenges in studying abortion is measuring its incidence. There was no national surveillance system for abortion until 1973, the year of the US Supreme Court decision in Roe versus Wade. In that year the Alan Guttmacher Institute (now the Guttmacher Institute) published its first national estimate of abortions by state. The Guttmacher survey of abortion providers was conducted annually from 1973 to 1988 with exception of 1983. After 1988, however, the periodicity of the survey was increased to every 4 years: 1992, 1996, 2000, 2004, and 2008.

The second major population-based source comes from the Centers for Disease Control and Prevention (CDC). The CDC collects data from state health departments and reports abortions by state, year, and several demographic factors: age, race, marital status, gestational age, type of procedure, parity, and previous induced abortions. There are two advantages to the CDC data. First, the availability of abortion by characteristics of the patient enable studies of policies based on age or gestational age. Second, data are available annually, whereas the Guttmacher Institute reports data periodically. As with data from the Guttmacher Institute, the CDC reports abortions by state of occurrence. In addition, the total number of abortions as reported by the CDC is approximately 15% lower than that reported by the Guttmacher Institute, and the degree of undercounting varies substantially by state. Further, not all states report abortions to the CDC or abortions cross-classified by characteristics of the patient; California and Florida are two populous and notable examples. Finally, the limited cross-tabulation of the data prevents analyses by race or by gestational age.

Although the Guttmacher Institute's periodic survey of abortion providers yields the most widely accepted estimate of the number of abortions, they have two important limitations for policy evaluations. First, abortions are tallied according to the state in which they occur and not according to the state in which a woman resides; and second, data are not available by age or any other characteristic at the state level. To overcome these limitations, Guttmacher researchers have applied the distribution of abortions by state and age as reported by the CDC to estimate the number of abortions by age. They also use information from the CDC on the proportion of abortions provided to nonresidents in a state along with other sources to estimate abortions by state of residence. Thus, it is important to remember that Guttmacher's report of abortions by state of residence are an estimate and that they are unlikely to accurately measure cross-state travel by subpopulations in response to a change in policy. This is an important drawback, which is often ignored.

The third major source of data is state health departments. The CDC uses these same data in its surveillance reports. The major advantage of obtaining them directly from the state is that some states make available to researchers individual-level data on induced abortions, which allows for a more refined aggregation of data than what is available from the CDC. This can substantially improve the internal and external validity of an analysis (the ability to measure what one sets out to measure). The two major drawbacks to these data are similar to those stated above: completeness of reporting varies by state and residents who leave their state for an abortion are rarely counted by the state in which they reside. However, the latter drawback can be overcome if researchers are able to secure data from neighboring states.

The lack of data by state of residence is a major limitation. Studies of parental involvement (PI) laws and mandatory delay statutes based on data by state of occurrence will overestimate the decline in abortions associated with the laws, not only because residents leave the state for an abortion but also because nonresidents stop entering the state for an abortion. Studies of PI laws in the 1980s and the early 1990s were particularly vulnerable to this source of bias, as only 13 states had such laws in 1988. This made travel outside one's state of residence feasible. More recent evaluations are less vulnerable to this source of bias because 35 states, including almost all states in the South and Midwest, now have PI laws. This makes traveling to a state without a law very challenging.

Other information on abortion is available from population-based surveys. The National Longitudinal Survey of Youth 1979 and 1997 ask respondents about previous abortions. The National Survey of Family Growth also queries respondents about past abortions. However, surveys grossly underestimate the number of abortions as many women do not report them. Moreover, the underreporting is not random: young, poor, and minority women appear to underreport more than other demographic groups. This has greatly limited the use of these data to evaluate policy.

Another source of data on the characteristics of women who have abortions comes from the Guttmacher Institute and its periodic survey of abortion patients. Using a sample of nationally representative abortion clinics, researchers survey patients waiting to have an abortion. The data are weighted to be nationally representative. In addition to data on age and race, they have information on income and insurance status.

Demographics

The abortion rate is defined as number of induced abortions per 1000 women of 15–44 years of age. In the US in 1973, there were 744 600 abortions and the rate was 16.3. As shown in **Figure 1**, the abortion rate rose after the decision in Roe versus Wade and peaked in 1981 at 29.3. It has fallen almost continuously since then. In 2008, there were 1 212 400 abortions and the rate stood at 19.6.

Table 1 shows the abortion rates by characteristics of the patients in 1998 and 2007 based on the CDCs annual surveillance of reporting states. The abortion rate was greatest for women of 20 to 24 years of age at 35.6 per 1000 in 1998 and 30.0 per 1000 in 2007. The teen abortion rates fell 25.3% over the same period from 19.8 to 14.8. The abortion rate of Blacks (37.8 in 1998) was more than 3 times that of Whites, and the

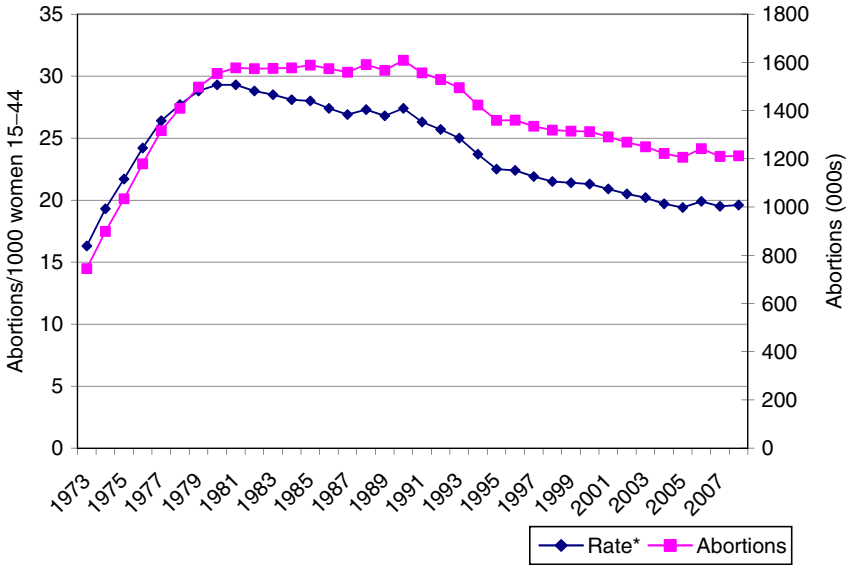


Figure 1 Number of abortions (in thousands) and abortion rate in the US, 1973–2008. Reproduced from Jones, R. K. and Kooistra, K. (2011). Abortion incidence and access to services in the United States, 2008. *Perspectives on Sexual and Reproductive Health* **43**(1), 41–50.

Table 1 The US abortion rates by characteristics of patients, 1998 and 2007

	1998	*2007*	*% Change*
Age (years)[a]			
<15	1.9	1.2	−36.8
15–19	19.8	14.8	−25.3
20–24	35.6	30.0	−15.7
25–29	24.2	22.0	−9.1
30–34	13.6	13.7	0.7
35–39	7.3	7.9	8.2
40>	2.5	2.7	8.0
Race[b]			
White	12.4	10.9	−12.1
Black	37.8	33.5	−11.4
Other	25.7	23.3	−9.3
Ethnicity[b]			
Hispanic	27.1	22.2	−18.1
Non-Hispanic	16.9	15.1	−10.7
Gestation (%)[c]			
<=13 wks	90.6	91.6	1.1
<=8	55.8	63.6	14.0
9–13	34.8	28.0	−19.5
>13 wks	9.4	8.4	−10.6
14–15	3.4	3.3	−2.9
16–17	2.1	1.8	−14.3
18–20	2.3	2.0	−13.0
>=21	1.5	1.3	−13.3

[a]Abortions per 1000 women of the specific age.
[b]Abortions per 1000 race or ethnic-specific women.
[c]Percent distribution of abortions.
Source: Reproduced from Centers for Disease Control and Prevention (2011). Abortion Surveillance – US, 2007. *Morbidity and Mortality Weekly Report* **60**(1), 1–39.

Table 2 Abortion rates by socioeconomic characteristics of patients, 1994 and 2000

	1994	*2000*	*% Change*
Education[a]			
Not HS grad	22	23	7
HS grad/GED	20	20	1
Some college	29	26	−12
College grad	19	13	−30
Poverty status[±]			
<100%	36	44	−12.1
100–199%	31	38	−11.4
200–299%	25	21	−9.3
>=300%	16	10	−9.3
Medicaid coverage[±]			
Yes	50	57	14
No	20	18	−12

[a]Abortions per 1000 women 15–44 in the respective category.
Abbreviations: GED, general educational development; HS, high school.
Source: Reproduced from Jones, R. K., Darroch, J. E. and Henshaw, S. K. (2002). Patterns in the socioeconomic characteristics of women obtaining abortions in 2000–2001. *Perspectives on Sexual and Reproductive Health* **34**(5), 226–235.

rate for Hispanics more than 2 times that of Whites in both 1998 and 2007. Abortion rates have fallen for all three groups since 1998. More than 90% of abortions are performed at 13 weeks or less gestation. The percent of abortions less than or equal to 8 weeks gestation has risen from 55.8% in 1998 to 63.6% in 2007. This coincides with the growth in medical abortions which accounted for almost 14% of abortions among reporting states. The percent of abortions at or after 21 weeks gestation fell from 1.5% to 1.3% between 1998 and 2007.

Abortions by education, poverty status, and insurance coverage are shown in Table 2. These estimates are from the Guttmacher Institute's periodic survey of abortion patients and are weighted to be nationally representative. Several figures stand out. First, differences by poverty status are striking. In the year 2000, women from families with less than 100% of

the federal poverty level in that year have more than four times the abortion rate of women from families at 300% or more of the federal poverty level. The abortion rate of women with Medicaid coverage, at 57 per 1000 Medicaid recipients, are even higher than those of women in poverty. However, differences in abortion rates by education, a permanent measure of human capital, are much more muted.

In sum, age, race, and income are the three most important correlates of abortion rates in the US. They suggest that young, Black women in poverty are at much higher risk of an unintended pregnancy than their older, White counterparts. The figures also underscore the importance of abortion as a method of fertility control among young, poor, and minority women. If the elasticity of demand for abortion services is greater among less advantaged groups, then policies that raise the cost or lessen the availability of abortion services are likely to impact these groups more than women whose rate of unintended pregnancy is less.

Overview of Studies on Health

Studies of the relationship between abortion and health have progressed with advances in the field of applied microeconometrics. Borrowing from the medical sciences, random control trials (RCTs) have become the gold standard. RCTs remain rare in economics, but their acknowledged quality has pushed researchers to design studies with strong internal validity and transparent sources of identification. In this section, the author reviews the evolution of studies linking abortion and health through the improvement in research design. Early studies of abortion and health relied on cross-sectional variation to identify an association. The second phase of studies on abortion and health leveraged panel data and changes in policy to understand the determinants of abortion and its impact on fertility. A related group of studies used panel methods to estimate the cohort effect of abortion legalization on broad measures of well-being. The most recent set of studies has employed abortion legalization as an instrument for births in an effort to estimate changes not only in the health of birth cohorts exposed to legalized abortion *in utero* but also to estimate the potential health of children that were not born.

Early Studies of Abortion and Health

As noted above, Grossman and Jacobowitz (1981) were the first to use the household production function framework to associate access to abortion with infant health. The empirical work involved regressions of county-level neonatal mortality rates averaged over 3 years from 1970 to 1972 on measures of the cost of fertility control and other inputs into the production of health. They used two measures of abortion availability: Dichotomous indicators of whether the county was in a state that had reformed or legalized abortion by 1970 and the 3 year average of the state abortion rate (abortions per 1000 live births) from 1970 to 1972. And they applied coefficients from the cross-sectional regression to estimate the reduction in neonatal mortality attributable to each input. Overall the model could explain between 35% and 53% of the decline in neonatal mortality between 1971 and 1977. However the most striking result was that the abortion rate accounted for more than 50 of the explained decline for both White and non-Whites.

A series of papers followed the Grossman and Jacobowitz (1981) framework but with more recent data and greater attention to the endogeneity of abortion in the production of infant health. In one study economists estimated the reduced form production function of infant health. The outcome was again the county-level neonatal mortality rate averaged over 3 years (1976–78). They included proxies for the price of inputs such as the number of abortion providers in the county or the number of maternal and child health clinics. The results suggested that an increase in the number of abortion providers was strongly associated with decreases in neonatal mortality. Other economists used the county-level neonatal mortality rate in an effort to estimate the structural production function of infant health. They were interested in the pathways through which abortion affected survival. Thus, they also estimated structural models of low birth weight and preterm births. They included the abortion rate as well as the number of teenage users of family planning clinics as determinants of each outcome. They used two-stage least squares (TSLS) to account for the endogeneity of the abortion rate with number of abortion providers per county as an instrument (more on the instruments below). They found that state-level abortion rates were inversely correlated with neonatal mortality, low birth weight, and preterm birth. Moreover, they argued that abortion improved newborn survival by lowering the incidence of low birth weight births. Others followed this approach by estimating structural models of infant survival. However, their objective was to understand the relative contribution of government programs. These include participation in the Supplemental Nutrition Program for Women, Infants, and Children (WIC), inpatient days in neonatal intensive care units, use of family planning clinics, as well as maternal and child health clinics. As did other economists, these authors used TSLS with the availability of clinics, abortion providers, and neonatal beds as instruments. They reported that the abortion rate explained approximately half of the decline in neonatal mortality between 1964 and 1977 accounted for by the model.

The aforementioned studies used aggregate data to correlate the abortion rate with county-level measures of health. All reasoned that areas with higher abortion rates had a more optimal distribution of birth outcomes as less healthy or desired fetuses were aborted. An ecological approach appeared the only way to associate abortion to health. At the individual level, a pregnancy that is terminated is eliminated from the sample of births. There seemed to be no individual-level analog to the aggregate analysis. However, in two papers, economists applied the emerging econometrics on censored samples to analyze the effect of pregnancy resolution on birth outcomes (Grossman and Joyce, 1990). In both papers, the authors used individual-level data on births and abortions in New York City. The birth and abortion files contained information on age, race, marital status, parity, schooling, as well as measures of the availability of family planning and abortion services by neighborhood. The authors concatenated the files

to create a sample of pregnancies that resulted in either an induced abortion or a live birth. They argued that the sample of births represented a nonrandom draw from the population of pregnancies. In one paper, the authors used the decision to give birth conditional on pregnancy as an expression of wantedness. Women who were selected in the birth sample were more likely to obtain timely prenatal care than those who aborted had they instead carried to term. They estimated the observed counterfactual by using the inverse Mill's ratio to obtain the expected number of months a woman would have delayed prenatal care had she not aborted. The difference between the expected and actual months of delay for women with the same observables became an estimate of the impact of 'wantedness' on the demand for health-producing inputs. They found that women who had a greater probability of giving birth had less than expected delay in prenatal care.

Grossman and Joyce (1990) extended the model to include birth outcomes while treating prenatal care as an endogenous input into the production of health. They also provide a framework that signed the effect of changes in the cost of abortion, the cost of contraception, and underlying health endowment of the fetus. They treated contraception and abortion as substitutes. An increase in the cost of contraception or a decrease in the cost of abortion raises the probability of becoming pregnant. However, an increase in the cost of abortion holding the cost of contraception constant raises the probability of giving birth, conditional on becoming pregnant. For instance, assume that Black women face a higher cost of contraception due to less access and information. A decrease in the cost of abortion will lower the probability of giving birth conditional on pregnancy, increase the demand for healthy inputs, and increase birth weight. This is what the authors found for Black women but not for Whites.

These early papers were important because they tried to develop an empirical test of the association between abortion and health. They used the household production framework to incorporate the cost of fertility control in models of the quantity and quality of children. The statistical analyses became progressively more sophisticated as researchers applied recent advances in econometrics to account for the endogeneity of inputs. However, the identification strategies used then would never meet the standards of today. First, all data were cross-sectional. The lack of a panel precluded fixed effects, which would have limited the identifying variation to within-area changes in policy. Instead, authors compared the impact of abortion rates on birth outcomes in, for example, Utah relative to New York. Given the limited number of covariates, the likelihood of omitted variable bias was large. Even reduced-form analyses suffered from problems of endogeneity. The number of abortion providers in a state or county, for instance, represents the interplay of the supply and demand of abortion services instead of some exogenous measure of price. The sample selection models used by Joyce and Grossman (1990) were novel applications at the time but again lacked a credible identification strategy. More importantly, the robustness of these models depends on the availability of instruments that predict the probability of giving birth but which have no direct effect on the birth outcome. None of the instruments in the two papers could be credibly excluded from the birth outcome equation. Despite these serious drawbacks, this early work motivated subsequent studies that paid much greater attention to identification and for much of the 1990s focused on reduced-form policy questions.

A paper by economists in the mid-1990s provided a segue to the reduced-form policy-orientated papers that soon followed. The authors took the model of Grossman and Joyce (1990) as their starting point. They used individual data from the National Longitudinal Survey of Youth 1979 (NLSY79) to estimate the impact of the price of abortion on birth outcomes. State policies regarding the public financing of abortion through Medicaid served as proxies for the price of abortion in the reduced-form production function of infant health. They found no association between Medicaid financing restrictions and birth weight. In the second part of the paper, they estimated the birth probability equation and found a robust association between Medicaid financing of abortion and the decreased probability of giving birth. For Black women, the availability of Medicaid financing lowered the probability of birth by 0.10 over a mean of 0.88, which is a large effect.

Several features of the analysis are noteworthy. First, the authors used 10 years of data from the NLSY79 and were able to exploit changes in policy over time. Second, they focused on the reduced form instead of the structural production function of health. However, they used random effects instead of state-fixed effects to control for unobserved cross-state heterogeneity. A random effects specification assumes that unobserved state factors (the random effects) are uncorrelated with the policy under study; in this case Medicaid financing of abortions. This was unlikely because mostly liberal states continued to use public funds for abortion after the Hyde Amendment in 1976. In addition, there is very little within-state variation in Medicaid financing of abortion. The big changes in Medicaid came in the late 1970s with the Hyde Amendment. In other words, despite the use of longitudinal data, their policy estimates are essentially obtained from cross-sectional variation in Medicaid financing of abortion. Nevertheless, the paper represented a bridge to subsequent papers in the 1990s that took advantage of panel data with state-fixed effects to eliminate confounding from hard-to-measure differences between states and counties.

Abortion Policy and Fertility in the 1990s

Work on abortion and health in 1990s was shaped by the advances in applied microeconometrics. A series of seminal papers in the econometric literature described the conditions that must hold before instrumental variable methods would yield even limited estimates of treatment effects. The 1990s also saw more emphasis on transparent sources of variation and the quality of the comparison group. The difference-in-difference (DD) methodology became popular because it focused on the reduced form and plausible counterfactuals. There was also much more use of panel data given the attention to pre–post contrasts. Another development was interest in the effect of abortion policy on fertility. This relationship is key to the household production model. If researchers can not demonstrate a relationship between the price of fertility control and the number or timing of births,

then abortion may not play an important role in the quality–quantity trade-off envisioned by its early proponents.

The most important policy change in the US was the legalization of abortion. This occurred largely in two steps. From September of 1969 through December of 1970, abortion became de facto or *de jure* legal in 5 states (Alaska, California, Hawaii, New York, and Washington) and the District of Columbia (Lader, 1974). Abortion became legal nationally with the US Supreme Court decision in Roe versus Wade in January of 1973. The two-step process toward national legalization provided plausibly exogenous sources of variation with which to identify the effect of the availability of abortion services on fertility. An early paper looked at the impact of the legalization of abortion in New York on teen birth rates in New York City in the years before Roe. Lacking data from a control state, the authors used an interrupted time series analysis to estimate the monthly change in White and non-White teen births after abortion became legal in July of 1973. They found that White and non-White births fell 14% and 18%, respectively, in the 24 months after the law went into effect.

Levine *et al.* (1999), however, were the first to exploit the staggered process of legalization within a DD strategy to obtain the most credible estimates of the effect of a decrease in the price of fertility control on birth rates. Using natality data from all 50 states and the District of Columbia, they contrasted changes in fertility from 1961 to 1980 among the early versus the later legalizing states. Overall birth rates fell almost 5% more among women in the early compared with later legalizing states. However, when the authors took account of distance to the nearest legalizing states, the results showed that birth rates fell 10% among those that lived more than 750 miles away from the nearest state in which abortion was legal. Surprisingly, there was no distance gradient for those who lived within 750 miles. Specifically, birth rates fell 4.5% regardless of whether women resided 250 miles away or between 250 and 750 miles from a state with legalized abortion. The study was a classic example of a DD and provided convincing evidence that the early legalization of abortion had an immediate effect on fertility. Some of these same authors would further exploit this natural experiment to analyze changes in well-being associated with changes in fertility.

Post-Roe Policies

Although induced abortion was declared a fundamental right, it remained highly controversial. State governments moved quickly to find the legal limits of regulation. Three state policies have dominated both the political discourse and academic research. The first is the Hyde Amendment, which prohibited the use of federal funds to cover the cost of an abortion unless the mother's life is in danger. The second is PI laws which require that a physician notify or obtain consent from a parent or parents before performing an abortion on a minor, usually defined as girls less than 18 years of age. The third policy is a mandatory delay and counseling statute. This requires that women receive state-mandated information regarding the abortion procedure, the status of the fetus, and alternatives to abortion usually 24 h before the termination. Each policy has been used by economists to analyze changes primarily in abortion and birth rates, although some have looked at the reduced-form association with health. In this summary the focus is on a selected group of studies based on the quality of the design and their impact on subsequent work.

Medicaid

In 1976, Congress passed the Hyde Amendment, which bans federal funding of abortion in all but the most extreme circumstances. The statute prohibits expenditure of federal funds for abortion services except in cases where the continuation of the pregnancy threatened the woman's life. Currently, 17 states use their own funds to pay for all or most medically necessary abortions sought by Medicaid recipients.

The impact of Medicaid financing restrictions has been analyzed extensively. A review by researchers at the Guttmacher Institute in 2009 listed 37 studies related to the Hyde Amendment. In this article, the focus is on studies by economists that use panel data designs or that exploit a particularly unique experiment. The Journal of Health Economics published two studies of the Hyde Amendment in the same issue in the Winter of 1996. In both the studies, researchers used a panel of states. In one study, authors analyzed abortion rates from 1974 to 1988, whereas in the other researchers used data from 1977 to 1988. Both studies found that the restrictions were associated with a decline in abortion rates of between 3% and 5%. One group of researchers used TSLS to account for the endogeneity of abortion providers; however, the instruments were not convincing. The authors used the natural logarithm of the number of hospitals to predict the natural logarithm of abortion providers and yet many hospitals provided abortions, which undermined the exclusion restriction. The other group of researchers analyzed birth and pregnancy rates in addition to abortion rates. They found that increases in the cost of an abortion lowered birth rates in models that used a 1-year lag in the Medicaid restrictions. Moreover, the decline in births was greater than the fall in abortion rates. The latter finding is hard to reconcile for as it suggests that the decline in births not only offsets the likely rise among some women who carry to term but also induces an even larger group to avoid pregnancy altogether.

Arguably the best 'natural experiment' of the Medicaid financing of abortions occurred in North Carolina (Cook *et al.*, 1999). The State allocated a fixed sum of funds to be used by poor women for abortions as a substitute for resources restricted by the Hyde Amendment. However, between 1978 and 1994, the fund expired five times before the end of the fiscal year in June. The cutoff occurred once in months of December, January, and March and twice in the month of February. The authors found that the cutoff was associated with a fall in abortions and a commensurate rise in births. The effects were greater for Blacks than for Whites and for women with less than 12 years of schooling compared with those with more. Specifically, abortion among Blacks fell 9.5% overall, whereas births rose by 4.7%. In absolute terms, there was a one-to-one correspondence between the fall in abortions and rise in births among Blacks.

The study from North Carolina is particularly convincing. The timing of the funding cutoff varied by year and month and thus would have been difficult for a woman to anticipate.

The authors found no jump in abortions in July as the fund was replenished. The fall in abortions coincided with a rise in births, and effects were greater among groups with higher rates of poverty. The study in North Carolina provides a useful contrast to the previous studies of publicly funded abortions in the US. There is an important trade-off between internal and external validity in these studies, which will be relevant in the discussions that follow. The study in North Carolina has the stronger internal validity, but it pertains to a single state. Nevertheless, the funding cutoff occurred five times, which strengthened the design considerably. However, the panel data studies have the advantage of analyzing changes in 50 states with more than 34 'natural experiments.' However, the number of experiments is misleading. There is limited state variation in the timing of Medicaid funding restraints as the vast majority of restrictions went into effect in 1977 or 1981. Finally, the natural experiment in North Carolina was only able to address short-term changes in abortion and births, whereas the panel studies were able to test for longer term impacts, which may dissipate over time as women adjust to the restrictive funding environment. Despite these caveats, a clear conclusion is that the cutoff of public funding for abortions reduced abortion rates among poor women. The first-order effect should be a rise in births, for which the study in North Carolina provides convincing evidence.

Parental involvement laws

The Supreme Court's decisions in Planned Parenthood of Central Missouri versus Danforth in 1976 and Bellotti versus Baird in 1979 made it constitutional for states to require minors seeking abortions to obtain parental consent or to notify their parents provided that there is an alternative approval mechanism such as a court bypass procedure. Thirty-eight states currently require parental consent or notification of at least one parent or in some instances other adults such as a grandparent or guardians.

Evaluation of PI laws on abortion and births has been hampered by limited data. Ideally, researchers would like age-specific abortion rates by state of residence from 1974 to 2008. These data do not exist. The CDC collects abortions by age for approximately 40 states, but they refer to abortions by state of occurrence. The Guttmacher Institute has used the CDC data to estimate abortions by state of residence, but the Guttmacher researchers acknowledge that their estimates do not take into account travel by subgroups. This becomes a major source of bias in studies of PI laws because resident minors leave the state in response to a PI requirement and nonresident minors stop coming into the state. Abortions by state occurrence will show a substantial drop in abortions to minors when in fact many abortions to minors that would have occurred in the state before the law are performed in other states after the law. This has been demonstrated repeatedly (Cartoff and Klerman, 1986). A second important issue is that researchers have used abortions and birth rates of 18- and 19-year olds as either a counterfactual for changes in birth and abortion rates for minors or as a falsification test. However, the most affected group of minors is 17-year olds. They have the most pregnancies and they are the least willing to involve their parents. Yet, three-quarters of minors who are 17 years of age when they become pregnant will give birth as 18-year olds. As a result, a comparison group of 18-year olds in a DD analysis is contaminated because it includes a large proportion of girls who were exposed to the PI law during pregnancy when they were 17 years of age. Similarly, a falsification test in which the birth rates of 18- or 19-year olds is regressed on a PI law may show little change or even a rise in births. Here too the test is compromised because the 17-year olds who were exposed to the law as minors gave birth when they were 18 years of age.

As with Medicaid financing restrictions, economists have tended to use panel data of state abortion rates to evaluate PI laws. One author reported that PI laws were associated with a 20% fall in the abortion rate of teens of 15–19 years of age. The major limitations were that the author used CDC occurrence data from 1978 to 1990, which fails to account for travel by resident and nonresident minors and the author included 18- and 19-year olds who were unaffected by the law. Another economist used Guttmacher data on teen abortion rates by state of residence for 1985, 1988, 1992, and 1996. He reported a 15% decline in the abortion rate of minors. However, his data do not take into account movement across borders and he only had 4 years of nonconsecutive data. Two economists analyzed data from three states: South Carolina, Tennessee, and Virginia. They found little association with the conditional probability of abortion given pregnancy. They attributed the null finding to travel by minors out-of-state. However, pregnancy resolution as an outcome was uninformative about possible decreases in pregnancy in response to the law. Two other economists analyzed county birth rates from 1973 to 1988. They found that PI laws were associated with a 3% decrease in the birth rate of minors but a 2% decrease in the birth rate of teens of age 18 and 19 years. In absolute terms, however, the fall in the older teen birth rate exceeded that of minors, a result that could be interpreted as a relative rise in the birth rate of minors.

Finally, a study in Texas was able to overcome a number of the empirical challenges that have hampered previous studies (Joyce *et al.*, 2006). First, the authors had data on abortions to residents of Texas. Second, they were able to collect data from the neighboring states as to the number of Texas minors that went out of state after the law. Few minors left Texas because all of the border states except New Mexico enforced a PI law. Third, the authors measured abortions and births by age at conception, which minimized the misclassification bias in previous work. They found that the Texas notification law was associated with a 16% fall in abortion rates among minors who were 17 years and 6–9 months of age at conception and a 4% rise in births. Subsequent work demonstrated that some minors who were almost 18 yeas of age when they conceived waited until they were 18 years of age to abort, even if the delay caused them to terminate substantially later in pregnancy. Finally, they showed that using age at the time of the abortion or birth and ignoring the misclassification resulted in a much larger fall in abortions with no rise in births. This provides some explanation for the findings by other economists who reported no change in births associated with PI laws. In all the other studies authors used age-specific birth rates based on the teen's age at the time of birth and not at conception.

The studies of Texas by Joyce and colleagues are to the PI literature what the study by Cook *et al.* (1999) is to the

literature on Medicaid financed abortions. Both studies have strong internal validity, given the design and quality of data, but both pertain to a single state, which limits their external validity. Studies that use state panels with many law changes would seem superior, but less accurate data on residents and the difficulty of accounting for trends in the outcomes have undermined their internal validity. This trade-off between internal and external validity continues in the studies of mandatory delay and counseling laws as will be shown next.

Mandatory delay and counseling

Many states require a waiting period between the time a woman has been counseled about her abortion and the actual procedure. About 23 states require a mandatory waiting period of 24 h. Utah requires a waiting period of 72, another state 18 h, and one state requires that counseling take place on a day before the abortion but did not specify the length of the waiting period. Four other states had mandatory counseling and waiting period laws whose enforcement had been enjoined. These laws specify that certain information must be given or offered to the women at the initial visit. The required counseling usually includes, among other things, the gestational age of the fetus, information about fetal development, the risks of abortion and childbirth, and resources available for pregnant low-income women. Some mandatory counseling and waiting period laws stipulate or have been interpreted to mean that a woman can be counseled via mail or phone about her procedure; others require that the woman be counseled in person, which usually means she must visit the facility twice – once for counseling and again for the procedure.

The constitutionality of mandatory delay statutes was not confirmed until the 1992 US Supreme Court decision Planned Parenthood of Pennsylvania versus Casey. Thus, there have been relatively few studies and few have found any significant impact of these policies on abortion and birth rates. One problem has been the use of state panels through 1997 or 1998. These studies were statistically underpowered as only a small percentage of women in these panels were exposed to the law. Another reason why these laws have had relatively little impact is because most states allow information to be given over the phone or the internet. This imposes relatively little burden on either the patient or the clinic and would only affect abortions if the required information was persuasive. A recent case-study analysis in Texas found no change in the abortion rate of Texas residents after the state required a 24 h delay and mandated information in January of 2004. The law did not have an inperson requirement as women could obtain the information over the internet (Colman and Joyce, 2011). In contrast, states that require that patients receive the mandated information in person, at least 24 h before the procedure, have demonstrated a greater impact on abortion rates. The burden of an inperson statute is potentially substantial if it necessitates that a woman who lives far from the clinic stay overnight. Mississippi provides such a case. The state imposed a mandatory delay and counseling law with an inperson requirement in August of 1992. Three studies of the law's impact, all using different counterfactuals, found that the law was associated with approximately a 10% decrease in abortion rates, an increase in second trimester abortion rates, and a substantial rise in women leaving the state for an abortion. The key to each study was the quality of the data. Researchers were able to measure abortions to residents of Mississippi obtained in other states. They also had data on the gestational age of the fetus at the time of the termination. However, as with Medicaid financing of abortions and PI laws, the external validity of studies based on a single state is a key limitation.

What conclusion can bedrawn from analyses of state policies in the post-Roe era? The first is that raising the cost of abortion affects behavior. Abortion rates fall, women travel to less restrictive states, and abortions occur later in pregnancy. What is less clear is the magnitude of these changes. The impact of a policy depends on the availability of alternatives. Very poor women may be unable to raise the necessary funds for an abortion. If minors have to travel hundreds of miles to find an abortion provider in a state without a parental notification statute, then they may carry the pregnancy to term. If women must see a physician twice and wait at least 24 h between visits before a procedure can go forward, then her termination is likely to be delayed. Measuring the impact of these policies on births is more challenging. Statistical power is limited. If the birth rate is approximately 3- to 4- times the abortion rates, then even a 10% decrease in abortion would at most result in a 2.5% increase in births. If some women respond to the new law by avoiding pregnancy, the increase will be even less.

The small change in births induced by these policies makes it very difficult to detect changes in health associated with each. The finding from studies report changes in suicide, maltreatment of children, and homicide associated with these laws are implausible. The reduced-form strategy used in many of these studies is vulnerable to omitted variable bias. One researcher, for example, reports that Medicaid restrictions increase suicides among women but mandatory delay laws protect against suicide. Two other economists report an increase of 30–60% in child abuse victims associated with mandatory delay laws. The rationale is that mandatory delay laws result in more unwanted children, but they never show that mandatory delay laws increase birth rates. Another study found that PI laws increase rates of gonorrhea among women less than 20 years of age compared with women 20 years of age and older from 1981 to 1998. However, it has been difficult to show that PI laws had any impact on abortion rates in the 1980s and the early 1990s and so any effect of sexually transmitted diseases is suspect. Moreover, data on sexually transmitted diseases by race are poorly reported in the US. In large racially diverse states, race was unknown in 30–40% of reported cases of gonorrhea.

In the next section the issue of abortion and health will be taken up but with the next generation of studies. The research designs improve. There is more attention to the credibility of the 'first-stage' and the quality of the instruments. The underlying theory can still be traced to the quantity–quality model of household production, but there is less interest in theory and more emphasis on the empirics.

Back to the Future: Roe versus Wade as an Instrument

Advances in research design and insistence on greater rigor in the application of instrumental variables greatly has improved

applied economics since the late 1990s. The literature on abortion and health was similarly affected. Researchers realized that changes in policies regarding Medicaid financing of abortion, PI laws, and mandatory delay statutes did not alter the timing or number of children sufficiently to power analyses of maternal health and child well-being. Thus, researchers returned to abortion legalization in the US and abroad in which there was greater evidence of changes in fertility associated with the more dramatic fall in the price of fertility control. Two papers led the way. In the first, researchers used the legalization of abortion in the US as an instrument for teen childbearing in models of schooling and labor market outcomes. With data from the 1980 Public Use Microdata Samples (PUMS) from the US Census , the authors showed that the longer a teen was exposed to legalized abortion, the lower the likelihood of becoming a teen mother or married before the age of 20 years. The impact of legalization on childbearing was substantially greater among Blacks than among Whites. The racial pattern persisted in the reduced-form models of high school graduation, college attendance, and labor force participation. The authors then used exposure to legalized abortion as an instrument for Black teen out-of-wedlock childbearing in models of school, work, and poverty. They did not pursue a similar analysis for Whites because there was no reduced-form evidence to support it. The results were large. Teen motherhood reduced college entrance by 20 percentage points when estimated by ordinary least squares (OLS) but by 56 percentage points when estimated by TSLS. Differences between OLS and TSLS for labor force participation were even greater. The authors concluded that on balance the data suggested that abortion legalization increased schooling and employment among Black women. Nevertheless, the authors noted that despite the change teen fertility, it was difficult to detect the consequences of teen childbearing even with large samples from the US Census. They go on to encourage researchers to find other sources of exogenous variation in fertility in order to identify the effects of teen childbearing on downstream outcomes.

In the same year, Gruber *et al.* (1999) published an important paper entitled, 'Abortion Legalization and Child Living Circumstances: Who is the Marginal Child?' They too used the 1980 PUMS to analyze changes in the health and well-being of cohorts born before and after the legalization of abortion. Legalized abortion, they argued, changed the distribution of women who gave birth which, in turn, altered the average circumstances under which subsequent cohorts of children were raised. Improved circumstances after Roe would be evidence of positive selection. They also noted that increases in the average circumstance of a cohort implied that the conditions of the marginal child, the one who would have been born had the women not ended the pregnancy, would have to have been worse for average well-being to rise.

The authors estimated both reduced-form and structural models of child well-being using the two phases of abortion legalization in the early 1970s. The reduced form showed that the average change in each outcome was associated with increased access to legalized abortion. In these regressions, the authors found that the rate of low birth weight birth associated with pre-Roe legalization fell from 7.7% to 7.6%, whereas infant mortality dropped from 1.9 per 1000 live births to 1.86 per 1000. The reduced-form results also suggested that children after legalization were less likely to live with a single parent, to live in poverty, or to receive welfare. Effect sizes were approximately 3% of the mean for each outcome. Changes in well-being associated with the marginal child were much larger. The TSLS estimates suggest that the probability of dying in the first year was 40% greater for the marginal child, although the rate of low birth weight was 14% greater. The results by race were less consistent. Although the impact of abortion legalization on the birth rates of non-Whites was twice as large as on Whites, none of the reduced-form estimates of changes in non-White living circumstances or infant health were associated with abortion legalization. The same was true for the marginal child as estimated by TSLS.

In a sequel to the marginal child, the researchers analyzed the impact of abortion legalization on adult outcomes with data from the 2000 census. As before, cohorts pertained to individuals born between 1965 and 1979 and who were 21 to 35 years of age as of the 2000 census. As in Gruber *et al.* (1999), they regressed measures of well-being on the two-phases of legalized abortion in 1970s. The outcomes include the percent in poverty, in single-parent household, on welfare, incarcerated, employed, a high school dropout and a non-college graduate. In only 2 of the 7 outcomes was there an association with early legalization and in only 3 of the outcomes was there any association with all phases of legalization. In the TSLS models in which each outcome was regressed on the birth rate instrumented by the cost of abortion, less than half the outcomes were associated with worse conditions for the marginal child.

The 'marginal child' papers provided a novel and more general empirical framework for estimating the impact of abortion legalization on the child that was not born. Instead, of only associating abortion legalization with average changes in affected cohorts, these authors provided a clever method of estimating the counterfactual outcome. There are, however, important limitations to the empirical work and results. First, in both papers, the authors could not separate age from period effects because they only had data on each outcome at a single point in time. The inclusion of state-specific quadratics in age may have accounted for some of the variation in period effects, but period effects can be very powerful determinants of crime, employment, single parenthood, etc. Second, a lack of selection effects among non-Whites is difficult to explain, especially in light of other work that demonstrated robust effects of abortion legalization on education and employment among Black women. Not only did the legalization of abortion affect non-White fertility more than Whites, but also the non-Whites are more likely to be incarcerated, on welfare, single parents, and high school dropouts. If abortion is improving the circumstances of White children, indicative of positive selection, why would an even greater relative and absolute decrease in fertility among non-Whites not affect their circumstances even more? Either there is negative selection among non-Whites or unmeasured period effects are confounding estimates. Third, it is difficult to interpret the first-stage estimates in this study. There are many interactions in which the omitted category is obscure and the exclusion restrictions are hard to justify. Despite these issues, the marginal child papers were an important advance in the literature.

Abortion and crime

Clearly, the most sensational association with abortion came from Donohue and Levitt's (2001) paper linking the legalization of abortion to the precipitous drop in crime. The mechanism was not novel. Citing Grossman and Jacobowitz (1981) and Gruber *et al.* (1999), Donohue and Levitt (2001) argued that the child who was not born would have grown up in worse living circumstances, received less parental support, and as a result would have been more prone to criminal behavior as a teen and adult. The paper received remarkable attention in the popular press and its basic finding reached an even broader audience with the publication of Levitt and Dubner (2005) book, Freakonomics. The empirics were simple. The authors regressed total crime rates on lags of the abortion rate adjusted for state and year-fixed effects. They also regressed age-specific arrest rates for those of 15–24 years of age on the lagged abortion rate. In both specifications they found that abortion rates could explain upward of 50% of the decrease in crime in 1990s.

The results were quickly challenged. It was straightforward to show that their story did not line up with simple plots of age-specific homicide rates (Joyce, 2009). For instance, homicide rates soared between 1985 and 1992 among young, African-American males in large urban areas and then dropped almost as precipitously thereafter. There were relatively modest changes in murder rates among other groups who were also exposed to legalized abortion *in utero*. Most criminologists attributed the increases in homicides to the crack cocaine epidemic which spurred a rise in gang violence. However, no credible data on crack-cocaine use by state, year, and age existed which created a potentially significant omitted variable problem. This was aptly demonstrated by two economists who first replicated Donohue and Levitt's regressions but then added state–year interactions. The association with the abortion rate fell by 50–60%. Another economist used a triple difference strategy to eliminate the confounding effect of crack cocaine by comparing the crime rates of 19-year olds born before abortion was legalized to that of 17-year olds born just after. Both groups experienced the same period effects (i.e., the crack-cocaine epidemic) but only the younger cohort was exposed to legalized abortion *in utero* (Joyce, 2009). Joyce found no association between legalized abortion and crime. A full airing of the debate is beyond the scope of this article. Regardless of the ultimate judgment of the Donohue and Levitt thesis, their work stimulated further research. Economists examined the association between legalized abortion and drug use, whereas others correlated legalized abortion with teen pregnancy, a female proxy for delinquent behavior. Economists also convincingly linked legalized abortion to sexually transmitted diseases. The strength of these papers rested on use of abortion legalization as the identifying source of variation. Legalization, much more than subsequent policies regulating abortion, had a clear, measurable impact on fertility. And yet the challenge in all these papers is identification of a cohort effect amidst often powerful period effects. In the case of abortion and crime, it was the crack epidemic of the late 1980s and the early 1990s that confounded estimates. With teen pregnancy, it was welfare reform and the expanding economy in the 1990s. Thus, studies that analyzed changes in outcomes around the time of legalization are more convincing because the confounding from period effects is arguably more easily controlled. Even with more proximate outcomes, the health effects of abortion are exceedingly difficult to identify. Recall that Gruber *et al.* (1999) found exceedingly modest declines in low birth weight and infant mortality among cohorts exposed versus unexposed to legalized abortion. In fact, more recent research suggests that the 1–2% declines in their paper are probably too small to be detected with the proper adjustment of the standard errors.

One paper illustrates just how difficult it can be to associate even dramatic changes in the cost of fertility control with well-being (Pop-Eleches, 2006). In December of 1966, Romania outlawed abortion and all methods of fertility control in response to the declining birth rate in the country. The result was an immediate doubling of the birth rate from 14.3 births per 1000 population to 27.4 a year later. The author used this unprecedented fertility shock to estimate its impact on the educational and labor market outcomes of the birth cohorts born just before and after the ban. The overall result was an increase in well-being, a result directly at odds with the US experience. The seemingly contradictory finding resulted from the positive increase in childbearing among families of higher socioeconomic status. Once the author adjusted for the composition change, exposure to the ban was associated with decreased schooling. The author interpreted the latter effect as the negative impact of unwantedness. The author found no association with labor market outcomes. The author also reported a 27% increase in infant mortality and a 30% increase in low birth weight. The changes in infant health were relatively short lived and thus may have been caused in part by lack of prenatal and obstetric services.

The increase in fertility in Romania was 20 times the decrease observed with abortion legalization in US and yet, even with such a huge jump in the birth rate, changes in well-being were somewhat modest or relatively short lived. This underscores the point made previously: detecting cohort effects on downstream outcomes is extremely challenging. Without large, exogenous shocks, distinguishing cohort from age and period effects may exceed researchers' ability to detect them with extant data.

Summary

The Romanian study provides an appropriate bookend to the work of Grossman and Jacobowitz (1981). The 25-year interval saw a large body of research devoted to identifying an empirical link between abortion and well-being. A tentative conclusion would argue for a positive association between the availability of legalized abortion services and increases in the health and well-being of the exposed cohorts. But even this modest assessment comes with many caveats. The early cross-sectional estimates must be discounted because the potential for confounding is overwhelming. Reduced-form estimates based on panel data that exploit change in policies such as parental involvement laws or Medicaid financing restrictions lack a sufficiently robust first stage to identify effects on health. The return to the early years of abortion legalization improved the first stage, but even then, statistically significant findings

were not consistent and the most sensational estimates with respect to homicide have been largely discredited. Thus, the author ends with the Romania study for it provided the outsized experiment so valued in applied microeconometrics. But even in this case, the association between large changes in fertility and more schooling among the affected cohorts was modest. This suggests that long-term effects of changes in the cost of fertility control on the well-being of affected cohorts may well exist, but effects are probably too small and data too imprecise to identify them econometrically.

See also: Fertility and Population in Developing Countries. Health Care Demand, Empirical Determinants of. Instrumental Variables: Informing Policy. Observational Studies in Economic Evaluation. Panel Data and Difference-in-Differences Estimation

References

Akerlof, G., Yellen, J. and Katz, M. (1996). An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* **111**(2), 277–317.

Becker, G. S. and Lewis, H. G. (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy* **81**, S279–S288.

Brown, S. S. and Eisenberg, L. (1995). *The best intentions: Unintended pregnancy and the well-being of children and families*. Washington, DC: National Academy Press.

Cartoff, V. G. and Klerman, L. V. (1986). Parental consent for abortion: Impact of the Massachusetts law. *American Journal of Public Health* **76**(4), 397–400.

Colman, S. and Joyce, T. (2011). Regulating abortion: Impact on patients and providers in Texas. *Journal of Policy Analysis and Management* **30**(4), 775–797.

Cook, P. J., Parnell, A. M., Moore, M. J. and Pagnini, D. (1999). The effects of short-term variation in abortion funding on pregnancy outcomes. *Journal of Health Economics* **18**(2), 241–257.

Donohue, J. and Levitt, S. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics* **116**(2), 379–420.

Garrow, D. J. (1998). *Liberty and sexuality: The right to privacy and the making of Roe V Wade*. Berkeley, CA: University of California Press.

Grossman, M. and Jacobowitz, S. (1981). Variations in infant mortality rates among counties of the United States: The roles of public policies and programs. *Demography* **18**(4), 695–713.

Grossman, M. and Joyce, T. (1990). Unobservables, pregnancy resolutions, and birthweight production functions in New York City. *Journal of Political Economy* **98**, 983–1007.

Gruber, J., Levine, P. and Staiger, D. (1999). Legalized abortion and child living circumstances: Who is the marginal child. *Quarterly Journal of Economics* **114**(1), 263–291.

Joyce, T. (2009). A simple test of abortion and crime. *Review of Economics and Statistics* **91**(1), 112–123.

Joyce, T., Kaestner, R. and Colman, S. (2006). Changes in abortions and births following Texas's Parental Notification Law. *New England Journal of Medicine* **354**(10), 1031–1038.

Lader, L. (1974). *Abortion II: Making the revolution*. New York: Beacon.

Levine, P. B., Staiger, D., Kane, T. J. and Zimmerman, D. J. (1999). Roe v. Wade and American fertility. *American Journal of Public Health* **89**(2), 199–203.

Levitt, S. and Dubner, S. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: Harper Collins.

Lundberg, S. and Plotnick, R. D. (1990). Effects of state welfare, abortion and family planning policies on premarital childbearing among white adolescents. *Family Planning Perspectives* **22**(6), 246–275.

Pop-Eleches, C. (2006). The impact of an abortion ban on socioeconomic outcomes of children: Evidence from Romania. *Journal of Political Economy* **114**(4), 744–773.

Relevant Websites

http://www.abortion.com/
Abortion.

http://www.cdc.gov/reproductivehealth/data_stats/Abortion.htm
Centers for Disease Control and Prevention.

www.guttmacher.org
Guttmacher Institute.

http://www.naral.org/
NARAL pro-Choice America.

Access and Health Insurance

M Grignon, McMaster University, Hamilton, ON, Canada

Abbreviations

ACSC	Ambulatory care sensitive condition.
HIE	Health insurance experiment.
TANF	Temporary assistance for needy families (welfare scheme in the US).

Glossary

Adverse events Negative outcomes of treatments, such as death or rehospitalization.
Ambulatory care Care provided outside hospitals to patients who are not bedridden and live in the community.
Attrition Reduction in number of a sample of respondents to a repeated survey (from initial survey year to subsequent ones).
Catastrophic care Care that is needed to prevent death or extreme disability.
Cost sharing, copayments, and coinsurance These three terms are used interchangeably in this article, to mean a payment made at the point of use by the patient that is not reimbursed by any health insurance.
Exogenous A variable is exogenous if it is not a function of other parameters or variables in the model.
Income effect Change in consumption of a good as a result of a change real income.
Instruments or instrumental variables Variables used as proxy of factors which are suspected of being not entirely exogenous. The instrument correlates with the factor but its influence on the dependent variable is exogenous.
Longitudinal studies Studies in which the same individual subject is observed repeatedly over time.
Marginal value The maximum value attached to a little more or less of a good, service or desired characteristic.
Moral hazard Moral hazard refers to the possibility that insured individuals will behave in such a way after an insured event has occurred that will increase the claim cost to insurers, partly because the user-price of care is lower through insurance and demand may therefore rise.
Out-of-pocket Amount of money spent directly by a patient at the point of use and is not reimbursed by insurance (see cost sharing: what is not covered by any insurance plan).
Social epidemiologists Social epidemiologists are interested in the social determinants of the distribution of health in a population.
Social experiment A field experiment (not in the laboratory) to answer an economic or social policy questions.
Subsidy Part of the price of a service that is covered by an insurance plan or a public agency.

Introduction

It is evident that lack of (or poor) insurance coverage is a barrier to access healthcare. Evidence that insurance status is linked to access to healthcare seems overwhelming: those with insurance always use substantially more than those without.

Economists tend to be more skeptical, for the following two reasons: they question the causality behind the observed link between coverage and utilization; they question the inference from differences in utilization to differences in access to care. The causality issue is currently not important and it is summarized briefly in Section Health Insurance Increases Utilization. The distinction between utilization and access is currently a matter of scientific investigation among economists and social epidemiologists, and this review of the literature will mostly focus on this issue. Section Interpreting the Causal Effect of Insurance: Moral Hazard or Access? summarizes the theoretical debate on the inference question, which can be described as follows: Is the difference in utilization resulting from insurance coverage a matter of moral hazard – the insured use more than they need – or access – the uninsured do not use what they need? It is shown that the empirical answer depends on how healthcare need is defined and measured. Sections Effect of Insurance on the Subjective Assessment of Unmet Need by Survey Respondents, Insurance and Utilization of Medically Necessary Care, and Effect of Insurance on Health Outcomes: Adverse Events and General Health and Mortality then review the empirical evidence on the impact of insurance on the utilization of care that is needed, using three different definitions of need. In Section Effect of Insurance on the Subjective Assessment of Unmet Need by Survey Respondents, a subjective definition (what is perceived as unmet need) is used; in Section Insurance and Utilization of Medically Necessary Care, a more objective definition of need as what is clinically recommended to survive or maintain good health is used; last, in Section Effect of Insurance on Health Outcomes: Adverse Events and General

Health and Mortality, an outcome-oriented definition of need and evidence on the effect of lack of coverage on mortality and health status is used. Section Policy Implications concludes and draws policy recommendations.

Health Insurance Increases Utilization

The causality issue is as follows: When we observe differences across insurance it is noticed that individuals are not assigned to a given health insurance status but they make their own decisions on whether to be insured or not. Of course, these decisions are constrained, by how much individuals can spend overall compared to the price of health insurance, but, nevertheless, individuals at the same level of income and faced with the same premiums make different decisions regarding coverage (Bundorf and Pauly, 2006). If that decision is somehow linked to their utilization of healthcare services in a way that is not observed (in the survey used by the analyst), the correlation between insurance status and utilization may be spurious and it would be wrong to infer causality from it. For example, if individuals were to buy health insurance only because they wanted to commit to visit a doctor once a year, and get their tension and cholesterol checked, the correlation between insurance status and utilization of these services would be perfect. However, that would not mean that covering the uninsured would change their behavior: if the reason why they do not buy insurance is as they do not value the services it covers, they then might not be interested even if the services were free of charge at the point of use.

One way to address the issue is to run a social experiment: the health insurance experiment (HIE), conducted by the RAND Corporation randomly assigned approximately 2000 households to a variety of plans with varying cost-sharing arrangements (Newhouse and the Insurance Experiment Group, 1993). Because individuals were assigned to the plans rather than choosing them, any difference in utilization can be safely interpreted as causal. The results from that social experiment indicate a clear causality from coverage to utilization: individuals assigned to plans with lower copayments used more outpatient services, prescription drugs, and even inpatient services. The latter finding has been recently disputed by Nyman (2007), who argued that it is an artifact because of attrition (those who are poorly covered through the experiment and need hospital care quit the experiment and revert to their former plan); Newhouse et al. (2008) responded that subjects have no incentive doing that because they are more than compensated for the loss if (and only if) they stay in the experiment. It is true that the attrition rate was much higher in the higher coinsurance plan than in the free plan but it remains undecided whether subjects left the experiment (although they had no interest doing it) when in need of hospital care and not well covered (Nyman's suggestion) or whether they left for other reasons (the HIE Group's response to Nyman).

Beside social experiments, which are costly and constrained by ethical issues (it is not feasible to assign subjects to no coverage at all and some stop loss must be put in place, which does not allow the researchers to test the effect of not being insured), economists use a variety of econometric strategies to test causal inference in observational studies and all find a causal link from insurance status to utilization pattern.

Interpreting the Causal Effect of Insurance: Moral Hazard or Access?

It is evident that coverage influences utilization and it can be said that not being insured causes lower levels of utilization of healthcare services. The remaining issue is one of interpretation: Do the uninsured use less because they cannot afford the services when they are ill? Or do the uninsured buy exactly the amount of healthcare they need, whereas the insured overconsume healthcare because they do not have to pay for it at the point of use? Or is it that both interpretations are partially true: Some among the insured 'overconsume' and some among the uninsured cannot access the care they need. To understand the issues underlying the difference in interpretations we need to go back to the economic theory of health insurance and introduce concepts such as moral hazard. As will be clear at the end of this section, a key concept for the understanding of the access versus moral hazard controversy is the concept of need: if we could tell what is needed and what is a matter of preference in healthcare services utilization, we could tell which part of the variation in utilization across insurance status is a problem of access for the uninsured and which is moral hazard of the insured.

Andersen (1995), and most social epidemiologists, equated access to utilization: if one uses fewer services it is because they cannot use as much. He distinguishes between 'potential access' (enabling factors such as availability of services, coverage, regular source of care, travel costs, and waiting time) and 'realized access' (actual utilization). But the economists disagree on the proposed theory. As noted by Hurley (2000), access is a process-oriented concept and is unrelated to actual use: the difference between such a conception and Andersen's is that, for a given level of accessibility, individuals with different preferences make different choices. For most economists, access is similar to 'opportunity,' and individuals are always free to use opportunities as they see fit. Some of the difference between the insured and the uninsured is a matter of access (the medical need of the uninsured is not met), and some is a matter of want (the insured use nonneeded care).

The objective is of course to evaluate the respective roles of access and want in the difference in utilization across insurance status. To do so, one needs to understand the way health insurance works and interferes with decisions made by individuals regarding their utilization of healthcare services. The following is drawn from Nyman (2003).

Although standard (nonhealth) insurance pays a lump sum in case a detrimental event occurs (life insurance pays a given sum in case the insured dies), health insurance typically pays back through reduced prices of healthcare. Being covered by health insurance, therefore, means gaining access to discounted healthcare services. Some plans have a limit on reimbursement, but most public plans do not set such limits on reimbursements for acute care (hospitalizations, visits to a family doctor, and drugs prescribed by a doctor).

As a result, insured individuals live 'in a different world' than the uninsured, a world with lower prices of healthcare services. Proponents of the moral hazard hypothesis posit that because the uninsured are faced with the true price of healthcare, they buy units of healthcare services until they reach a level at which the marginal value of an extra unit is less than the price they have to pay. The insured do the same, but because they face a lower price of healthcare they buy more than what would satisfy them (to be exact, what would maximize their satisfaction) if they were not insured. The analysis of health insurance is similar to the analysis of subsidies for specific goods (e.g., food): when a price is artificially lowered, individuals do not get the right information about the relative values of goods and favor the subsidized one to the detriment of nonsubsidized goods.

The economic theory of health insurance is not only about this substitution effect but also involves what economists call income effects: If we compare two individuals with the same level of income, one benefiting from a discount on the price of one specific good but not the other one, it is clear that the former has more purchasing power than the latter. In that sense, they are richer and can make the decision to allocate that extra purchasing power as they see fit. If they decide to buy more healthcare services, because they are sick and made richer by their health insurance coverage, they are not substituting away from other potential uses of their money. They make a rational decision to allocate their extra purchasing power where it is needed.

The moral hazard story goes as follows: "Being insured means I will take advantage of lower prices of healthcare to use more of them, whether I am sick or not, need it or not. It is the fact that they are cheaper than if I was uninsured that motivates me the most."

The income transfer story is as follows: "Being insured means that when sick and in need of care, I will be richer than if I was uninsured. I will then spend more on healthcare because this is what I need to do (I am sick) and I can afford it. It shows clearly that the 'income effect' is the translation in economic theory of the access problem of social epidemiology."

It is of course impossible to separate these two mechanisms empirically on the basis of the difference in utilization across insurance status: they both predict the exact same difference in utilization.

The only notion being observed that would allow to separate the two mechanisms is 'need': Recall that the income effect occurs because the insured benefits from an income transfer when sick, whereas the substitution effect is independent of health states. One useful way to look at access versus moral hazard would, therefore, be to look at the differential effect of coverage on care that is 'needed' versus care that one could go without.

So far, we have only moved the question one step further and still need to define what 'needed' means in healthcare. As shown by Culyer (1998) and the literature on equity in healthcare utilization, need is an elusive concept, and it is impossible to provide a theoretical definition of need that would satisfy most. Rather, need is defined as how it is measured in empirical studies.

How do we measure need? Here, three ways of defining needed care are suggested:

- Subjective: Do they feel they could not access care they needed?
- Objective, process-oriented: Needed care is the type of care that is clinically necessary to maintain health.
- Objective, outcome-oriented: Access barrier can be inferred from lower utilization if and only if lack of coverage causes poorer health outcomes.

These questions were investigated in the RAND HIE: the objective was not only to measure the causal link between coverage and utilization but also to describe which services were underused by the less well covered (or overused by the better covered) and to measure the impact of being less well covered on health (a 2–4 years follow-up was included in the experiment). It is very often stated that the RAND shows a strong difference in utilization as a result of differences in coverage but no difference at all in health outcomes. Some use that often stated conclusion to infer that 100% of the difference in utilization is because of moral hazard and nothing to access problems. Interestingly, this is not the interpretation of the HIE group members themselves: first, they show that the insured utilize more of both clinically recommended and futile care than the uninsured, implying that the difference is due in part to both access problems and moral hazard. Second, they observe that in some groups (the poor and the sick) being less well covered has consequences on health. However, the effect is offset on average because the better covered also seem to suffer (surprisingly) from 'too much healthcare.' The combination of these two effects is the often cited 'no effect on health' but the RAND experiment itself does not conclude to the absence of a link between being less well covered and deteriorating health. In a sense, there must be an effect because one of the result of the RAND is that those in the plans with higher copayments used less inpatient care, and it is hard to imagine that the better covered would be admitted to a hospital to receive treatments with absolutely no effect on their health, simply for the sake of staying in a hospital.

Effect of Insurance on the Subjective Assessment of Unmet Need by Survey Respondents

A simple way to assess needed care is to directly ask respondents of a survey to state whether they had to forgo care they needed in the recent past (typically 12 months). The price to pay for such simplicity is the subjective component of the perception of need: if subjective perceptions of need correlate in a systematic way with decisions not to buy insurance, the value of such subjective assessment is low. Also, it must be noted that unmet needed care can be the result of many factors beyond lack of insurance (lack of time, procrastination, and fear).

An idea to test a causal link between coverage and perception of unmet need that should not be affected by systematic variations in how subjective need is defined is to take advantage of exogenous changes in health insurance coverage.

One such shock is the 1996 Reform of Welfare in the US that led to reductions in the caseload of the temporary assistance for needy families (TANF). Women who lost TANF also lost public health insurance after 12 months and follow ups show substantial increases in self-reported unmet need for a variety of healthcare services.

Insurance and Utilization of Medically Necessary Care

To overcome the subjectivity of self-reported unmet need, we can define needed care as what is necessary to maintain health. A stringent definition is that care is needed if and only if not receiving it would lead to death or severe disability, and the evidence on the causal effect of coverage on utilization of such care is reviewed (see Section Care That is Needed in Life-Threatening Situations or When Quality of Life Would Be Greatly Affected without Treatment). A more lenient definition is that care is needed as long as clinical consensus is that not receiving that type of care would affect intermediary health outcomes and the evidence based on that clinical definition of need is reviewed in Section Differences in Utilization of Clinically Recommended Care.

Care That is Needed in Life-Threatening Situations or When Quality of Life Would Be Greatly Affected without Treatment

A first approach is to describe what individuals facing a health shock (an illness or injury necessitating treatment if the patient wants to recover) do when they are not covered. Most of the literature on insurance and the economic consequences of health shocks is recent and from low- and middle-income countries; the literature on health shocks in rich countries is mostly about health and labor supply, and the case of the uninsured is less often considered because in most rich countries, to the possible exception of the US, public insurance covers potentially catastrophic health shocks.

In low- and middle-income countries (Ethiopia, Vietnam, and Laos), the uninsured pay for medical care in case of health shocks necessitating catastrophic spending through informal insurance mechanisms (microfinance schemes, informal lending, or transfers), drawing from their assets and savings, or cutting back on other consumption items. The only exceptions seem to be China, where the uninsured spend less out-of-pocket than the insured in case of health shocks, and Thailand, where the poor who need treatment for end-stage renal disease use therapeutic strategies or less frequent dialysis, which have side effects but keep them alive.

In the US, bankruptcy can be used to protect assets in case of large medical bills. Approximately 1 million households filed for bankruptcy caused by medical bills in excess of US$1000 in the US in 2001. Bankruptcy is not enough, though, and 61% of them also had to cut back healthcare.

In the US, as in China, the uninsured spend less out of pocket than the insured in case of a severe health shock, suggesting that lack of insurance makes medical services less affordable and, therefore, reduces access. Of course, another way to spend less out of pocket is to receive care free of charge, through charity. It is documented that public and not-for-profit hospitals in the US deliver care free of charge to patients unable to pay for care in cases of severe illnesses and accidents.

A less stringent definition of health shocks is 'nonavoidable hospitalizations.' These are hospitalizations that cannot be avoided by effective, timely, and continuous outpatient (ambulatory) medical care for certain chronic conditions – they are also called admissions for non-ambulatory care sensitive conditions (non-ACSCs). Among adults, their necessary character can be disputed: for instance, a cataract excision is a non-ACSC (no primary care can really prevent cataract), can be 'necessary' in some cases (to cure near blindness) but can also be discretionary in other cases (when vision quality is diminished); similarly, a hip replacement can be needed (the patient cannot walk without it) or discretionary (the patient can walk but feels some pain or discomfort). However, in the case of children (younger than 15 years of age), it can be argued that what is not preventable is more likely to be needed to prevent future health problems.

A study of non-ACSC pediatric admissions from 1983 to 1996 based on the US National Hospital Discharge Survey uses exogenous expansions of the Medicaid program between 1983 and 1996 (increase in children population covered by 16% points overall but at different times in different states) to estimate a causal link, rather than a simple correlation, between Medicaid coverage and use of hospital care for non-ACSC. If utilization of non-ACSC hospitalizations increases with enrollment in Medicaid, this is an indication of a causal link between lack of coverage and difficulties to access needed care. They find that Medicaid expansions led to an increase in non-ACSC admissions: any increase in enrollment by 1% increases the probability of admission for a non-ACSC by 0.81%. Therefore, there was an access problem to inpatient care for children without insurance before the expansion. When admitted, these newly covered children also receive more procedures than when they were not covered.

Similarly, the implementation of a universal National Health Insurance for the elderly in Taiwan had a stronger effect on low- and middle-income elderly than on high-income elderly individuals, suggesting that there was an access problem linked to ability to pay for treatment without insurance.

Differences in Utilization of Clinically Recommended Care

Although ambulatory care services are less expensive, some authors consider that they are 'needed' when proven to be effective, in the sense that not using them negatively affects health. As a result, if the uninsured can be shown to use less preventive services than the insured that could be interpreted as a problem of access to care. What is known on the causal effect of insurance on the utilization of clinically recommended services (such as mammography) or intermediary clinical outcomes (such as blood pressure) is now reviewed.

Changes in insurance status in longitudinal studies identify both a causal effect of copayments on mammography and a causal effect of loss of coverage on postemergency room visit to an ambulatory care doctor in the US.

Levy and Meltzer (2001, 2004, 2008) reviewed studies on insurance and intermediary health outcomes. Studies testing for a causal link are selected. Some of the studies reviewed in Levy and Meltzer will be reviewed in Section Effect of Insurance on Health Outcomes: Adverse Events and General Health and Mortality (those on final outcomes such as mortality, self-assessed health, or functional ability). They show a clear effect of loss of coverage on blood pressure, but some of these studies cannot conclude at any substantial effect of coverage on intermediary health outcomes.

Effect of Insurance on Health Outcomes: Adverse Events and General Health and Mortality

Introduction of copayments in public schemes (medication insurance for the elderly and welfare recipients in Quebec in 1996 or the California Public Employees Retirement System (CalPERS) in 2001) reduces utilization and substantially increases the probability of adverse events (more than double in Quebec).

Moving to studies testing the effect of insurance on mortality and general health; most studies do not measure the effect of insurance on utilization and infer access problems directly from detrimental effect of lack of insurance on health outcomes. Historical data (European countries in 1870–1914) show that an increase of 10% points in the proportion of population covered by health insurance led to a reduction in mortality by 0.9–1.6 per 1000. The 1.6 effect is certainly implausibly high but it should be kept in mind that expansions of coverage were usually targeted at individuals toward the lower end of the income distribution, where mortality was very high and at a time when their income did not allow them any contact with a doctor. As a result, these estimates are of effects at the maximum rate of return of coverage on access and of access on health. On the contrary, the introduction of Medicare in 1965 had no discernible effect on the change in mortality around 1970: Regions in the US with lower rates of insurance after the age of 65 years did not see any more substantial decrease in their mortality than regions with higher rates (which were less affected by Medicare as a result). The fact that Canadian provinces did not implement universal coverage at the same time (between 1962 and 1972) can be used to identify a significant effect of universal coverage: a reduction of 4% in infant mortality and of 1.3% in low birth weight.

Another approach uses the exogenous discontinuity in insurance status for most Americans when they turn 65 years: There is indeed a decrease in the mortality rate (compared to the trend before the age of 65 years) of approximately 13%, but it is hard to attribute it entirely to Medicare (Americans tend to retire at the age of 65 years as well, which can be good for health). Moreover, the effect does not vary at all across race and location or self-employed status although insurance status pre-Medicare varies substantially across these variables. A randomized trial in Oregon studies the effect of getting coverage on health outcomes (30 000 low-income individuals were randomly selected to benefit from Medicaid coverage and 10 000 applied – these are compared to similar individuals on the waiting list who were not selected) and finds an effect on self-assessed health at 1 year follow-up. The data are still under analysis and more should be known soon about objective measures such as blood pressure.

Studies using instruments (variables that affect health through insurance but are not subject to the endogeneity issue of insurance status, such as spouse's union status, immigration status, and number of years in the US, work loss in the previous 5 years, or state-level unionization rates or Medicaid eligibility and generosity of benefits) find large and significant effects of insurance on health (self-assessed health, general mortality, and human immunodeficiency virus-acquired immunodeficiency syndrome-related mortality), but the quality of the instruments can be discussed.

One particular relationship has been studied in more detail and remains disputed in the empirical literature: the effect of insurance on infant and children health. The effect of expansions of health insurance for pregnant women, infants, and children in the 1980s (1979–92) in the US on birth outcomes and children health is estimated as strong and negative on mortality (expansions yielded a decrease in mortality by almost 40%) by Currie and Gruber (1996). However, Dave et al. (2008) rightly pointed out that this is implausibly high. Their objection is that the quasiexperiment is not methodologically sound: if some unobserved variable explains that states where efforts on public maternal and natal health were made also were those states where Medicaid expansions took off first, using eligibility by year and state will overestimate the effect of insurance on mortality).

The study of expansions in insurance for infant and pregnant women finds a weak effect on birth weight (likely because of crowding out: overall, the expansions led to only a 10% points increase in the proportion insured) but a substantial effect on infant mortality (expansions decreased it by 8.5%). Last, the study of expansions in insurance for pregnant women on infant mortality found that the effect was strong for infants whose mother lived closest to a hitech hospital. It is also found that better educated women (not dropouts or teen mothers) actually used less hitech care (notably caesarian section) after the expansions, likely because of the fact that they switched from a private insurance to Medicaid, but without any notable effect on their infant's health (a case of futile care because of private insurance and generous coverage).

Overall, the lack of insurance increases the probability of adverse events and is the cause of poorer self-assessed health and higher infant mortality. Its effect on adult mortality and low birth weight is less clearly documented.

Policy Implications

It can be safely concluded that access problems are part of the difference in utilization across insurance status: It is not only about moral hazard and the difference also stems from the fact that the uninsured do not benefit from an income transfer when sick and, as a result, cannot access needed medical care. They access charity care if the intervention is a matter of survival or to prevent disability, delay recommended care such as follow-up after emergency admission or ambulatory care after new symptoms of a chronic condition,

and are much less likely to be screened for cancer, or have their blood pressure or cholesterol measured. As a result, not being insured has consequences on health, documented by downstream adverse events, self-assessed health, and, possibly, longevity.

From a normative perspective, this means that some of the difference in utilization between the insured and the uninsured is welcome: it does not mean that the insured spend 'too much' on care because they are overcovered but that the sick who are insured can use the income transfer they receive from the healthy to access needed care. This review shows that inpatient services that are nonavoidable and expensive should enter a universal plan; it also shows that preventive services and ambulatory care services that meet clinical recommendations should also be covered for the less well-off, who cannot afford it if not covered. There is no literature that would allow us to determine what is not affordable if not covered. It has been suggested in some countries with public insurance that affordability should be the main criterion for coverage: for instance, the 'bouclier sanitaire' (health shield) discussed in France in 2007–08 was a project to replace the current universal public plan with various copayments for various services (and exemptions for the chronically ill) with a full coverage plan with a deductible set at 10% of income. In Ontario, Canada, the same idea forms the basis of a tax deduction for those who have to spend more than a share of their taxable income on prescription drugs out of the pocket. The issues with such attempts at solving moral hazard (through deductible) and access (universal coverage and no copayment beyond the deductible) are threefold: first, there is no clear definition of affordability and the 10% threshold is rather arbitrary (Glied 2008); letting physicians determine what is needed without imposing any cost sharing on patients seems to be a more promising avenue to solve moral hazard and access simultaneously (the difficulty being to provide doctors with the right incentives to deliver services that are needed only). Second, the chronically ill with low or middle income will reach the deductible every year and will be penalized for being chronically ill though they are not at fault. In the US, this would prove a progress compared to the situation before the latest reform (where preexisting condition were a cause for exclusion of coverage), but in most European countries and Canada that would be a regression. Third, the deductible set at 10% of income would not address the issue of access to preventive care: in the case of preventive care, the issue does not seem to be that individuals cannot pay for it (except the very poor), but rather that the benefits (positive effect on health) accrue in the future, whereas the cost is borne immediately.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Demand for Insurance That Nudges Demand. Health and Health Care, Need for. Health Care Demand, Empirical Determinants of. Health Insurance and Health. Managed Care. Moral Hazard. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Value-Based Insurance Design

Reference

Andersen, R. M. (1995). Revisiting the behavioral model and access to medical care: Does it matter? *Journal of Health and Social Behavior* **36**(1), 1–10.

Bundorf, M. K. and Pauly, M. V. (2006). Is health insurance affordable for the uninsured? *Journal of Health Economics* **25**(4), 650–673.

Culyer, A. J. (1998). Need – Is a consensus possible? *Journal of Medical Ethics* **24**, 77–80.

Currie, J. and Gruber, J. (1996). Health insurance eligibility, utilization of medical care, and child health. *Quarterly Journal of Economics* **111**(2), 431–466.

Dave, D. M., Decker, S., Kaestner, R. and Simon, K. I. (2008). Re-examining the effects of Medicaid expansions for pregnant women. *NBER Working Paper 14591*. Cambridge, MA: National Bureau of Economic Research.

Glied, S. (2008). Universal public health insurance and private coverage: Externalities in health care consumption. *Canadian Public Policy* **34**(3), 345–357.

Hurley, J. (2000). An overview of the normative economics of the health sector. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1, Part 1, ch. 2, pp. 55–118. North Holland: Elsevier.

Levy, H. and Meltzer, D. (2001). What do we really know about whether health insurance affects health? University of Chicago School of Public Health, Unpublished document, December.

Levy, H. and Meltzer, D. (2004). What do we really know about whether health insurance affects health? In McLaughlin, C. (ed.) *Health policy and the uninsured: setting the agenda*, pp. 179–204. Washington, DC: Urban Institute Press.

Levy, H. and Meltzer, D. (2008). The impact of health insurance on health. *Annual Review of Public Health* **29**, 399–409.

Newhouse, J. P., Brook, R. H., Duan, N., et al. (2008). Commentary: Attrition in the RAND health insurance experiment: A response to nyman. *Journal of Health Politics, Policy and Law* **33**(2), 295–308.

Newhouse, J. P. and the Insurance Experiment Group (1993). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge: Harvard University Press.

Nyman, J. A. (2003). *The theory of demand for health insurance*. Stanford, CA: Stanford University Press.

Nyman, J. A. (2007). American health policy: Cracks in the foundation. *Journal of Health Politics, Policy and Law* **32**(5), 759–783.

Further Reading

Culyer, A. J. and Wagstaff, A. (1993). Equity and equality in health and health care. *Journal of Health Economics* **12**, 431–457.

Addiction

MC Auld, University of Victoria, Victoria, BC, Canada
JA Matheson, University of Leicester, Leicester, England, UK

Glossary

Dynamic rationality A decision process such that a plan made in the present for a future period is consistent in the sense that the plan remains optimal when the future period arrives.

Exponential discounting Discounting future costs or benefits through a process in which the rate of time preference does not depend on the time interval between the moment of choice and the actual events.

External cost An involuntary cost that is imposed on a third party. For example, second-hand smoke from cigarettes, or traffic accidents resulting from alcohol-impaired driving.

Hyperbolic discounting Discounting future costs or benefits through a process in which the rate of time preference depends on the time interval between the moment of choice and the actual events, specifically, the instantaneous rate of time preference for a choice τ time units away can be expressed as $\gamma(1+\alpha\tau)^{-1}$, where γ and α are positive parameters.

Mental accounting The process by which an individual weighs the costs and benefits of an action or a consumption choice.

Normative implications Logical conclusions from a theory, which refer to the actions that should be taken by a welfare-maximizing policy maker.

Present bias The tendency to overweigh benefits or costs that are incurred in the present relative to those which are incurred in the future. Present bias suggests individuals do not discount exponentially.

Rational choice Behavioral patterns that are minimally consistent in the sense that if A is selected over B and B is selected over C, then A must be selected over C. May loosely be considered as behaviors intended to achieve some goal through weighing off broadly defined costs and benefits.

Reinforcement Consuming more of an addictive good today will increase the value given to consumption of the addictive good tomorrow.

Tolerance The consumption of a given amount of an addictive good in the future will yield less satisfaction, the higher is consumption of the addictive good today.

Utility A numerical representation of preferences in which more preferred consumption choices are given a higher number than less preferred consumption choices.

Utility projection bias The tendency of an individual to incorrectly predict that future preferences will closely resemble current preferences.

Introduction

What do economists add to the multidisciplinary discussion of addiction? In this article, economic theories of addiction, statistical evidence produced by economists on addictive behaviors, and resulting policy implications are described.

The manner in which economists approach addictive behaviors differs in some ways from the approaches of other disciplines. Medical and public health research often views addiction as, by definition, maladaptive. Addicts passively submit to urges rather than actively make rational consumption decisions. Consumption of an addictive good is itself beyond the control of the individual. The National Institute on Drug Abuse uses the following definition:

> Addiction is defined as a chronic, relapsing brain disease that is characterized by compulsive drug seeking and use, despite harmful consequences. It is considered a brain disease because drugs change the brain – they change its structure and how it works. These brain changes can be long lasting, and can lead to the harmful behaviors seen in people who abuse drugs.

By this definition, addiction is characterized by physiological changes and research often focuses on the neurological and psychological mechanisms underlying those changes (see Redish *et al.*, 2008 for a cross-disciplinary review of addiction research). Alcohol and other drug addictions are found to cause physical changes in body functioning, such as reductions in functioning of neurotransmitter activity like dopamine, and these neurotransmitters are part of the brain's reward system (Koob and Le Moal, 2008). These physiological changes are often observed in conjunction with, and indeed difficult to disentangle from, psychological changes such as increased depression and anxiety (Newlin, 2008).

Economists differ in generally focusing on models intended to reveal how social phenomena involving addictive behaviors emerge, which requires models suitable for investigating the manner in which addicts alter their behaviors as incentives change. By how much do smokers change their cigarette consumption if tobacco taxes increase, and over what time period? Do illicit drug addicts change their behavior as criminal penalties imposed on drug possession vary, and if so, how is the market for illicit drugs affected? What are the

private and social costs of addictive behaviors? How are addictive behaviors related to income? Which policies tend to reduce harms to addicts and to nonaddicts? These sorts of questions are better addressed using a combination of abstract behavioral models combined with statistical evidence on addictive behaviors, prices, and other incentives than by detailed exploration of physiological mechanisms.

Following Becker and Murphy (1988), economists often use the following definition:

Addiction: A good or activity is addictive for a given person at a given time if an increase in the person's consumption today causes an increase in consumption tomorrow, other things equal.

Loosely speaking, you are addicted to cigarettes in the economic sense if smoking more today causes you to smoke (or want to smoke) more tomorrow. Increased consumption of a nonaddictive good, however, does not cause you to want to consume more today if you happened to consume more of it yesterday; your desire to drink milk today is independent of your past milk consumption. Note that this notion of addiction does not require the addiction to operate through an action of the drug on the brain, although it is consistent with such an action. Nor does this definition require that the activity is maladaptive; a person may be addicted in the economic sense to, for example, health-enhancing exercise. Finally, whether a given good or activity is addictive may vary across people and over time within a given person's life.

The economic definition of addiction is a purely behavioral definition, as opposed to alternate conceptions involving physiological processes. Nonetheless, in Becker and Murphy's canonical model, people exhibit reinforcement and tolerance, elements of alternate conceptions of addiction. Reinforcement here means that increasing consumption of an addictive good today increases the marginal value that is given to consumption of the addictive good tomorrow. Tolerance suggests that consuming a given amount of the addictive good today yields less utility when consumption of the addictive good yesterday was higher.

The implications of this apparently straightforward notion of addiction are surprisingly complex. In the next Section Perfectly Rational Addiction, the canonical addiction model in economics, the rational addiction model of Becker and Murphy (1988), is discussed. This model is highly stylized, imposing strong assumptions about preferences and information, but the model is able to mimic many aspects of addictive behavior, make predictions that are possibly surprising but verified by evidence, and provide a framework for empirical analysis of taxation and other policies intended to limit consumption of addictive goods. Research building on this framework to incorporate more realistic behavioral and information assumptions is considered in Sections Imperfectly Rational Models of Addiction and Irrational Models of Addiction. Following Cawley (2008) economic models are distinguished as falling into one of the three categories: models of perfect rationality, models of imperfect rationality, and models of irrationality. Finally, in Sections Empirical Evidence and Policy Implications of Addiction Perspectives the statistical evidence and policy implications stemming from this line of research are discussed.

Perfectly Rational Addiction

Economic models typically assume that people have well-defined goals and tend to make decisions that further those goals. For example, one's goal as a commuter driving home from work may be to choose a route to minimize your driving time, and model worlds with many drivers each attempting to achieve that goal are used to predict how changes in a road system would affect traffic patterns. People in a model are 'rational' if they make decisions that are consistent with their goals. It is important to emphasize that 'rational' in this context is a technical jargon: It loosely means people weigh the benefits and the costs of a given action when making their decision, but it does not make any judgment about what defines a cost or a benefit per se (Mas-Colell *et al.*, 1995).

Consider a simple example of a model of consumer choice invoking this rationality assumption. A person can buy cigarettes or various other goods and services with a given amount of money. Given income and the prices, all affordable combinations of cigarettes and other goods define a 'menu' from which the person must choose. If the price of cigarettes is US$10 per pack and people have US$100 to spend, then 11 packs of cigarettes is not on their menu. If they can consistently rank these options from most to least preferred, which implies that if they rank A higher than B and B higher than C, they must rank A higher than C, then they are rational in the economic sense of the word. Given their rankings, how their choices will vary as the economic environment varies can be predicted. Whether a given change in economic environment makes the person better off in a well-defined sense – that is, makes an option available which is preferred to the current choice – can also be deduced from the model.

The rational addiction model extends this sort of analysis to capture special properties of addictive goods and activities. Canonical models of consumer choice take preferences as given: At some time, for example, a consumer has preferences over cigarettes and other goods and services, and chooses accordingly. The rational addiction model is dynamic; it is a model of decisions and outcomes over time, a complication which is necessary to capture the idea that addictive behaviors today affect behavior and outcomes in the future. Preferences over the addictive good and other goods and services at a given time are endogenous in the rational addiction model, as they depend on previous behavior.

The standard rational addiction model makes strong assumptions over this dynamic process, although these assumptions are weaker and more realistic than had been previously invoked in the literature. Before Becker and Murphy (1988) some economists had attempted to model consumption of addictive goods as 'habits' that have some but not all features of addictive behaviors. In these models, how much you smoke today depends on how much you have smoked in the past, but you do not take into account that your consumption in the future will change if you choose to smoke more today. People in these models are myopic and naive: They are constantly surprised when they discover that how much they smoked yesterday has changed their desire for cigarettes today.

Becker and Murphy (1988) consider the other extreme case: Instead of completely failing to understand that

tomorrow's outcomes depend on today's behaviors, Becker and Murphy consider a world in which people understand this relationship perfectly. They model addiction as stock, not unlike a capital stock, that increases or decreases over time according to the flow of consumption. Abstinence leads to depreciation in the addictive stock over time – if you quit smoking today, your level of addiction to cigarettes will decay over time. This decay is offset by consuming the addictive good – smoking more today increases your stock of addiction. Whether you become more or less addicted over time depends on whether you consume enough of the addictive good to offset the decay in your addiction. This model is intended to capture stylized facts about the dynamics of addiction: Addiction does not start or stop instantaneously, rather, an addiction is built up over time through use of the addictive substance and addiction decays over time with abstinence or decreased consumption. The rational addicts choose current consumption being fully aware of how their behavior today affects their stock of addiction, and thus their behavior, tomorrow.

Becker and Murphy prove that people in such a world will display behaviors that are typically associated with addiction. The model predicts that a rational addict builds tolerance and goes through withdrawal. Sufficiently strong addictions generate 'cold turkey' quitting behavior as opposed to quitting slowly by gradually decreasing consumption over time. Further, the model allows economists to predict how addicts will respond to a change in the price of the addictive good, and hence provides a lens through which tax policy toward tobacco, alcohol, and other addictive goods can be viewed. Finally, the model generates a falsifiable prediction about behavioral responses to price changes: An anticipated future increase in the price of the addictive good will cause rationally addicted people to immediately reduce consumption of the addictive good. This effect follows from the rational addict understanding that a future price increase will make consuming at current levels in the future more costly, and the pain of withdrawal is diminished if consumption is reduced by small amounts over time rather than a large amount in the future. Likewise, a price decrease in the past will result in more past consumption, leading to a higher addictive stock, and greater consumption levels today. A consequence of this model is that all prices, past, current, and future, influence the person's current consumption decision.

An extension of the rational addiction framework to multiple addictive behaviors can also explain cyclical binging and abstinence. Palacios-Huerta (in press) shows binging behavior is a prediction of the rational addiction model in which there are multiple, substitutable, addictive goods. For example, consider someone who is addicted to both cannabis and to alcohol but considers them substitutes. If the two behaviors deplete one's health stock in different ways (one through liver damage and the other through lung damage), binging behavior will result as individuals alternate between the two activities, binging on alcohol while lung health recovers and binging on cannabis while liver health recovers.

These models help us understand addictive behaviors in realistic settings in which there is a complicated relationship between consumption of various drugs and policies, which may fail if they ignore these relationships.

The rational addiction model has a number of important consequences with respect to how the policy is evaluated and implemented. First, people who discount the future heavier are more likely to engage in addictive behaviors. This is particularly relevant in explaining why smoking uptake is so much higher among youth rather than adults. Second, the full effect of a permanent change in prices on individual behavior cannot be judged in the short run. Facing a price increase, the addict will reduce consumption gradually over time. Becker and Murphy predict that in the long run addiction leads to a more price-responsive demand, a hypothesis that is confirmed in the empirical literature discussed in Section Imperfectly Rational Models of Addiction. Finally, announcing future change in tax policies will impact current consumption.

The Becker and Murphy (1988) model is subject to two major criticisms. The first is that the model predicts that addicts will never regret their choices. A large body of evidence falsifies this prediction. The second criticism is that strong assumptions are made about information. In particular, people can accurately predict the future effects of their current consumption. A much-criticized welfare implication follows: The rational consumer always makes optimal consumption choices, and policy interventions designed to deter consumption of addictive goods are generally welfare reducing.

Imperfectly Rational Models of Addiction

Several extensions followed the Becker and Murphy (1988) model to address the restrictive assumptions of perfect rationality. Two assumptions that have received attention are the assumption of perfect information and foresight and the assumption of exponential discounting. Models that address these concerns otherwise follow a common strategy to Becker and Murphy (1988); people continue to make decisions that they believe – at the time the decision is made – are in their best interest.

The perfectly rational consumer correctly predicts the effect that consumption of an addictive good will have on their behavior in the future. However, this assumption is contrary to evidence that suggests people are very poor judges of their future preferences and tastes: People tend to bias estimates of their future tastes toward being like their current tastes (Loewenstein *et al.*, 2003). This utility projection bias is particularly troublesome when nonaddicted people need to make judgments about the impact that consumption of an addictive good will have on their future preferences. Badger *et al.* (2007) show that even seasoned heroin addicts underestimate the influence of their addiction on behavior. To address this issue, Orphanides and Zervos (1995) extend the rational addiction model to allow people to be uncertain about how addictive they will find a good or activity. People update their beliefs about the addictiveness of the good by observing the actions of those around them and through their own experimentation. People try their first cigarette, for example, without knowing how addictive they will find smoking. In this model, addicts may regret their past choices even though they make the best choices they can with the information available at the time. Some people will underestimate their potential for addiction and regret having become an addict.

For most addictive goods, such as cigarettes or narcotics, consumption leads to an immediate benefit while the cost, such as poor health, is realized in the future. For this reason the manner in which people discount the future has important consequences for the rational addiction model. Dynamic rationality implies that people discount exponentially and consistently. That is, a predetermined and constant rate of discount is applied to every period in the future. Models of hyperbolic discounting instead assume that people have a present bias, applying a larger discount rate to events far in the future than events that are to occur sooner. A number of controlled experiments find that hyperbolic discounting is a more accurate depiction of behavior than exponential discounting (for a review of the evidence on hyperbolic discounting see Frederick *et al.* (2002)). If this type of discounting accurately reflects the decision process, then people will underweigh the future costs of their actions at the time decisions are made. Gruber and Koszegi (2001) extend the perfectly rational model to include hyperbolic discounting. This model yields dramatically different normative implications than the Becker and Murphy (1988) framework, as there is an 'internality' – one's smoking today harms one's future self, and one's present self and one's future self are, in effect, in conflict.

In the canonical rational addiction model, and some extensions thereof, people make lifetime consumption plans and adjust them as new information is revealed. A different approach to modeling the behavior of a rational addict is taken by Gul and Pesendorfer (2007) who consider people who make a consumption plan, but need to exert costly self-control to see it through. Consider a rational alcoholic who determines an optimal consumption plan of four drinks per day. According to the Becker and Murphy framework, absent any changes in information, the rational alcoholic will see this plan through, consuming four and only four drinks daily. However, such a plan requires self-control. The temptation of having extra alcohol in the house may cause the alcoholic to deviate from the four drink per day plan, and instead have five or six drinks. This deviation from the plan in the current period makes self-control in future periods even more difficult. In this framework, an addictive good is harmful if people experience an ever-widening gap between their planned optimal consumption and their actual consumption. Like the Becker and Murphy model, addicts in this model respond to anticipated future price increases by decreasing current consumption patterns, and may exhibit binging and abstinence cycles. However, unlike the Becker and Murphy model, this model can explain the use of short-term commitment devices, such as rehabilitation centers, by addicts.

Irrational Models of Addiction

Most researchers outside the field of economics do not think about addiction in a rational decision framework. This largely follows from an empirical anomaly: Addicts commonly express a strong desire to reduce or stop their consumption of addictive goods but fail to follow through. The ability of addiction to override rationality is captured in a statement made by David Kessler, former commissioner of the Food and Drug Administration: "Once they have started smoking regularly, most smokers are in effect deprived of the choice to stop smoking" (statement to the House Subcommittee on Health and the Environment 25 March 1994).

The economist views a consumer as irrational if decision making ignores relevant information and incentives. For example, models of myopic decision making can be thought of as irrational; people do not consider how current decisions will impact future outcomes. It has been argued that addiction leads to a failure in the processing of information, and therefore causes the addict to deviate from rational decision making. Clinical evidence suggests that addicts exhibit a bias in their mental accounting, placing too little weight on the negative consequences of their behavior (see Tomer (2001) for a discussion). Further, the observed procrastination of addicts, wishing to quit but continually putting off action, suggests that rational behavior does not fully capture addictive behavior.

Even if the consumption of addictive substances is irrational, surely addicts are rational in some facets of their lives. Bernheim and Rangel (2004) model a 'cue-triggered' decision process built on three premises. First, the consumption of addictive goods by an addict is often a mistake. Second, increased consumption of addictive goods makes addicts more sensitive to random environmental cues that trigger mistaken consumption. Third, addicts understand the cue-triggered process and take steps to manage their susceptibility. The cue-trigger model draws on neurological evidence that addictive substances interfere with the operation of pleasure and reward processes in the brain. In this model, people face a dynamic decision process in which environmental cues trigger a 'hot' decision-making mode during which the substance is consumed regardless of relevant incentives and information. When operating in a 'cold' mode people fully consider the current and future consequences of their actions, including how decisions influence the likelihood of being cued into a hot mode. For example, if stressful circumstances exacerbate the cravings associated with cigarette addiction, then a person trying to quit smoking will likely take steps to avoid stressful circumstances. Addicts in this model are aware of their propensity to make consumption mistakes and will take steps to precommit to future consumption and mitigate cues.

Empirical Evidence

How well do any of the models previously discussed capture the behavior of addicts? In this section, the econometric literature on addictive behaviors are discussed.

An advantage of the rational addiction model is that it provides a framework in which to develop statistical models of the consumption of addictive goods (Becker *et al.*, 1994). A key insight from this framework, and a testable implication, is that past, current, and future prices will all affect consumption behavior. From models that estimate the size of these effects researchers can predict how policy changes such as tax increases or decreases will impact across people and across time. These models are estimated using either aggregate or individual-level data on consumption of

addictive goods, prices, incomes, and other determinants of consumption. The addictive good under scrutiny varies across studies: There are many studies of tobacco and smoking behavior; other possibly addictive goods that have been empirically examined in this framework include alcohol, marijuana, cocaine, gambling, and even coffee.

The key and oft-replicated finding from the empirical literature on addictive goods is that people, even addicts, respond to an increase in the current price of addictive good by decreasing current consumption (Chaloupka and Warner, 2000; Gallet and List, 2002; DeCicca *et al.*, 2008; Sen *et al.*, 2010). If the consumption of addictive goods were an entirely irrational behavior, then consumption would not vary systematically and predictably with prices. Contrary to irrationality, it is well established that consumption of addictive goods responds to price incentives. This implies that the consumption of addictive goods is, at least to some degree, rational.

Much of the empirical literature considers one addictive good or activity in isolation, but some work attempts to model joint consumption of multiple addictive goods. Generally, a change in the price of one addictive good will affect consumption of all addictive goods. Examples include Dinardo and Lemieux (2001), who present statistical evidence suggesting that youths substitute alcohol and cannabis, and Cameron and Williams (2001), who estimate own- and cross-price effects in demand for alcohol, tobacco, and cannabis and find that alcohol and cannabis may be substitutes, whereas alcohol and cigarettes are complements. Jofre-Bonet and Petry (2008) document a complex pattern of substitutes and compliments between various addictive substances for heroin and cocaine addicts. They find that heroin and cocaine addicts use marijuana, valium, and cigarettes as substitutes.

The intertemporal influence of prices on behavior constitutes the main estimable difference between nonaddictive goods and addictive goods: The consumption of nonaddictive goods is not influenced by past or future prices. Using this testable hypothesis, many papers claim to find strong evidence of rational addiction, even for goods such as coffee (Olekalns and Bardsley, 1996). However, Auld and Grootendorst (2004) demonstrate that using aggregate data (e.g., total cigarette sales by the US state over time) to estimate addiction models tends to yield spurious evidence in favor of addiction; these methods are biased in favor of finding evidence of addiction even when the good under scrutiny is actually nonaddictive. This problem can be avoided by using individual-level data or using quasi-experimental empirical strategies. For example, Gruber and Koszegi (2001) use the preannouncement of state excise taxes on tobacco and show that smokers are forward looking in their behavior.

Similarly, statistical models show that past consumption affects future consumption in the manner predicted by rational addiction models, with an effect that diminishes over time (Gilleskie and Strumpf, 2005). The effect of past consumption on current behavior has also been found to vary markedly across people in the manner predicted by economic theory (Auld, 2005). Keeler *et al.* (1999) find that smokers respond to price incentives and that smokers with higher socioeconomic status are more likely to quit, all of which is predicted by the rational addiction model.

The empirical literature has had less success in cleanly distinguishing between different models of addiction. Goldfarb *et al.* (2001) note that commonly used empirical methods in this literature cannot be used to support or refute rational models over nonrational models. In particular, all economic models of addiction predict the observed responsiveness to prices. Levy (in press) extends the empirical literature by deriving the conditions under which the perfectly rational model of addiction can be tested against models that exhibit present bias and utility projection bias. Further, he derives estimating conditions that allow him to distinguish between the two forms of bias. Using data from the US National Health Interview Survey he finds that observed behavior strongly rejects perfect rationality, and estimates of projection bias and utility bias are strong and consistent with previous studies of nonaddictive behaviors. Consistent with the existence of these biases, Gruber and Mullainathan (2005) find that tobacco taxes increase self-reported happiness for people with a high propensity to smoke. This is suggestive that taxes are correcting for an internality.

Policy Implications of Addiction Perspectives

The extent to which people operate in the perfectly rational framework of Becker and Murphy has important normative implications that impact policy. Under the assumptions of the perfectly rational framework, people consume addictive goods according to their individual preferences and policy interventions are welfare improving only to the extent that they account for externalities associated with addictive consumption. For example, policy to reduce alcohol consumption is only welfare improving to the extent that it reduces externalities (involuntary benefits or, here, costs imposed on third parties), such as traffic accidents and violent crime.

However, even small departures from perfect rationality may imply a greater role for policy (Laux, 2000; Suranovic *et al.*, 1999). Policy intervention can be welfare enhancing when people have incorrect or insufficient information, or if the decision-making process is in part driven by irrational behavior such that 'internalities' (costs a person imposes on their future self as a result of irrational behavior) result. However, the specific type of policy intervention that should be implemented depends to a large extent on the model of consumer behavior. Further, it should be cautioned that policies designed to correct internalities are by definition paternalistic and hence controversial (Viscusi, 2002).

Taxation

One oft-suggested tool for intervention policy is taxation. There are sound reasons to tax addictive goods that do not hinge on their addictive property. The external costs of some addictive goods, such as second-hand smoke from cigarettes, can and should be internalized with taxes. Generating government revenue by taxing inelastically demanded goods creates fewer market distortions than taxing goods with elastic demand. Therefore, addictive goods with inelastic demand should be heavily taxed for revenue creation. These arguments

do not rest on improving the welfare of potential addicts per se.

Addiction itself has no clear-cut implication for tax policy because different models generate different optimal tax policies. For example, if people have time-inconsistent preferences, such as in hyperbolic discounting models, or incorrectly forecast utility with a present bias, then the optimal tax will be higher than those predicted by perfectly rational models of addiction with only externalities (Gruber and Koszegi, 2001; Levy, in press). Present bias and utility-projection bias mean that people place too little importance on, or systematically misjudge, how current behavior will impact their future selves. Therefore, taxes on addictive goods can enhance welfare by forcing people to internalize the impact of their current behavior on their future selves. In a simulation of their hyperbolic discounting model, Gruber and Koszegi (2001) estimate that a tax of at least US$1.00 per pack of cigarettes should be applied to correct the present bias in discounting. With both a utility projection bias and a present bias in discounting, Levy (in press) estimates that an optimal corrective tax should be set considerably higher.

Not all economic models of addiction imply corrective tax policy to improve the well-being of addicts and potential addicts. In the temptation model of Gul and Pesendorfer (2007), individuals optimally consume the addictive good given the temptation they face and their ability to commit to future consumption. In this framework, tax policy, when used alone, is always welfare reducing: A tax increases the cost of consuming the addictive good but does not remove or reduce temptation. Likewise, in the cue-triggered decision-making model, Bernheim and Rangel (2004) find that taxation of addictive goods may be harmful, as it may do little to change the consumption behavior of addicts and instead crowds out consumption of nonaddictive goods. Even if taxation is beneficial, Bernheim and Rangel find that banning consumption of the addictive good may be a superior policy to taxation.

Bans

Bans and restrictions are perhaps the most commonly used policy intervention with respect to addictive substances. Many models of imperfect rationality and irrational behavior predict that bans can be welfare improving. Gul and Pesendorfer (2007) show that prohibitive policies are always welfare improving because they limit the opportunity to make addictive consumption choices, thereby reducing temptation. A partial ban, say in the workplace but not at home, is considered by de Bartolome and Irvine (in press) who model the short-run behavior of an addict. The addict likes higher overall consumption of the addictive good but dislikes variance in consumption throughout the day. The workplace ban reduces daily consumption through the addict's dislike of variance; reductions in workplace consumption of the addictive good are not fully reallocated to consumption at home. These models, however, are not designed to evaluate the overall implications of prohibitions and, in particular, do not attempt to assess unintended consequences of prohibitions (Miron and Zweibel, 1995) nor the operation of black markets (Lee, 1993), so these policy implications must be considered as only part of a much larger story.

Partial bans in the form of controlled distribution offer another policy instrument. In a cue-triggered model of behavior these can be used to improve welfare. Specifically, when distribution is controlled in such a way that addicts are forced to 'stock-up' in cold states, rather than make purchases as hot states arise, they will choose the optimal level of consumption for their future selves. Partial bans allow the cold state decision maker to commit to hot state consumption. Such policy could potentially be achieved through the use of prescriptions or time-specific restrictions on sales.

Information and Insurance

When people lack information about their susceptibility to addiction, public provision of accurate information about addictive goods can enhance welfare (Orphanides and Zervos, 1995). Further, continued research and dissemination of information on the assessment of individual risk with respect to addiction can be welfare enhancing, even when people know the true distribution of risk across the population. Such efforts will assist people in better assessing their uncertain susceptibility to addiction. The need for accurate information also means that there is a welfare case to be made for restricting misleading advertising campaigns (Orphanides and Zervos, 1995). Similarly, limiting cue use in advertising for addictive goods is potentially beneficial (Bernheim and Rangel, 2004).

When uncertainty exists about susceptibility to addiction or the environmental cues which an individual will face in the future, there is an opportunity for a welfare-enhancing policy intervention through insurance provision. This insurance may come in the form of subsidization for rehabilitation and withdrawal treatment. It should be noted that the moral hazard and asymmetric information problems that accompany this market are nontrivial (Orphanides and Zervos, 1995).

Finally, Tomer (2001) argues that even with full information addicts may incorrectly weigh the costs and benefits associated with their behavior, placing too little weight on the negative consequences of their actions. In this way, continued addiction may be the result of systematic mental accounting of errors in which the addict places too little weight on the potential loss of family, friends, and other forms of social capital, and too much weight on the immediate cravings associated with addition. In this case, interventions by family and friends, to make the benefits of abstinence salient, will be welfare improving. Such interventions are commonly used in cases of severe addiction.

Summary

Economists approach addiction from a behavioral point of view and with a focus on assessing and measuring the effects of policy interventions, such as taxation and prohibitions. The canonical model, Becker and Murphy's (1988) rational addiction model, considers a world in which people are aware that their consumption of addictive goods today will affect their behavior in the future and make choices accordingly. This model provides a framework to analyze addictive behaviors and

has led to a large and detailed body of empirical evidence. The model has also been extended in many ways to incorporate more realistic psychological, physiological, and social aspects.

The standard model makes several predictions that are falsified, notably including the prediction that addicts do not regret their past decisions. A number of theoretical investigations relax or otherwise modify the assumptions of the standard model to address this failing. In these models, people may not know themselves well enough to predict whether they will find some good or activity addictive, or they may have self-control problems that prevent them from quitting a harmful addiction even though they realize that addiction is harmful. Policy implications vary across theoretical models as the assumptions driving the model vary, so the theoretical literature has not come to a consensus on optimal policy toward addictive goods. Current research continues to incorporate results from other disciplines, such as neuroscience, into economic models.

Economists have also produced a large body of statistical evidence detailing what kind of people consume various addictive goods, the extent to which people respond to changes in the price of addictive goods, and how consumption varies with prices, income, and other incentives over short and long time periods. This literature shows that addicts do respond to prices and other incentives, that past consumption of addictive goods causes current consumption of addictive goods, and that consumption of a given addictive good is best understood as a part of a profile of consumption of various addictive goods rather than in isolation, for example, policy makers should consider the effects of a change in heroin policy on alcohol consumption in addition to heroin consumption.

Acknowledgment

Auld thanks the Center for Addictions Research of British Columbia for financial support.

See also: Alcohol. Illegal Drug Use, Health Effects of. Smoking, Economics of

References

Auld, M. C. (2005). Causal effect of early initiation on adolescent smoking patterns. *Canadian Journal of Economics* **38**(3), 709–734.

Auld, M. C. and Grootendorst, P. (2004). An empirical analysis of milk addiction. *Journal of Health Economics* **23**(6), 1117–1133.

Badger, G. J., Bickel, W. K., Giordano, L. A., Jacobs, E. A. and Loewenstein, G. (2007). Altered states: The impact of immediate craving on the valuation of current and future opioids. *Journal of Health Economics* **26**(5), 865–876.

de Bartolome, C. and Irvine, I. J. (in press). The economics of smoking bans. *Working Paper no. 201027*. Geary Institute, University College Dublin.

Becker, G., Grossman, M. and Murphy, K. (1994). An empirical analysis of cigarette addiction. *American Economic Review* **84**(3), 396–418.

Becker, G. and Murphy, K. (1988). A theory of rational addiction. *Journal of Political Economy* **96**(4), 675–700.

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review* **94**(5), 1558–1590.

Cameron, L. and Williams, J. (2001). Cannabis, alcohol, and cigarettes: Substitutes or complements? *Economic Record* **77**(236), 19–34.

Cawley, J. (2008). Reefer madness, Frank the tank or pretty woman: To what extent do addictive behaviors respond to incentives? In Sloan, F. A. and Kasper, H. (eds.) *Incentives and choice in health care*. Cambridge, MA: MIT Press.

Chaloupka, F. and Warner, K. (2000). The economics of smoking. In Culyer, A. and Newhouse, J. (eds.) *Handbook of health economics* 1(B), pp. 1539–1627. North Holland: Elsevier.

DeCicca, P., Kenkel, D. and Mathios, A. (2008). Cigarette taxes and the transition from youth to adult smoking: Smoking initiation, cessation and participation. *Journal of Health Economics* **27**(4), 904–917.

Dinardo, J. and Lemieux, T. (2001). Alcohol, marijuana, and American youth: The unintended consequences of government regulation. *Journal of Health Economics* **20**(6), 991–1010.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature* **40**(2), 351–401.

Gallet, C. and List, J. (2002). Cigarette demand: A meta-analysis of elasticities. *Health Economics* **12**(10), 821–835.

Gilleskie, D. and Strumpf, K. (2005). The behavioral dynamics of youth smoking. *Journal of Human Resources* **40**(4), 822–866.

Goldfarb, R. S., Leonard, T. C. and Suranovic, S. M. (2001). Are rival theories of smoking underdetermined? *Journal of Economic Methodology* **8**(2), 229–251.

Gruber, J. and Koszegi, B. (2001). Is addiction rational? Theory and evidence. *Quarterly Journal of Economics* **116**(4), 1261–1303.

Gruber, J. H. and Mullainathan, S. (2005). Do cigarette taxes make smokers happier? *The B.E. Journal of Economics Analysis & Policy* **5**(1), 1–45.

Gul, F. and Pesendorfer, W. (2007). Harmful addiction. *Review of Economic Studies* **74**(1), 147–172.

Jofre-Bonet, M. and Petry, N. M. (2008). Trading apples for oranges? Results of an experiment on the effects of heroin and cocaine price changes on addicts polydrug use. *Journal of Economic Behavior and Organization* **66**(2), 281–311.

Keeler, T. E., Marciniak, M. and Hu, T. (1999). Rational addiction and smoking cessation: An empirical study. *Journal of Socio-Economics* **28**(5), 633–643.

Koob, G. F. and Le Moal, M. (2008). Addiction and the brain antireward system. *Annual Review of Psychology* **59**, 29–53.

Laux, F. L. (2000). Addiction as a market failure: using rational addiction results to justify tobacco regulation. *Journal of Health Economics* **19**(4), 421–437.

Lee, L. W. (1993). Would harassing drug users work? *Journal of Political Economy* **101**(5), 939–959.

Levy, M. (in press). An empirical analysis of biases in cigarette addiction. Working Paper.

Loewenstein, C., O'Donoghue, T. and Rabin, M. (2003). Projection bias in future utility. *Quarterly Journal of Economics* **118**(4), 1209–1248.

Mas-Colell, A., Whinston, M. and Green, J. (1995). *Microeconomic theory*. Oxford: Oxford University Press.

Miron, J. and Zweibel, J. (1995). The economic case against drug prohibition. *Journal of Economic Perspectives* **9**(4), 175–192.

Newlin, D. B. (2008). Are "physiological" and "psychological" addiction really different? Well, no!... um. er, yes? *Substance Use and Misuse* **43**(7), 967–971.

Orphanides, A. and Zervos, D. (1995). Rational addiction with learning and regret. *Journal of Political Economy* **103**(4), 739–758.

Olekalns, N. and Bardsley, P. (1996). Rational addiction to caffeine: An analysis of coffee consumption. *Journal of Political Economy* **104**(5), 1100–1104.

Palacios-Huerta, I. (in press). Multiple additions. *Working Paper 2001–20*. Department of Economics, Brown University.

Redish, A. D., Jensen, A. and Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences* **31**, 415–487.

Sen, A., Ariizumi, H. and Driambe, D. (2010). Do changes in cigarette taxes impact youth smoking? Evidence from Canadian provinces. *Forum for Health Economics and Policy* **13**(2), Aricle 12.

Suranovic, S., Goldfarb, R. and Leonard, T. (1999). An economic theory of cigarette addiction. *Journal of Health Economics* **18**, 1–29.

Tomer, J. F. (2001). Addictions are not rational: A socio-economic model of addictive behavior. *Journal of Socio-Economics* **33**, 243–261.

Viscusi, W. K. (2002). The new cigarette paternalism. Regulation *Winter 2002–2003* 58–64.

Further Reading

Heyman, G. M. (2009). *Addiction: A disorder of choice*. Cambridge, MA: Harvard University Press.

Adoption of New Technologies, Using Economic Evaluation

S Bryan, University of British Columbia, Vancouver, BC, Canada; Vancouver Coastal Health Research Institute, Vancouver, BC, Canada, and University of Aberdeen, Aberdeen, UK
I Williams, University of Birmingham, Birmingham, UK

Glossary

Acceptability The requirement that economic analyses provide information that is seen by end-users to be relevant and appropriate to the decisions they face, takes into account relevant contextual factors, and is delivered in a timely fashion.
Accessibility The requirement that economic analyses can readily be understood and interpreted by end-users.
Cost-effectiveness acceptability curve (CEAC) The CEAC plots the probability that the intervention in question is cost-effective against a range of possible threshold values.
Coverage A decision to 'cover' a technology indicates that its cost will be reimbursed as part of an insurance package.
Interactive model of research utilization A model in which policy formulation is understood as a nonlinear process involving multiple agents and influences.
Net-benefit statistic The net-benefit statistic expresses the additional health effects in monetary units by using an estimate of the 'maximum willingness to pay' per unit of health gain.
Problem-solving model of research utilization A model in which empirical and analytical evidence is applied directly to a policy problem, enabling the optimal solution to be identified and implemented.

Introduction

The overarching central issue addressed by the discipline of economics is resource scarcity. In one sense or another, all economists are working on questions that have some connection to scarcity and limits. Thus, the primary purpose of economic analysis, and cost-benefit and cost-effectiveness analysis (CEA) in particular, is to support decision-making necessitated by the scarcity problem. Therefore, economic evaluation information is generated with the direct intention of influencing policy – but is that objective achieved? This is the central question addressed in this article.

The policy frame here relates to decisions on coverage of medical interventions. A decision to 'cover' a technology indicates that its cost will be reimbursed as part of an insurance package, and so it involves setting limits on the health care services that can be accessed or provided. Coverage decisions are taken in health systems where private insurance is widely seen and in systems dominated by publicly funded insurance programs.

This article initially provides a definition of economic evaluation typically undertaken to inform coverage decisions and then introduces a case study, the UK's National Institute for Health and Clinical Excellence (NICE). The problem, reflected in the lack of use of such information, is then outlined, with supporting evidence from the published literature presented. The article then provides a discussion of how some of the barriers and obstacles to use might be overcome.

Normative Economic Evaluation

Much economic evaluation work in health care, seeking to support coverage decision making, has a 'normative' bent. That is, the role of the economist has been to indicate the nature of the resource allocation decision that ought to be followed if certain objectives are to be achieved. An important prerequisite for such a normative stance is that the analyst has a good understanding of the objective function (i.e., what should the health service be seeking to achieve?) and the decision rules to be applied. As Culyer (1973) points out, the process of agreeing objectives is not necessarily straightforward:

> In the real world … policy makers and most other people who seek economic advice do not have well-articulated ideas of their objectives. One of the first tasks of a cost-benefit analyst, for example, is usually to seek to clarify the objectives – even to suggest some.
> Culyer (1973, p. 254)

Many health economists have taken Culyer at his word, proposing an objective of maximising population health benefits and, although there are those who argue for a broader set of objectives, the proposition does receive some support from policy makers and the public more generally. The difficulties and disputes arise primarily around attempts to measure health. Over the course of the past 20 years or so the subdiscipline of health economics has had a methodological focus on health measurement and valuation. The result is a measure of health that can be operationalized for use in policy making, that is, the quality-adjusted life-year (QALY). The decision rule, therefore, is to invest in those technologies that produce the largest QALY gains for a given level of cost. To inform such decisions, normative analyses tend to provide results in the form of an incremental cost-effectiveness ratio (ICER), a net-benefit statistic and a cost-effectiveness acceptability curve (CEAC).

- The ICER reports the ratio of additional costs to additional health effects associated with a new intervention (e.g., cost per QALY gained).
- The net-benefit statistic expresses the additional health effects in monetary units by using an estimate of the

 doi:10.1016/B978-0-12-375678-7.01414-0

'maximum willingness to pay' per unit of health gain, where available.

- The CEAC plots the probability that the intervention in question is cost-effective against a range of possible threshold values to define cost-effectiveness.

A National Institute for Health and Clinical Excellence Case Study

Perhaps the most researched example of use of economic evaluation in coverage decision making is the UK's NICE. In many respects, NICE has set the standard for evidence-informed coverage decision making and openness to the application of economic analyses.

The Institute, established in 1999, has as one of its functions the appraisal of new and existing health technologies. Coverage decisions made by NICE are based on explicit criteria and are informed by evidence, including an economic evaluation. The evidence is interpreted and considered by the Technology Appraisal Committee, and that Committee formulates recommendations and guidance on the use of the technology in the National Health Service (NHS) in England and Wales.

There can be no doubt that the technology appraisal decisions at NICE are driven in large part by the results of economic analyses. This was stated explicitly by the Institute's Chairman, Sir Michael Rawlins, who stated that in determining its guidance, NICE would take six matters into account, including both clinical and cost-effectiveness (Rawlins and Culyer, 2004). Further, in the Secretary of State's Direction to NICE when it was established in 1999, the intent was clearly stated: NICE should consider the broad balance of clinical benefits and costs.

As a crude example to demonstrate that cost-effectiveness drives decisions, in the appraisal of statin therapy for secondary prevention of coronary heart disease, the ICER ranged from £10 000 to £16 000 per QALY gained and the guidance from NICE states: 'Statin therapy is recommended for adults with clinical evidence of coronary vascular disease' (NICE, 2006). However, when the ICER is much less favorable, in the case of Anakinra for rheumatoid arthritis the ICER was in the region of £105 000 per QALY gained, the guidance tends to be negative: "Anakinra should not normally be used as a treatment for rheumatoid arthritis. It should only be given to people who are taking part in a study on how well it works in the long term" (NICE, 2003).

This general picture is supported by the analyses of decisions taken by NICE and other agencies presented by Clement *et al.* (2009, p. 1437): agencies such as NICE make "recommendations that are consistent with evidence on effectiveness and cost-effectiveness but that other factors are often important." Qualitative work by Bryan *et al.* (2007, p. 41) tells a very similar story – examples of quotes from NICE committee members:

> "I think economic evaluation was regarded as being important from day one."

> "It [the CEA] seems to me to be the clincher really. If it's too high then it's not going to get funded."

The Problem

The NICE story is positive but it is important to understand that it is an outlier in terms of policy use of economic evaluation in health care. The broader literature on this topic has a consistent refrain, with concern expressed regarding the usefulness, or more precisely the lack thereof, of CEAs when applied in decision making processes. Responses to this concern have tended to centre on questions of how evaluation research by health economists can be made more useful and accessible to policy makers.

As a framework for considering these issues, the authors have previously grouped barriers to the use of economic analyses in health care decision-making under two headings: accessibility and acceptability. The accessibility concern includes issues such as interpretation difficulties, the aggregation of results, difficulties in accessing information, shortage of relevant skills, etc. Under an acceptability or relevance banner, a whole range of barriers might be considered relating to the timeliness of information provision, and the quality and nature of the information.

Thus, if one accepts this framework, the necessary requirements for economic evaluation evidence to be used in decision-making, relate both to accessibility and to acceptability. For the information to be accessible, it is required that the results of the economic analyses can readily be understood and interpreted by end-users. This is mainly concerned with issues of the presentation of information. For the information to be acceptable, it is necessary that economic analyses provide information that is seen by end-users to be relevant (i.e., providing data on parameters that are likely to influence the decision of the policy maker), information that is appropriate to the decisions they face, taking into account relevant contextual factors (e.g., budgetary arrangements commonly seen in the NHS), and that such analyses are seen as providing information in a timely fashion.

This article will now summarize the main themes that emerge from the published literature on this topic. The authors will then return to NICE and reflect further on its use of economic evaluation in light of these accessibility and acceptability criteria. The article will conclude with reflections of going forward, drawing on contributions from a more 'positive' approach to economics.

Empirical Work

This part of the article discusses the work of others who have researched the use of economic evaluation in health care decision making. A formal review of literature in this area has been published by Williams *et al.* (2008) and this article draws, in part, from that work.

The vast majority of empirical work in this field was conducted from the mid-1990s onwards. In terms of method,

there are three strands to the empirical literature:

- Surveys and questionnaires.
- Studies specifically of the NICE appraisals process, drawing solely on secondary sources.
- A prospective, case study approach, represented by a single study.

One of the most innovative pieces of research, going beyond surveys and interviews, was conducted by McDonald (2002). Based within an English Health Authority, she offered health economics support as a participant observer of a Coronary Heart Disease Strategy. She found that CEA was not geared toward assisting in the decision making processes prevalent at local levels of the NHS in England. This work highlighted barriers beyond those identified in previous UK studies. These are discussed below.

In a US context, use of formal CEA in technology coverage decisions is, if anything, even less commonly seen.

Successful application of CEA to policy has thus proved to be a challenge to decision makers across a range of health care systems. This low level of use occurs despite evidence suggesting that decision makers appreciate the potential value of cost-effectiveness information to the policy.

Studies of NICE have largely relied on data collected from secondary sources. Although these vary in approach to data analysis, each identifies CEA as a prominent feature in the Institute's work, in contrast to decision makers from all other studies.

Barriers to the Use of Economic Evaluation

Research indicates a plethora of active barriers to use of CEA. In relation to accessibility, there are three dimensions reported as significant within the literature. The first relates to the shortage of relevant analyses. Early studies in particular emphasize the difficulties decision makers face in obtaining economic evaluations. The second barrier derives from uncertainty or ignorance over how and from where existing studies can be accessed. This is compounded by the funding and access difficulties inherent in commissioning a new CEA that can be delivered in a timely manner. Finally, and – within this category of barriers – most consistently, studies demonstrated a lack of expertise in comprehension and interpretation. It is clear from studies at local levels that decision makers struggle to understand health economic analyses including the concepts and language used, and the presentational styles adopted.

These problems of accessibility are compounded by barriers relating to the perceived acceptability and ease of implementation of CEA. A small number of studies indicated that perceived methodological flaws were a major impediment to utilization. More commonly, studies found that decision makers did not always consider the source of CEAs to be independent. The pharmaceutical industry has been active in using CEAs to promote their products and studies repeatedly emphasize the distrust this engenders in decision makers.

Studies employing qualitative methods have uncovered factors relating to the complexity and interactive nature of the decision making environment, and therefore the competing drivers of decisions. Far from reflecting a problem-solving research-led model, health care decision making is subject to multiple influencing factors including: political considerations, administrative arrangements, equity concerns, societal opinion and the values and attitudes of decision makers. Interestingly, this multiplicity of competing considerations was also indicated in more recent quantitative analysis of NICE decisions.

The study by McDonald (2002) uncovered fundamental value conflict between decision makers' guiding principles and those underpinning normative health economics. She reinforces the assertion that single objectives are not routinely present in decision making and details instances of decision making which could not be said to be following any single maximization principle. As a participant observer, her attempts to introduce a rational, problem-solving approach to resource allocation resulted in a 'paralysis' caused, in part, by complex funding constraints. Rational approaches to policy formulation were considered by decision makers to be less satisfactory than standard nonrational practices of 'muddling through' in a context of resource scarcity.

Finally, studies from across the range of methodological types suggest that decision makers perceive recommendations from CEAs to be difficult to implement. For example, budget holders operating within short-term budgeting cycles may be under pressure to contain cost over and above promoting efficiency and others experience difficulties redirecting resources across inflexible financial structures. Such barriers have been expressed in terms of the savings identified in economic evaluations being unrealisable in practice. Health economists are then accused of being ill informed on structural aspects of health systems.

Overall, the literature reveals a growing realization that interventions by health economists in the area of research utilization have neither addressed the totality of factors which influence policy makers nor accounted for the complexity of health care decision making processes.

Prescriptions for Improvement

Typically, the published research draws on a similar range of potential solutions to the problem of low levels of usage. These include the need to standardize and improve methods of CEA and to increase the available evidence base for decision makers both in terms of volume and timeliness. A strong strand within prescriptions for greater usage focused on education and training for decision makers so that CEA can be better accessed, understood and applied.

Overall, responses to reported barriers tended to centre on questions of how research by health economists can be made more useful and accessible to policy makers. Prescriptions for overcoming accessibility barriers usually involve a combination of increasing resources, improving the means of communication with decision makers, and providing decision makers with training in interpreting health economics.

However, it is less clear from the literature how barriers relating to organizational and political context are to be addressed. There is little, for example, by way of prescriptions for shaping the health care system in order to incentivize and facilitate the use of CEA. Indeed, one study author, McDonald (2002), is pessimistic as to the appropriateness of seeking to

increase the use of CEA. Her argument is that, as a result of the complex and sometimes perverse structures of the English NHS, it is unhelpful to prescribe rational frameworks for NHS decision makers because this serves only to highlight to decision makers the gap between the rationalist ideal and the structural and political reality of the system.

Further National Institute for Health and Clinical Excellence Reflections

This part of the article draws on the authors' qualitative empirical work looking at the challenges for NICE in making full use of economic evaluations. Although issues of accessibility, broadly speaking, are not acute at the national level in the UK, organizations like NICE still have some important issues to address in this field. The NICE Appraisals Committee is in the highly unusual situation of having, for every topic they consider, an economic analysis undertaken specifically for their purposes. Thus, they avoid the frequently cited problems encountered by those working at a local level in the NHS of not being able to access cost-effectiveness (CE) information in a timely manner.

In terms of the challenge of interpreting CEAs, the qualitative study uncovered poor levels of understanding of CE information. The extent to which this is a serious barrier depends, to some extent, on the role NICE Committee members are expected to play and the overall approach to decision making being adopted. If all Committee members have a vote on the policy decision then they all need to understand all relevant information presented, including the CEA. A failing on the part of analysts that was revealed from the authors' research concerned the presentational style of CE studies. The highly technical nature of the CE studies being undertaken for NICE, and their presentational style, make for difficulties in understanding for the noneconomist. The need for improvements in the presentation of CE studies was a strong message from the authors' work.

A commonly cited acceptability concern with the CEAs is that they fail explicitly to consider the opportunity costs of the decisions being made. In the authors' research this was raised by a number of committee members including both health economists and health care managers. The CEA at NICE typically presents the problem in terms of a one-off decision concerning the coverage of a given health technology, commonly a new drug. No explicit consideration is therefore given to the sacrifice that would be required in order for the additional resources to be made available (assuming that the incremental cost is positive). An attempt to negate this problem involves use of a CE threshold, and defining technologies that have ICERs that fall below the threshold as cost-effective uses of NHS resources (regardless of their true opportunity cost). This issue has been highlighted by other commentators. However, although the necessity of using a CE threshold was acknowledged by most of the authors' research subjects, it was also viewed as problematic because the basis for the threshold value or range is very unclear.

In summary, the data from the authors' qualitative work with NICE suggest that for analyses to be viewed as acceptable, it is necessary that they provide information: (1) that end-users see as relevant (i.e., providing data on parameters that are likely to influence the decision of the policy maker), (2) that is appropriate to the decisions being faced, taking into account relevant contextual factors (e.g., budgetary arrangements commonly seen in the NHS), and (3) that can inform implementation of decisions in a complex decision making environment.

The Research-Practice Divide

This article has explored some of the reasons for the moderate impact of economic evaluation on health policy. There is little dispute that such findings are a source of concern to the discipline of health economics and that for such analyses to be a valuable decision making tools then change of some form is required. Commentators have identified weaknesses in methodologies adopted in economic analyses and there have been concerted attempts to improve their quality through, for example, the development of methodological standards. Difficulties in implementation may also derive from limits to the generalizeability of studies, resulting from factors such as: variations in disease epidemiology, relative prices, levels of health care resources, organizational arrangements, and clinical practice patterns.

However, one of the most challenging issues is contextual and relates to the difficulty in implementing hypothetical savings predicted by CEAs. It has been noted that the erroneous assumption of incremental divisibility of interventions and their benefits underpins many CEAs. Adang *et al.* (2005) have developed checklists to address the issue of reallocating resources within a real world context in order to get better information as to whether savings can indeed be made.

Important as these developments undoubtedly are, they also need to be accompanied by a concerted attempt to understand the differences in respective domains of 'research' and 'practice'. Much valuable work has been done on techniques for reducing or bridging the gap between the 'two communities' of researchers and decision makers. A review of studies by Innvaer *et al.* (2002) suggests that 'personal contact' between researchers and decision makers is one of the most commonly reported facilitators of evidence-based decision making. Lavis *et al.* (2003) argue that such interaction enables researchers to improve the production of analyses although simultaneously enhancing their adaptation by policy makers. However, these prescriptions for closer contact between researchers and decision makers also need to avoid naivety: it has been seen that other barriers exist. Also, incentives and rewards for researchers are less likely to recognize the value of incremental influence than they are outcomes that have a more direct influence on policy formation. In other words, the academic institutional environment in which economic evaluations are produced is not always conducive to such an interactive approach.

Much of the health economics literature to date has concentrated on barriers of accessibility of CEA results. This suggests a view that improvement in the process by which evaluations are communicated to decision makers, and the latter's capacity to understand their recommendations, ought to be the focus of attention and activity if impact is to be maximized. In other words, the emphasis is on tweaking the process

at both ends in order to support rational implementation of research findings. A focus on barriers to the acceptability of economic evaluation directs us away from such an approach. Instead, it is seen that there is substantive disjuncture between researchers and decision makers in terms of objective functions, institutional contexts and professional value systems. The literature in this area charts a growing realization of the conditions and contingencies of the health decision making environment. There has been a move away from an assumption of policy involving simple, rational choices to a realization of an interactive process with competing aims and considerations. Issues such as system rigidities, value conflict and competing objectives are difficult to overcome as this requires broader changes to the macropolitical and institutional environment of health care policy making.

A More 'Positive' Approach?

In contrast to the default normative approach taken in economic evaluation in health care, a positive analysis would simply generate information on the likely costs and benefits associated with alternative courses of action. Dowie (1996) describes such research as knowledge-generating, as opposed to decision-making. A distinguishing feature of positive analyses is that there is no *a priori* objective specified. Such analyses might involve the use of profile or cost consequence approaches to reporting results. This is where the predicted impacts of the intervention in question are detailed, possibly in a tabular form, without any attempt to summarize or aggregate across different dimensions. Kernick (2000) is a strong advocate of such an approach:

> Cost consequence analysis emphasises the importance of presenting data on costs and benefits in disaggregated form, implying a recognition of the value judgement from decision makers and an acceptance that benefits and disadvantages cannot always be condensed into a single output measure.
>
> Kernick (2000, p.314)

Traditional economic evaluation work evokes a conception of research utilization defined by Weiss (1979) as the 'problem-solving model'. In this model empirical and analytical evidence is applied directly to a policy problem and supplies the information required to enable the optimal solution to be identified and implemented. For the problem-solving model to apply, the recommendations of a normative economic analysis, for example, would need to be implemented directly by the relevant policy maker and would be seen as the driving force behind the decision reached. As Weiss (1979) indicates:

> … when this imagery of research utilisation prevails, the usual prescription for improving the use of research is to improve the means of communication to policy makers.
>
> Weiss (1979, p.428)

However, there are a number of weaknesses with the problem-solving model. For example, some have called into question the likelihood of establishing a single, agreed objective. Although many economists may adopt a normative view that the problem-solving model has much to recommend it, it has to be recognized that, the real world rarely lives up that aspiration. For example, in a review of UK studies into factors effecting evidence-based policy-making, Elliott and Popay (2000) conclude that many policy problems are often intractable or not clearly enough delineated to be tackled directly and comprehensively. They also find that research evidence is frequently unlikely to be sufficiently clear-cut and unambiguous to translate directly into policy. They also call into question the assumption of a straightforward policy process in the problem-solving model and conclude that dissemination of health services research results has been hampered by a preoccupation with the rational, problem-solving model. In these circumstances, Weiss's 'interactive' model of research utilization, in which policy formulation is understood as a nonlinear process involving multiple agents and influences, has far greater descriptive validity.

The distinction between problem-solving and interactive models of research utilization correlates, to some extent, with the binary of normative and positive approaches to health economic analyses. The requirement for agreement of purpose and objectives between researcher and decision maker is a defining premise of both normative economic evaluation and problem-solving conceptions of policy research utilization. Positive approaches to evaluation, however, may be seen as more helpful to decision makers involved in policy processes that are marked by interaction and competing or multiple objectives. An understanding by the analyst of the nature of the policy environment into which the analyses are being placed is required. This will allow more informed choice to be made concerning the appropriate approaches to analysis and presentation of results.

In highlighting the failure of health economists to consider issues of the acceptability of the data they generate, Kernick (2000) argues that:

> The history of any movement determines its structure and the way in which meaning is generated within it. Health economists tend to adopt a straightforward view … Just as the NHS was configured in part to reflect the needs of doctors and not patients, the development of health economics was set to reflect the requirements of the academic discipline and not the realities of the emerging healthcare environment.
>
> Kernick (2000, p.312)

Conclusions

And so to conclude, the driving force behind the push to make more use of economic analyses in health care resource allocation decisions is the desire to make decision processes, and the decisions themselves, more rational. In turn, greater rationality in the system contributes to openness and transparency, and so necessitates that the information on which decisions are based is accessible to a wide audience – the more accessible the information used in decision-making, the easier it is to be inclusive in the decision-making process and the more transparent is the basis on which the decision is made.

This accessibility concern represents one of the challenges to the health economics community in terms of producing

evidence that is more reflective of real world practices but also highlights a potential training agenda: clinical and managerial decision makers in health care require some level of expertise and understanding of economic evaluation in order to provide input into the decision making process. Additional areas of focus for health economists include the need to overcome perceived weaknesses in the methods of their analyses, and the need to work with those at the front-line in health care to ensure alignment between the health maximization objectives often assumed in economic analyses and the broad range of other objectives facing decision-makers in reality. That is not to suggest that the decision-maker always 'knows best' but analyses based on false assumptions regarding objectives serve no purpose.

See also: Analysing Heterogeneity to Support Decision Making. Budget-Impact Analysis. Cost-Effectiveness Modeling Using Health State Utility Values. Cost–Value Analysis. Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties. Economic Evaluation of Public Health Interventions: Methodological Challenges. Efficiency in Health Care, Concepts of. Health and Its Value: Overview. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Information Analysis, Value of. Managed Care. Measuring Equality and Equity in Health and Health Care. Multiattribute Utility Instruments: Condition-Specific Versions. Observational Studies in Economic Evaluation. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Priority Setting in Public Health. Problem Structuring for Health Economic Model Development. Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation. Quality-Adjusted Life-Years. Searching and Reviewing Nonclinical Evidence for Economic Evaluation. Statistical Issues in Economic Evaluations. Synthesizing Clinical Evidence for Economic Evaluation. Time Preference and Discounting. Value of Drugs in Practice. Value-Based Insurance Design. Valuing Health States, Techniques for. Valuing Informal Care for Economic Evaluation. Welfarism and Extra-Welfarism. Willingness to Pay for Health

References

Adang, E., Voordijk, L., van der Wilt, G. and Ament, A. (2005). Cost-effectiveness analysis in relation to budgetary constraints and reallocative restrictions. *Health Policy* **74**, 146–156.

Bryan, S., Williams, I. and McIver, S. (2007). Seeing the NICE side of cost-effectiveness analysis: A qualitative investigation of the use of CEA in NICE technology appraisals. *Health Economics* **16**, 179–193.

Clement, F. M., Harris, A., Li, J. J., et al. (2009). Using effectiveness and cost-effectiveness to make drug coverage decisions. *Journal of the American Medical Association* **302**(13), 1437–1443.

Culyer, A. J. (1973). *The economics of social policy.* London: Martin Robertson and Company Ltd.

Dowie, J. (1996). The research-practice gap and the role of decision analysis in closing it. *Health Care Analysis* **4**, 5–18.

Elliott, H. and Popay, J. (2000). How are policy makers using evidence? Models of research utilisation and local NHS policy making. *Journal of Epidemiology and Community Health* **54**, 461–468.

Innvaer, S., Vist, G., Trommald, M. and Oxman, A. D. (2002). Health policy-makers' perceptions of their use of evidence: A systematic review. *Journal of Health Services Research and Policy* **7**(4), 239–245.

Kernick, D. P. (2000). The impact of health economics on healthcare delivery. *PharmacoEconomics* **18**(4), 311–315.

Lavis, J. N., Robertson, D., Woodside, J. M., McLeod, B. and Abelson, J. (2003). How can research organizations more effectively transfer research knowledge to decision makers? *Milbank Quarterly* **81**(2), 221–248.

McDonald, R. (2002). *Using health economics in health services: Rationing rationally?* 1st ed. Buckingham: Open University Press.

National Institute for Health & Clinical Excellence (2003). *The clinical effectiveness and cost effectiveness of anakinra for rheumatoid arthritis.* London, UK: NICE.

National Institute for Health & Clinical Excellence (2006). *Statins for the prevention of cardiovascular events in patients at increased risk of developing cardiovascular disease or those with established cardiovascular disease.* London, UK: NICE.

Rawlins, M. D. and Culyer, A. J. (2004). National Institute for Clinical Excellence and its value judgments. *British Medical Journal* **329**, 224–227.

Weiss, C. H. (1979). The many meanings of research utilization. *Public Administration Review* 426–431.

Williams, I., McIver, S., Moore, D. and Bryan, S. (2008). The use of economic evaluations in NHS decision-making: A review and empirical investigation. *Health Technology Assessment* **12**(7), 1–193.

Further Reading

Devlin, N. and Parkin, D. (2004). Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Economics* **13**, 437–452.

Hoffmann, C., Stoykova, B. A., Nixon, J., et al. (2002). Do health-care decision makers find economic evaluations useful? The findings of Focus Group Research in UK Health Authorities. *Value in Health* **5**(2), 71–78.

Schlander, M. (2008). The use of cost-effectiveness by the National Institute for Health and Clinical Excellence (NICE): No(t yet an) exemplar of a deliberative process. *Journal of Medical Ethics* **34**, 534–539.

von der Schulenburg, J. M. G. (2000). *The influence of economic evaluation studies on health care decision-making.* Oxford: IOS Press.

Spath, H. M., Allenet, B. and Carrere, M. O. (2000). Using economic information in the health sector: The choice of which treatments to include in hospital treatment portfolios. *Journal d'Economie Medicale* **18**(3–4), 147–161.

Williams, I. and Bryan, S. (2007). Understanding the limited impact of economic evaluation in healthcare resource allocation: A conceptual framework. *Health Policy* **80**, 135–143.

Williams, I., Bryan, S. and McIver, S. (2007). Health technology coverage decisions: Evidence From The N.I.C.E. 'experiment' in the use of cost-effectiveness analysis. *Journal of Health Services Research and Policy* **12**(2), 73–79.

Relevant Websites

http://www.cadth.ca/
The Canadian Agency for Drugs and Technologies in Health.

http://www.crd.york.ac.uk/CRDWeb/AboutNHSEED.asp
The Centre for Reviews and Dissemination at the University of York.

https://research.tufts-nemc.org/cear4/default.aspx
The Center for the Evaluation of Value and Risk in Health at Tufts University Medical Center.

http://www.hta.ac.uk/pdfexecs/summ1207.pdf
The Health Technology Assessment Programme of the National Institute for Health Research.

http://www.nice.org.uk/
The National Institute for Health & Clinical Excellence.

Advertising as a Determinant of Health in the USA

DM Dave, Bentley University, Waltham, MA, USA
IR Kelly, Queens College of the City University of New York, Flushing, NY, USA

Overview

Advertising is ubiquitous, found on television and radio, newspapers and magazines, mail and flyers on the windshield, billboards and sports arenas, and now on the computer, and virtually no one is immune to being exposed to it. The American Marketing Association defines marketing, of which advertising is a subset, as "the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large" (Grewal and Levy, 2009). To advertise itself is simply "the action of calling something to the attention of the public especially by paid announcements" (Merriam-Webster, 2011). What distinguishes an advertisement from other forms of marketing is that: (1) someone has paid to get the message shown; (2) the message must be carried by a medium; (3) legally, the source must be known; and (4) it represents a persuasive form of communication (Grewal and Levy, 2009). This article will provide a survey of economic views of advertising in general, which will provide the context for a better understanding of the relevance of advertising for health behaviors and health care markets.

Modern advertising began early in the twentieth century with the advent of Kellogg cereals and Camel cigarettes (Bittlingmayer, 2008). It is a huge industry, currently with over 14 000 establishments (Bureau of the Census US Department of Commerce, 2007) and over $200 billion in expenditures (Bittlingmayer, 2008).

Why consumers respond to advertising will be analyzed in more detail in the next section. It is this question that economists ultimately seek to answer in the context of the separate views – that advertising is persuasive, informative, or simply complementary to the advertised product. A brief survey of these different views of advertising is provided in this article, which can help frame the relevance and public health consequences of advertising for health behaviors and healthcare markets. The reader is referred to Bagwell (2007) for an excellent, comprehensive review of the economics of advertising, and also Schmalensee (1972) for an earlier take. Elements of each of these views exist in most industries, with variations across industry. A firm may generally view advertising as capital (albeit intangible) that depreciates over time (Bagwell, 2007). Most empirical studies find that most of the effects of advertising are short-lived and that most effects of advertising depreciate within a year. There has therefore been limited empirical evidence for the 'goodwill effect' in advertising, causing a firm's current advertising to be influenced by past advertising (Bagwell, 2007).

The nature of advertising has changed dramatically over time with the advent of new technology and media. Although the means of advertising in healthcare markets can vary across firms and industries (in part because of advertising restrictions), conventional media include magazines, newspapers, billboards, radio, television, and direct mail. With 77% of households using the Internet (Statistical Abstract of the United States, 2009), the computer has also emerged as an important medium for advertising. In addition, firms are also increasingly relying on product placement in movies and video games, and other forms of digital media. Although the volume may presumably diminish the individual effect of an advertisement because it is difficult for a potential consumer to focus on more than one ad at once, online advertising can more effectively tailor ads to individuals.

The number of establishments classified as 'advertising agencies' in 2007 was 14 355, up from 13 879 in 1992. Advertising expenditures rose from $2.1 billion in 1940 to $237.4 billion in 2002 (Bittlingmayer, 2008). Note that the North American Industrial Classification System (NAICS) code used by the Economic Census for advertising agencies (which do not include 'related services' such as public relations) is 541 810, corresponding to the standard industrial classification code used before 1997 of 7311. According to the Census, "[t]his industry comprises establishments primarily engaged in creating advertising campaigns and placing such advertising in periodicals, newspapers, radio and television, or other media. These establishments are organized to provide a full range of services (i.e., through in-house capabilities or subcontracting), including advice, creative services, account management, production of advertising material, media planning, and buying (i.e., placing advertising)." The intensity of advertising is often measured by the advertising-to-sales ratio. Advertising-to-sales ratios for industries relevant to our discussion are shown in **Table 1**. The advertising-to-sales ratio for the pharmaceutical industry, especially, understates the level of promotional efforts because it does not include other forms of promotion such as sampling to physicians and other

Table 1 Advertising expenditures as a percent of sales for selected industries, 2010

Industry	*Ad-to-sales ratio, 2010*
Distilled and blended liquor	14.4
Food and kindred products	11.5
Eating and drinking places	10.2
Beverages	6.1
Pharmaceutical preparations	4.2
Malt beverages	3.7
Wine, brandy and brandy spirits	3.3
Misc food preps, kindred products	2.8
Food stores	1.7
Meat packing plants	1.4
Grocery stores	0.8
Bakery products	0.3
All industries combined	2.1

Source: Adapted from Schonfeld & Associates (2010). *Advertising Ratios and Budgets.* June 1.

 doi:10.1016/B978-0-12-375678-7.00318-7

providers and direct marketing to providers. In 2005, the pharmaceutical industry spent 20% of its sales on promotional activities. The Dorfman-Steiner (1954) condition for optimal advertising gives some insight as to why certain industries (or firms) may engage in higher levels of advertising:

$$\text{Advertising/Sales} = \varepsilon_{QA}/\varepsilon_{QP}$$

The condition positively relates advertising intensity, as measured by the advertising-to-sales ratio, to the elasticity of sales with respect to advertising (ε_{QA}) and negatively to the elasticity of sales with respect to price (ε_{QP}), expressed in absolute magnitudes. Thus, the more price-inelastic is the good, the higher is its advertising intensity, *ceteris paribus*. Alcohol, tobacco, and prescription drugs, for instance, are found to be relatively price-inelastic, and these industries also devote a relatively greater fraction of their sales to advertising and promotion.

Advertising in healthcare markets is controversial, especially when it has been found to raise the overall market for unhealthy behaviors (for instance, smoking or junk food) or found to contain deceptive or misleading information. Thus, inevitably, advertising must have a certain degree of oversight. Federal agencies that regulate advertising include the Federal Trade Commission (FTC), the Federal Communications Commission (FCC), and the Food and Drug Administration (FDA). Other agencies such as the Bureau of Alcohol, Tobacco, and Firearms also play a role in regulating advertising (Grewal and Levy, 2009). The FTC, established in 1914, enforces the truth in advertising laws and identifies deceptive practices. The FCC, established in 1934, "enforces restrictions on broadcasting material that promotes lotteries; cigarettes, little cigars, or smokeless tobacco products; or that perpetuates a fraud." It also enforces laws to prohibit or limit obscene, indecent, or profane language (Grewal and Levy, 2009). The FDA, established in 1930, regulates labeling, health claims, and required disclosure statements. Many are unaware that advertising for weight loss products (discussed in Section 'Conceptual Framework') is not 'drug advertising' according to the FDA; as a dietary supplement, it is classified as a food and faces fewer standards than other drugs (Grewal and Levy, 2009; Cawley *et al.*, 2010).

The article is organized as follows. Section 'Conceptual Framework' provides a conceptual framework outlining the economic views of advertising. Advertising in several health markets – particularly those pertaining to tobacco, alcohol, food, soft drinks, cereal, weight loss products, and prescription drugs – is analyzed in Section 'Advertising in Health Markets'. Section 'New Directions' provides a glimpse into directions for future research in the area, particularly surrounding the advent of online advertising and drawing insights from neuroeconomics to the study of advertising. Section 'Summary' concludes.

Conceptual Framework

It is often presumed that the average consumer is responsive to advertising and promotion. However, one of the key questions with respect to advertising by firms in markets for healthcare inputs is whether advertising raises 'selective' or brand-specific demand versus 'primary' or industry-wide demand (Borden, 1942). The answer to this question has normative implications and relevance for public health. For instance, is advertising by the cigarette industry combative and solely reflective of a market share transfer or does it also lead to an overall expansion of the market? This was one of the disputes that was central to the litigation initiated in 1999 by the US Department of Justice (DOJ) against cigarette manufacturers. As a starting point, it is helpful to draw upon three principal views that have emerged with respect to why consumers may respond to advertising: (1) persuasive, (2) informative, and (3) complementary.

Chamberlin (1933) integrates advertising into his theory of monopolistic competition, observing that advertising can help firms to differentiate their products and generate an outward shift in firm-level demand. According to Chamberlin, advertising impacts demand by altering consumers' tastes and preferences. Under this 'persuasion' hypothesis, brand-level demand would not only shift outward in response to advertising but also become relatively less elastic, possibly leading to higher prices. Advertising-induced product differentiation and creation of brand capital may deter entry and enhance the monopolistic power of incumbent firms, especially if these established firms also enjoy scale economies in advertising and production (Kaldor, 1950). Thus, under the persuasion view, advertising can have significant anticompetitive effects, a point that was also emphasized by Robinson (1933).

Chamberlin (1933) also pointed to the transfer of information to consumers as another explanation for why consumers respond to advertising. This informative view of advertising took on a formal expression in Ozga (1960) and Stigler (1961). In markets characterized by imperfect information, advertising can effectively reduce search costs by conveying direct or indirect information to consumers regarding the existence, quality, price, and other attributes of products. As Bagwell (2007) noted, in such markets, advertising emerges as an endogenous response and solution to the information asymmetry. In contrast to the persuasive view, advertising plays a more constructive role under the informative view, and may also have pro competitive effects. As consumers receive low-cost (relative to incurring search costs) information on products and brands, the firm's demand becomes relatively more elastic and price dispersion in the market is reduced. Advertising can thus promote competition among incumbent firms and facilitate the entry of new firms as well as the introduction of new products.

Nelson (1974) contended that even when advertising does not hold direct information content, it may still signal indirect information regarding product quality and firm attributes. For instance, advertising can signal that a firm is an efficient producer because these firms would benefit the most from expanding demand. Advertising can also enhance the match between products and buyers in markets where consumers have heterogeneous valuations. And, advertising may help consumers recollect their previous experience with the product and lead to repeat-business. Because this effect is more valuable for firms producing high-quality products, advertising may thus indirectly signal quality even for new consumers.

Nelson (1970) distinguished between search goods, wherein the consumer can determine quality before purchase though perhaps after incurring some search costs, and experience goods, wherein the consumer can assess quality only after consumption. Advertising addresses an informational imbalance for experience goods by providing indirect information content regarding quality, and advertising intensity is thus predicted to be higher for experience goods. In contrast, advertising for search goods (for instance, eyeglasses, consumer electronics, or credit cards) would be focused on providing direct information regarding price, location, availability, and product attributes.

Darby and Karni (1973) also found it useful to distinguish a third category of goods that have 'credence' attributes, for which the consumer is unable to accurately evaluate quality even post consumption. This market failure of imperfect information for experience and credence goods also potentially gives firms an incentive to engage in misleading advertising claims (Darby and Karni, 1973; Nelson, 1974). Where market-based mechanisms are unable to deter deceptive advertising, there is a role for government regulation and publicly funded dissipative counter-advertising.

Although the persuasive and informative views provide conflicting assessments of the role of advertising, the third view of advertising provides a framework under which advertising is complementary to the advertised product. That is, advertising does not need to exert any direct influence on consumer preferences, and it may or may not possess information content. Within a household production framework, Stigler and Becker (1977) modeled the advertised product with its associated advertising expenditures as inputs into the production function for each final commodity, implying a complementarity between the advertised product and its advertising. Under this framework, a higher level of advertising can raise demand because the consumer now believes that he can obtain a greater output of the final commodity from a given input of the advertised good. In a related but separate framework, Becker and Murphy (1993) directly modeled advertising as an input into the individual's utility function. Advertising raises demand in this framework by increasing the marginal utility of the advertised good. Note that this complementarity follows from the fact that there does not exist a separate market for advertising messages – considerable transactions and monitoring costs make it infeasible to separately sell advertising to consumers.

Both of these paradigms, which impart a complementary role to advertising, also bridge back to the informative view. For instance, if advertising enables consumers to produce information at lower cost (Verma, 1980), then consumers can indeed more efficiently convert market goods into valued final commodities, as assumed by Stigler and Becker (1977). And, even if advertising in uninformative, it may still play a constructive role because consumers may value it directly, as assumed by Becker and Murphy (1993).

The upshot of this discussion is that no single view of advertising is applicable in every setting. Furthermore, from a public health standpoint, the debate centers around whether advertising reflects a brand-switching process or a market expansion process, especially in relation to the market for unhealthy inputs such as cigarettes, underage drinking, and junk food – or in different terms, whether advertising is combative (predatory) or cooperative. Because advertising can affect both selective (brand-centric) as well as primary (market) demand under all three views, the question cannot be resolved based on theory alone and empirical evidence needs to bear upon the specific demand effects of advertising in various markets. With that said, markets for most healthcare inputs have some predominant experience attributes – such as tobacco and alcohol products, over-the-counter (OTC) and prescription medications, and snacks and beverages. Thus, advertising intensity for many of these goods tends to be higher relative to the average industry (2.1%; see **Table 1**). These views of advertising also highlight potential effects on price, which depend on the extent to which advertising expenditures raise operating costs, affect price elasticity of demand, and allow firms to take advantage of scale economies. Finally, the concentration effects of advertising – that is, whether it facilitates entry or whether it augments the monopoly power of established firms – depends on whether advertising is purely persuasive in nature and leads to spurious brand differentiation or whether it redresses imperfect information and makes demand more elastic.

Advertising in Health Markets

Tables 1 and **2** suggest that, with the exception of restaurants that tend to be more monopolistically competitive, industries that more heavily advertise generally tend to be more concentrated, with Herfindahl–Hirschman indices of at least 1000 (characteristic of mild concentration) or four-firm concentration ratios of at least 80% (characteristic of very concentrated industries). Scale economies in advertising exist, and larger firms are better able to spend on advertising. Studies by Kaldor and Silverman (1948) and Doyle (1968) supported the notion that advertising intensity and concentration are highly linked, leading to an oligopolistic structure (Bagwell, 2007). Nelson (1975) found a significant relationship between advertising intensity and concentration for search goods but not for durable goods or nondurable experience goods. The markets for tobacco, alcohol, food, soft drink, weight loss products, and prescription drugs are analyzed below in more detail.

Advertising of Tobacco

Rather than compete directly on price, firms in highly concentrated industries such as the cigarette industry often use advertising to differentiate their brands and increase sales. In 2005, cigarette manufacturers spent $13.1 billion (or approximately 10% of their sales) on advertising and promotion, making cigarettes among the most heavily advertised and promoted products in the US. As reported in **Table 3**, this level also represents a 111% increase in total marketing expenditures over the past decade. Cigarette manufacturers had relied heavily on television advertising in the 1960s, though the application of the Fairness Doctrine to cigarette advertising in 1967 and the mandated antismoking messages subsequently reduced the commercial value of televised ads. Following a

Table 2 Concentration ratios and Herfindahl–Hirschman indices for select industries, 2007

2007 NAICS Code	*Industry*	*Companies*	*Four-firm concentration ratio*	*HHI*
312 221	Cigarette manufacturing	20	97.8	na
3 122	Tobacco manufacturing	73	89.6	na
31 212	Breweries	373	89.5	na
311 221	Wet corn milling	33	83.8	2338.20
311 222	Soybean processing	68	81.5	1930.80
31 123	Breakfast cereal manufacturing	35	80.4	2425.50
311 821	Cookie and cracker manufacturing	303	69.3	1607.20
31 122	Starch and vegetable fats and oils manufacturing	195	67.2	1476.20
31 131	Sugar manufacturing	37	59.9	1097.50
312 111	Soft drink manufacturing	259	58.1	1094.50
31 191	Snack food manufacturing	470	53.2	1984.10
31 192	Coffee and tea manufacturing	337	43.3	763.1
325 412	Pharmaceutical preparation manufacturing	763	34.5	456.8
3 115	Dairy product manufacturing	1 073	23.5	290.7
3 114	Fruit and vegetable preserving and specialty food manufacturing	1 248	21.7	192.5
311	Food manufacturing	21 355	14.8	102.1

Abbreviation: na, not applicable.

Source: Adapted from US Census Bureau (2007). Concentration ratios. Available at: http://www.census.gov/econ/concentration.html (accessed 09.02.13).

Table 3 US Cigarette advertising and promotion activities (thousands of 2005 $)

Category	*1995*	*2000*	*2005*	*Growth (%) 1995–2005*
Newspapers	$24 241	$57 951	$1589	– 93
Magazines	$315 469	$330 881	$44 777	– 86
Outdoor	$346 928	$10 392	$9 821	– 97
Transit	$28 578	$4	$0	– 100
Point-of-sale	$328 383	$389 360	$182 193	– 45
Total advertising	**$1 043 599**	**$788 588**	**$238 380**	**– 77**
Promotional Allowances (paid to retail outlets for favorable product positioning)	$2 365 124	$4 391 314	$847 686	– 64
Sampling distribution (provision of free samples to the public)	$17 540	$25 053	$17 211	– 2
Specialty item distribution (provision of other free accessories)	$843 251	$367 805	$230 534	– 73
Public entertainment (cost of event sponsorship)	$140 297	$347 367	$244 802	74
Direct mail	$43 886	$104 232	$51 844	18
Coupons and retail value added (promotional price reductions, bonus cigarettes, other bonus)	$1 709 361	$4 665 909	$11 378 742	566
Other promotional activities (includes endorsements and internet promotions)	$42 697	$72 191	$101 759	138
Total promotion	**$5 162 156**	**$9 973 870**	**$12 872 578**	**149**
Total advertising and promotion	**$6 205 755**	**$10 762 458**	**$13 110 958**	**111**

Source: Adapted from Federal Trade Commission, Cigarette Report for 2006. Available at: http://www.ftc.gov/os/2009/08/090812cigarettereport.pdf (accessed 09.02.13).

voluntary industry ban in 1970, cigarette broadcast advertising was officially banned by the Public Health Cigarette Smoking Act starting in 1971. Advertising practices were further restricted by the 1998 Tobacco Master Settlement Agreement (MSA), which also banned most forms of outdoor advertising. Cigarette advertising in magazines with youth readership increased dramatically post-MSA, but then later fell after public pressure (Hamilton *et al.*, 2002) (also see **Table 3**). Since 1970, and particularly accelerating after the MSA, firms' total marketing budget has shifted away from media-based advertising in favor of other promotional activities (such as coupons, added bonuses, promotional allowances, and event sponsorships). There was also a proliferation of cigarette brands over this period in an effort by firms to segment the market and thereby enhance their monopolistic power. The Family Smoking Prevention and Tobacco Control Act, signed into law in 2009, currently gives the Food and Drug Administration (FDA) authority to regulate the content, marketing, and sale of tobacco products.

Saffer (2000) noted that advertising by the cigarette industry is "designed to create a fantasy of sophistication, pleasure, and social success" and generate a product personality that will appeal to specific market segments. In other words, such advertising contains persuasive attributes and could raise demand by generating potentially spurious brand differentiation. Consistent with this persuasive view of advertising, Brown (1978) found decreasing average costs and increasing returns to advertising capital with sales, implying

that advertising potentially creates substantial barriers to entry in the cigarette industry.

Given the external costs of smoking and related public health concerns, the key debate has understandably centered on whether and the extent to which cigarette advertising and promotion raise total cigarette consumption and expand the overall market. There is a large literature that has evaluated the effects of tobacco advertising and promotion on consumption outcomes. Rather than survey this literature (Chaloupka and Warner, 2000), the main findings and issues that have emerged from these studies are reviewed below. Empirical studies have been challenged in trying to isolate a marginal change in consumption when advertising and promotional activities of tobacco companies are at or close to the point of saturation (Ross and Chaloupka, 2002) and have produced mixed findings. Consider the advertising response function shown in **Figure 1**, which can apply to the national or local market level, and to the industry as a whole or at the brand level. Because of diminishing marginal product, the function flattens out at some point and consumption becomes increasingly less responsive to advertising. Diminishing returns may be unavoidable because the effectiveness of additional advertising will decrease once the most responsive buyers have already been reached. In the context of the informative view of advertising, as an increasing number of potential buyers receive information regarding the advertised product, additional advertising is less effective because an increasingly greater proportion of individuals who are exposed to the ads are already familiar with the product.

Earlier studies generally relied on annual or quarterly aggregated data at the national level and find either no effects or very small positive effects of advertising on cigarette consumption. This is perhaps to be expected because loss of variance at such a high level of aggregation makes it difficult to reliably identify effects. As cigarettes are heavily advertised and promoted, the marginal product of aggregate national advertising (measured at a range around A_1 in **Figure 1**) may be very small or zero. Estimates based on a single time-series of aggregate national data are also likely confounded with unobserved trends and the simultaneity between advertising and sales.

Subsequent studies based on local or individual-level cross-sectional or panel data are more indicative of advertising-induced primary market-expansion effects. These studies typically use local-level (for instance, gathered at the level of the state or metropolitan statistical area) advertising data, which have greater (and plausibly more exogenous) variation owing to differences in advertising costs across markets and because of pulsing (which is a burst of advertising, in a specific market, that lasts for a short time and then stops). Goel and Morey (1995), for instance, used annual state-level data spanning 1959–82 and found significant effects of lagged cigarette advertising on consumption. Roberts and Samuelson (1988) developed a model of non price competition for an oligopolistic industry and applied it to their study of the cigarette market, utilizing data for six firms spanning 1971–82. They concluded that "advertising primarily affects the size of market demand and does not alter firm market shares" (p. 215). In a study using individual-level data on 6700 youth, combined with measures of televised cigarette advertising, counter-advertising, and self-reported time spent watching television, Lewit *et al.* (1981) found that smoking ads on television are significantly associated with higher youth smoking.

Studies that examine the impact of advertising bans provide further evidence on whether cigarette advertising expands the overall market. These studies also bypass some of the limitations stemming from the simultaneity between advertising intensity and sales. However, the passage of advertising restrictions may not be strictly exogenous and depends on past trends in smoking prevalence. If advertising only leads to brand-switching with no primary effects on market demand, then advertising restrictions should not have any effects on consumption. Banning advertising on certain media would potentially shift the advertising response function downward, as shown in **Figure 1**. Even if an advertising ban does not reduce the total level of advertising, it will reduce the average and marginal effectiveness of advertising as firms substitute from the banned media to the non banned media. Increased use of non banned media reduces average and marginal effectiveness because of diminishing marginal product. If firms try to compensate for the advertising ban by increasing total advertising expenditures, this would correspond with a

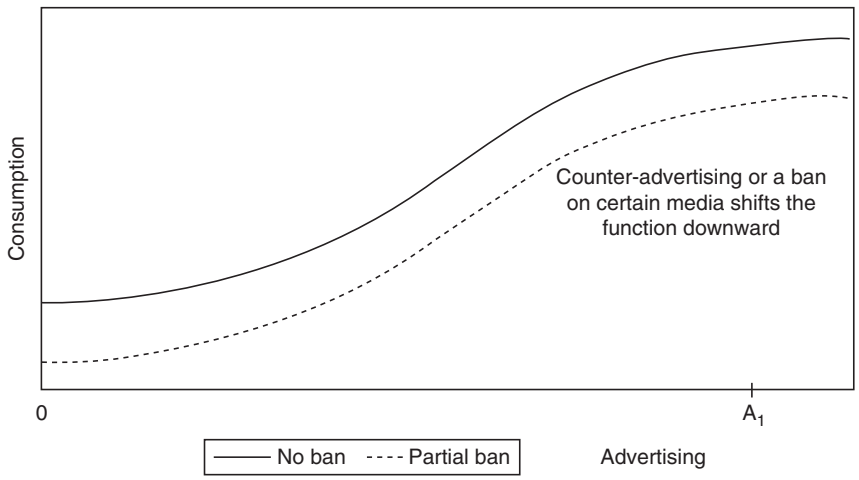


Figure 1 Advertising response function.

movement to a higher level of advertising on the lower advertising response function in **Figure 1**. **Table 3** provides some evidence that this may be the case for the cigarette industry. Consistent with an advertising-induced market expansion effect, Goel and Morey (1995) found that the broadcast ban on cigarette advertising lowered consumption. Saffer and Chaloupka (2000) studied the effects of tobacco advertising bans on tobacco consumption in 22 high-income countries over the period from 1970 to 1992. They found that although a limited set of advertising bans has little or no effect (because firms have many remaining media options), a comprehensive set of media bans can reduce tobacco consumption by 6–7%.

Cigarette brands may also have some credence attributes – wherein the consumer is not able to fully assess the product quality even after consumption. This can provide an incentive for firms to engage in potentially misleading advertising. For instance, the US Department of Justice maintained in a lawsuit filed in 1999 that the cigarette manufacturers falsely marketed and promoted their low-tar and light cigarette brands as being less harmful than conventional cigarettes. The consumer may be persuaded by these claims and would not be able to judge their veracity even post consumption at least over the short-term.

Given the possible market expansion effects of cigarette advertising and the presence of such misleading or imperfect product information, antismoking advertisements (or counter-advertising) have been undertaken by the public sector. Between 1967 and 1970, the Fairness Doctrine required broadcasters to donate air time to antismoking ads. At their peak, the ratio of antismoking ads to smoking ads was one-third (Saffer, 2000). Funds from the 1998 Master Settlement Agreement further provided for many state-initiated anti-tobacco campaigns. Studies have generally found such counter-advertising to be effective in reducing cigarette consumption. Emery *et al.* (2005), for instance, studied individual exposure to antitobacco advertising across the largest 75 media markets in 48 states between 1999 and 2000. They concluded that state-sponsored counter-advertising is associated with greater antitobacco sentiment and reduced smoking among youth. Interesting content analyses by Goldman and Glantz (1998) have suggested that the most effective antismoking messages focus on the tobacco industry's manipulation of its customers and the least effective are ads that portray smoking as unhealthy. This suggests that health-related messages currently may not be conveying any new information to consumers, and that the effectiveness of antismoking messages may derive from their directly counteracting the persuasive qualities of smoking ads and moderating the complementarity between smoking and smoking ads (for instance, through smoking ads portraying social prestige).

Advertising is also highly prevalent for products aimed at helping consumers quit smoking, such as nicotine-replacement therapy. Smoking-cessation products can be classified as experience goods because the consumer needs to use them before being able to assess their efficacy, and theory predicts a relatively high advertising intensity for experience goods. Avery *et al.* (2007) studied the market for such smoking cessation products and noted that the industry spent between 10% and 20% of its sales on advertising. They specifically studied the effects of magazine advertising of such products using individual-level data matched with salient individual-level measures of advertising exposure, paying careful attention to endogeneity concerns, and found that smokers who are exposed to more advertising are more likely to attempt to quit and are more likely to have successfully quit. Adopting the same identification strategy, Dave and Saffer (2013) also found that magazine advertising for smokeless tobacco (ST) products, which is one of the few conventional media available for manufacturers following bans in other media, leads to a higher probability of using ST. ST, which is safer than smoking though not completely safe, is also sometimes used as a cessation aid by smokers. Hence, the debate centers on the potential role of ST use and ST marketing as tools in an overall tobacco harm-reduction approach.

There is some indirect evidence on the competitive effects of advertising in the cigarette market. Brown (1978) found decreasing average costs and increasing returns to advertising, and concluded that advertising may create barriers to entry, based on data that preceded the 1970 television ban. Eckard (1991) utilized the television advertising ban as a natural experiment to study the effects of advertising, and found that concentration within the industry actually increased after the ban. This is in line with Thomas (1989), who found decreasing returns to scale with respect to advertising in the cigarette market, thus yielding a potential advantage to smaller firms with multiple brands. Indeed, the extent of brand proliferation and brand-level competition in the cigarette market is consistent with this finding.

In summary, the role of advertising in tobacco markets is controversial. The public health community contends that such advertising encourages smoking and particularly influences experimentation and smoking initiation among youth. The tobacco industry maintains that their advertising only affects selective demand through brand-switching and does not influence the overall size of the market. Manufacturers also suggest that their advertising provides important information content, for instance, regarding tar and nicotine (Chaloupka and Warner, 2000). Although earlier studies did not find significant market-level effects of cigarette advertising, more sophisticated analyses seem to indicate that advertising does impact primary demand. Further evidence gleaned from studies of advertising restrictions, antismoking ads, and advertising of smoking cessation products is also consistent with this market expansion effect. These studies also point to potential avenues through which advertising impacts the overall market demand, and these pathways are consistent with all three views of advertising discussed in Section 'Conceptual Framework.'

Advertising of Alcohol

Similar to the tobacco industry, the alcohol industry in the US is highly concentrated (see **Table 2**). The US brewing industry, for instance, is dominated by three firms, which account for almost 80% of beer sales. Beer brewers spent approximately $975 million in 2007 on advertising, with the top three firms accounting for 72% of these expenditures. Total advertising and promotional spending for all alcohol companies are on

the order of $4 billion (Jernigan and O'Hara, 2004). Advertising by the alcohol industry aims at raising sales through brand differentiation and customer loyalty, and advertising practices are self-regulated, primarily following a set of industry standards. For instance, industry guidelines allow alcohol-related ads to be placed in media where at least 70% of the audience is above the legal drinking age. Advertising messages also cannot directly appeal to under age youth. Some major broadcast networks adhere to a self-imposed ban on liquor advertising, though there are no such restrictions on cable networks.

The issues relating to the promotion and advertising of alcoholic beverages are similar to those discussed above with respect to tobacco, but with one exception. Unlike smoking, the majority of drinkers consume alcohol safely with little external harm. Thus, from a public health standpoint, the key debate with respect to market expansion has centered on problem drinking, which imposes considerable external costs (for instance, motor vehicle fatalities), and centered on the effects of advertising on youth drinking. On both of these fronts, although some studies have indicated that alcohol advertising is associated with more problem drinking and more underage drinking, the evidence is far from conclusive.

Anderson *et al.* (2009) reviewed 16 longitudinal studies that assessed adolescents' exposure to media-based advertising and their drinking behavior. They concluded in favor of evidence suggesting that exposure to advertising messages is associated with a higher likelihood that the adolescent will initiate drinking, and associated with higher drinking among baseline drinkers. Many of these reviewed studies, however, are based on small, often nonrepresentative, samples and utilize measures of recalled exposure to ads, which may be potentially confounded with unobserved predisposition toward drinking or pro drinking sentiment.

Saffer and Dave (2006) utilized cross-sectional data from the Monitoring the Future Surveys and longitudinal data from the National Longitudinal Survey of Youth (1997 cohort), both nationally representative, to study the effects of probable advertising exposure on adolescent drinking behavior. They bypassed the problems associated with self-recalled advertising exposure and instead exploited variation across and within markets with respect to the level of alcohol advertising in broadcast and print media. Estimates indicate significantly positive but relatively small effects of media advertising on alcohol participation and binge drinking (elasticity estimates of approximately 0.09 and 0.17, respectively), though there is some heterogeneity in this response across gender and racial groups. The authors simulated the effects of a 28% reduction in total alcohol advertising (based on the range observed in their data) and concluded that the reduction in advertising could decrease adolescent binge drinking from 12% to approximately 10% and decrease monthly alcohol participation from 25% to approximately 23%.

Experimental studies have investigated how individuals' drinking beliefs and behaviors respond to short-term advertising exposure in a controlled setting. Findings from this literature have been mixed. For instance, Lipsitz *et al.* (1993) alternately showed televised beer commercials, anti drinking public service announcements, and soft-drink commercials to three groups of fifth- and eighth-grade students. They did not find any significant differences in expectancies regarding drinking outcomes across any of the groups. Slater *et al.* (1997) examined the responses of high-school students to television beer advertisements embedded in sports or entertainment programs. They found that the responses were split along gender lines, with female students reacting more negatively to the beer advertisements than male students, especially when viewing sports content. The authors also found that white adolescents who responded favorably to the ads were more likely to report current drinking and future intentions to drink, though the effects were relatively small. It is difficult to disentangle causality in this study because favorable reaction to advertising may simply reflect the student's underlying predisposition to drinking.

Saffer and Dave (2006) also reviewed prior econometric studies on the effects of alcohol advertising on alcohol consumption for the general adult population, according to the source of variation in the advertising measure (time-series, cross-sectional, panel, advertising bans). The vast majority of these studies did not show any substantial positive effects of advertising on overall alcohol consumption. The bulk of these studies though have utilized national time-series data, which often lack variation and confound effects with other unobserved trends. However, given that most individuals consume alcohol without imposing external costs, the more relevant question concerns whether alcohol advertising impacts problem drinking per se. Saffer (1991, 1997) provided indirect evidence on this issue. The study found that countries that ban broadcast alcohol advertisements have lower rates of traffic fatalities as well as alcohol consumption (Saffer, 1991). Saffer (1997) studied the effects of broadcast and outdoor advertising in 75 media markets on motor vehicle fatalities. It was found that a total ban on alcohol advertising could save as many as 5000–10 000 lives, implying an advertising elasticity of between 0.12 and 0.25.

Econometric studies find more consistent and stronger evidence of brand-switching effects in the alcohol industry. Fisher and Cook (1995), for instance, analyzed US data spanning 1970–90 and did not find any evidence that advertising impacts overall alcohol consumption. However, they did find that increased liquor advertising is associated with a reduced consumption of wine, suggesting cross-beverage market share effects. Nelson and Moran (1995) further found that advertising reallocates inter brand market shares, and to a smaller extent also inter beverage market shares, consistent with Fisher and Cook (1995).

Broadcast advertising in the alcohol industry generally aims at brand differentiation, whereas price-based advertising is more common in the print media, especially newspapers. There is some evidence that such price-based advertising leads to pro competitive effects consistent with the informative view of advertising. Sass and Saurman (1995) indicated that large national brewers gain market share at the expense of smaller firms when states restrict advertising of retail prices. They found that the presence of restrictions on price advertising increased market concentration at the state level, both absolutely and relative to measures of national concentration. Additional restrictions on non price advertising did not affect market concentration. Milyo and Waldfogel (1999) exploited the US Supreme Court ruling that overturned Rhode Island's ban on price advertising of alcoholic beverages. Using

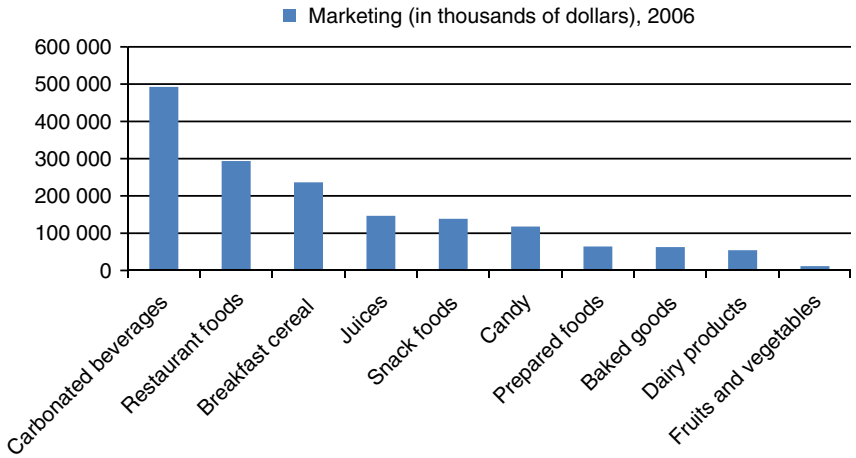


Figure 2 Total youth marketing for reported brands, 44 companies. Adapted from Federal Trade Commission (2008). *Marketing Food to Children and Adolscents: A Report to Congress.*

Massachusetts as a control, they found that price-based advertising substantially reduced the price of the advertised good, though there was little effect on the price of the non advertised good and no significant effect on price dispersion.

In summary, most evidence points to very weak or nonexistent advertising-induced market expansion effects in the alcohol industry. Several studies do not find strong positive effects of advertising on total alcohol consumption. There is some evidence that this overall nil effect may be masking salient effects for certain subpopulations. For instance, some studies have indicated that alcohol advertising increases indicators of problem drinking (for instance, motor vehicle fatalities) and drinking among adolescents, though in both of these cases the elasticity magnitudes are relatively small (and certainly smaller than estimated price responses). It should be noted that many of these econometric studies have estimated advertising effects conditional on price, which precludes one of the mechanisms through which advertising may impact primary or selective demand – that is, through changes in the retail price. This is especially relevant for price-based advertising. Tremblay and Okuyama (2001), for instance, made the point that if the elimination of advertising restrictions promotes price competition, then elimination of the self-imposed broadcast advertising ban in the liquor industry could cause alcohol consumption to rise even if advertising had no direct effect on market demand. There is more consistent evidence of advertising-induced brand-switching effects both at the brand level and the beverage level. This is in accord with the persuasive view of advertising, wherein the main role of advertising and promotion is to generate potentially spurious brand differentiation and enhance the brand's monopolistic power. At the same time, there is also some indication from studies based on cross-state restrictions of price-based advertising that such advertising can lower retail prices and have pro competitive effects.

Advertising of Food and Soft Drinks

Total marketing expenditures in 2006 for the food and beverage industry were highest in the carbonated beverages and restaurant foods categories, with $3.19 and $2.18 billion spent, respectively (Figure 2). Breakfast cereal ranked third in terms of marketing targeted at *youth* (ages 2–17), with $792 million spent. Overall, however, juice and non carbonated beverages and snack foods ranked higher than breakfast cereal, with $1.25 billion and $852 million spent, respectively. The levels of concentration across food and soft drink industries vary, with carbonated soft drink, cereal, and snack foods relatively concentrated compared to other food categories (see Table 2).

Similar to the tobacco and alcohol industries, the soft drink industry is relatively concentrated, with a Herfindahl–Hirschman index of 1094.5 in 2007 (Table 2). There is evidence that the soft drink industry might be more cooperative than predatory in nature, which would render them more likely to capture demand that does not exist rather than capturing a competing company's demand (Gasmi *et al.*, 1992). The Coca-Cola and Pepsi companies are the leading advertisers in the carbonated drink industry, and when the sugar rationing that was implemented in 1942 ended, soft drinks advertising on television experienced a significant increase (Wilcox *et al.*, 2009).

The breakfast cereal industry is characteristic of a very tight oligopoly, with a Herfindahl–Hirschman index of 2425.5 in 2007 (Table 2). There has been evidence that, within the cereal industry, incumbent firms often respond to the entry of new firms with advertising, in order to limit the sales of new entrants (Bagwell, 2007, p. 1729). This anticompetitive behavior may provide support for the persuasive view of advertising in this context, as opposed to the informative view. Yet Ippolito and Mathios (1990) suggested that, in response to growing evidence of fiber's potential cancer preventing benefit, a ban on advertising health claims for food products was lifted in 1985. As a result, consumption of cereal increased. (Kellogg had already begun its advertising campaign highlighting the link between fiber and cancer in October 1984, in violation of FDA policy.) The authors suggested that this lowered the search costs of obtaining health information.

In the food industry, it is not always clear whether advertising is *persuasive*, *informative*, or whether it has elements of

both. Glazer (1981), for example, examined the effect of an exogenous event on food prices: a newspaper strike in Queens and Long Island, NY, for 2 months in 1978. According to his study, the lack of information on prices during that time led to an increase in prices, perhaps indicating that in this context, advertising is informative (Bagwell, 2007). This may be because of the more competitive nature of the market analyzed. The marketing literature contends that *informative* advertising generally occurs in the early stages of a product's life cycle (to build brand awareness and generate demand); *persuasive* advertising occurs in the growth and early maturity stages of the product life cycle (when a product has gained a certain level of brand awareness); and *reminder* advertising – used to remind or prompt purchases – are for products that have gained market acceptance and are in the maturity stage of their life cycle (Grewal and Levy, 2009).

Some trends in food and beverage consumption are noteworthy (see Statistical Abstract of the US at http://www.census.gov/compendia/statab/cats/health_nutrition/food_consumption_and_nutrition.html). For example, per capita consumption of total fat increased from 56.9 lb in 1980 to 85.2 lb in 2008. Per capita consumption of carbonated soft drinks increased from 35.1 gallons in 1980 to 46.4 gallons in 2003. Whether the link between consumption and advertising is causal is discussed in more detail below, in the context of advertising exposure by children.

Researchers have estimated that children's exposure to advertising has increased from approximately 20 000 commercials in the late 1970s to over 40 000 commercials in the early 2000s (Kaiser Family Foundation, 2004). There is particular concern that food and beverage advertising targeted at children is harmful, as the nutritional content of these products is questionable, with most being high in fat, sugar, or sodium (Kaiser Family Foundation, 2004; Powell *et al.*, 2011). Children may not be rational decision-makers or may not be able to appropriately differentiate between advertising and regular programming on television. The exposure to advertising may lead to increased consumption of these products – suggesting that advertising may be cooperative, leading to an overall increase in consumption, rather than predatory or combative – and ultimately contributing to increased rates of childhood obesity.

There is strong suggestive evidence of the link between advertising and consumption or obesity (see the comprehensive reports by the Institute of Medicine, 2006, and the Kaiser Family Foundation, 2004, for excellent reviews of these studies), yet the potential endogeneity of advertising is an issue. Endogeneity may arise because of a firm wanting to locate in areas where demand is already high, which may support the informative view, as advertising is simply an endogenous response to imperfect consumer information (Bagwell, 2007). Moreover, the advertising/sales ratio may be influenced by profit margins and other variables (Bagwell, 2007). Higher levels of advertising may also be more feasible for firms that are concentrated and profitable.

Companies may also target areas where demand is low to capture additional demand, maybe revealing their cooperative nature. At the same time, companies may be cooperative in areas where demand is high, as mentioned above, to further increase demand on the intensive margin. If industry behavior is combative in this context, ordinary least squares estimates are likely biased upward.

Research suggests that food marketing can have a significant impact on consumption among children in the short-term (Epstein *et al.*, 2008; Halford *et al.*, 2004, 2007; Harris *et al.*, 2009) and the longer-term (Barr-Anderson *et al.*, 2009). One study found that adiposity in children increased with exposure to fast food advertising and that banning those advertising practices could reduce the incidence of childhood overweight by 18% (Chou *et al.*, 2008).

The Institute of Medicine (2006) report concluded that there was substantial evidence that "food and beverage marketing influences the preferences and purchase requests of children, influences consumption at least in the short term, is a likely contributor to less healthful diets, and may contribute to negative diet-related health outcomes and risks" (p. 307). The report goes on to say that "[n]ew research is needed on food and beverage marketing and its impact on diet and diet-related health and on improving measurement strategies for factors involved centrally in this research" (p. 309). In contrast to research in the tobacco and alcohol industries on the effects of advertising on consumption, research in this area is still in its infancy.

Chou *et al.* (2008) used an instrumental variables approach to carefully address the potential endogeneity of advertising, and found significant effects of televised fast-food restaurant advertising on body mass index (BMI) and obesity in children and adolescents, using the National Longitudinal Survey of Youth (children of the 1979 cohort and the 1997 cohort). The price of an advertisement and the number of households with a television in the market area served as instruments for fast food advertising. These instruments were found to be valid in that they strongly predicted advertising yet were legitimately excludable from the BMI equation. The authors then analyzed potential effects of two types of regulation: (1) treating food advertising as an ordinary business expense (and thus eliminating the tax deductibility of advertising) and (2) a complete advertising ban on television. Because the corporate income tax rate was 35%, elimination of the tax deductibility of food advertising costs would be equivalent to increasing the price of advertising by approximately 54%, which in turn would reduce fast-food restaurant messages seen on television by 40% and 33% for children and adolescents, respectively, and would reduce the number of overweight children and adolescents by 7% and 5%, respectively. A ban would reduce the number of overweight children aged 3–11 by 18% and the number of adolescents aged 12–18 by 14%. Yet this may be an overestimate; as Saffer (2000) had correctly pointed out, bans on advertising were only effective if they were comprehensive – covering all media, not simply television. Otherwise, the industry would simply shift its advertising expenditures to other media outlets.

Andreyeva *et al.* (2011) used the Early Childhood Longitudinal Survey (Kindergarten cohort) to show that soft drink and fast food television advertising is associated with increased consumption of soft drinks and fast food among elementary school children. They perform several robustness checks to address the potential endogeneity of advertising. Little effect was found for cereal advertising, which may be because of the strong correlation between cereal consumption

and having breakfast, which promotes reduced overall caloric intake.

In summary, most evidence points to advertising-induced expansion effects in the carbonated soft drink and fast-food restaurant industries, and to weak or nonexistent expansion effects in the cereal industry.

Compared to the cigarette and alcohol industries, the food and non alcoholic beverage industries face relatively little regulation. In light of potential adverse effects of advertising on obesity, however, some self-regulatory efforts have been put forth. One such effort is the 2006 Children's Food and Beverage Advertising Initiative (Council of Better Business Bureaus, 2009), whereby participating companies made efforts to improve the nutritional quality of foods marketed to children. Some question these self-regulatory efforts, though, arguing that a few nutritious products are introduced whereas the unhealthy products continued to be heavily marketed (Kunkel *et al.*, 2009).

Several industrialized countries such as Sweden, Norway, and Finland have banned commercial sponsorship of children's programs. Sweden also does not permit any television advertising targeting children under the age of 12 (Kaiser Family Foundation, 2004). There is no similar ban in the US. The FDA regulates and sets standards for the food industry in the US, and these standards vary by state. There has, however, been an increased focus on the potential effect of advertising on obesity in children. In the White House Task Force on Childhood Obesity Report to the President, the following recommendations related to marketing were made, suggesting that advertising in these industries affect childhood obesity:

- The food and beverage industry should extend its self-regulatory program to cover all forms of marketing to children, and food retailers should avoid in-store marketing that promotes unhealthy products to children (Recommendation 2.5).
- All media and entertainment companies should limit the licensing of their popular characters to food and beverage products that are healthy and consistent with science-based nutrition standards (Recommendation 2.6).
- The food and beverage industry and the media and entertainment industry should jointly adopt meaningful, uniform nutrition standards for marketing food and beverages to children, as well as a uniform standard for what constitutes marketing to children (Recommendation 2.7).
- Industry should provide technology to help consumers distinguish between advertisements for healthy and unhealthy foods and to limit their children's exposure to unhealthy food advertisements (Recommendation 2.8). (Solving the Problem of Childhood Obesity within a Generation, 2010).

The FCC has acknowledged the problem and has partnered with the FTC and the Task Force on Childhood Obesity. (See http://reboot.fcc.gov/parents/media-and-childhood-obesity.) Yet it generally remains the case that an advertisement must clearly misinform the consumer in order to be regulated. Increased government involvement in this context is an ongoing debate, with recent studies suggesting that Congress should become more involved by enforcing corporate accountability, changing how advertising is treated for tax purposes, encouraging alternative solutions to regulation, and utilizing the Interagency Working Group Proposal on Food Marketing to Children (Termini *et al.*, 2011).

Another issue that has been raised is the Federal government's role as advertiser: Beef. It's What's for Dinner; Pork. The Other White Meat; Got Milk?. Most of us have heard these slogans in advertisements for beef, pork, and milk. Yet many of us are unaware that they are sponsored by the Federal government, through its 'checkoff' programs (Wilde, 2007), overseen by the United States Department of Agriculture (USDA) starting 1996. (See the Commodity, Promotion, Research and Information Act of 1996: http://www.ams.usda.gov/AMSv1.0/getfiledDocName=STELPRD3479032.) Researchers such as Wilde question the government's well-funded federally sponsored checkoff programs, which "promote increased total consumption of beef, pork, and dairy products, including energy-dense foods such as bacon cheeseburgers, barbecue pork ribs, pizza, and butter" (Wilde, 2006). At the same time, the USDA's Dietary Guidelines recommend a balanced diet with higher levels of whole grains, fruits, vegetables, fish, and low-fat dairy products consumption, which are not advertised by the government to the same degree.

Although weight loss products (discussed in the next section) go relatively unregulated, nutritional claims for food and beverages have been addressed with regulations on food labels, which can be viewed as an indirect form of advertising. Using the National Health Interview Survey, Variyam and Cawley (2006) showed that the implementation of new nutritional labels as a result of the Nutrition Labeling and Education Act of 1990 (effective in 1994) was associated with a decrease in body weight and the probability of obesity.

More recently, calorie posting for chain restaurants (with 20 or more stores in a state) was mandated, starting with New York City in 2008 and eventually becoming a requirement for all states as part of the new health care law (Adamy, 2010; Bollinger *et al.*, 2011). Approximately 20 cities or states mandated calorie postings on menus after New York City (Adamy, 2010). Preliminary studies for New York have shown mixed effects on consumption: Bollinger *et al.* (2011) used data from Starbucks to find that the average number of calories per transaction falls, while Elbel *et al.* (2009) compared New York to New Jersey to find no significant difference.

Advertising of Weight Loss Products

Perhaps one of the most striking examples of deceptive (and yet acceptable) advertising is in the OTC weight loss drug industry. Using magazine and television ads to determine effects on consumption, Cawley *et al.* (2010) showed that people are not as responsive to clearly deceptive advertising compared with nondeceptive advertising. They concluded that although nondeceptive advertising may be more cooperative in nature, deceptive advertising may be more combative in nature, or have no apparent effect.

Research in this area is new, and yet a striking 20.6% of women and 9.7% of men have used OTC weight loss products (Cawley *et al.*, 2010) at some point in their lives. As mentioned in Section 'Overview,' consumers are also ill-informed about government regulation, with half of all consumers under the

impression that these weight loss products are approved for safety and efficacy by the FDA before being sold to the public (Cawley *et al.*, 2010). These OTC weight loss products may accurately be placed in the aforementioned third category of goods that have 'credence' attributes (Darby and Karni, 1973), for which the consumer is unable to accurately evaluate quality even after consuming the good. For instance, since medications have person-specific effects, a consumer may not be able to judge their true effectiveness even after consuming them. These attributes, combined with high turnover of firms in this industry, makes deceptive advertising possible.

Although the FTC Act prohibits 'unfair or deceptive acts or practices,' including both misstatement of facts and failure to disclose important information that consumers should know, it does not prohibit 'puffery' – claims that are so exaggerated that they are clearly incorrect, and no reasonable person would truly believe them. Puffery is defined as "the legal exaggeration of praise, stopping just short of deception, lavished on a product" (Grewal and Levy, 2009).

Advertising of Prescription Drugs

Between 1980 and 2009, expenditures on prescription (Rx) drugs in the US increased from $12 billion to $250 billion, representing an increase of 1974% (see Figure 3).

Most of the increase until the mid-1990s followed the growth in national health expenditures (NHE). However, since around 1995 spending on Rx drugs has outpaced the growth in NHE, making it one of the fastest growing components of health care costs. Consequently, the share of drug spending in NHE roughly doubled between 1994 and 2004, from 5% to 10% (Centers for Medicare and Medicaid Services - CMS; see Figure 4).

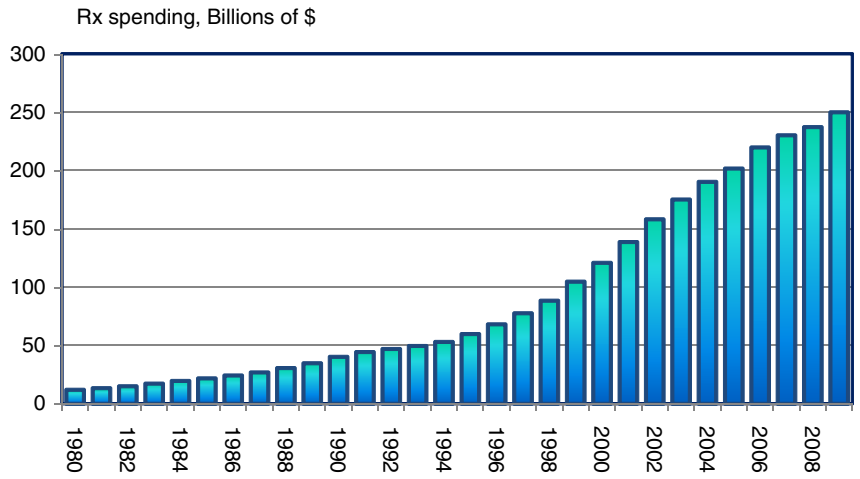


Figure 3 US prescription drug spending. Adapted from Centers for Medicare and Medicaid Services (CMS).

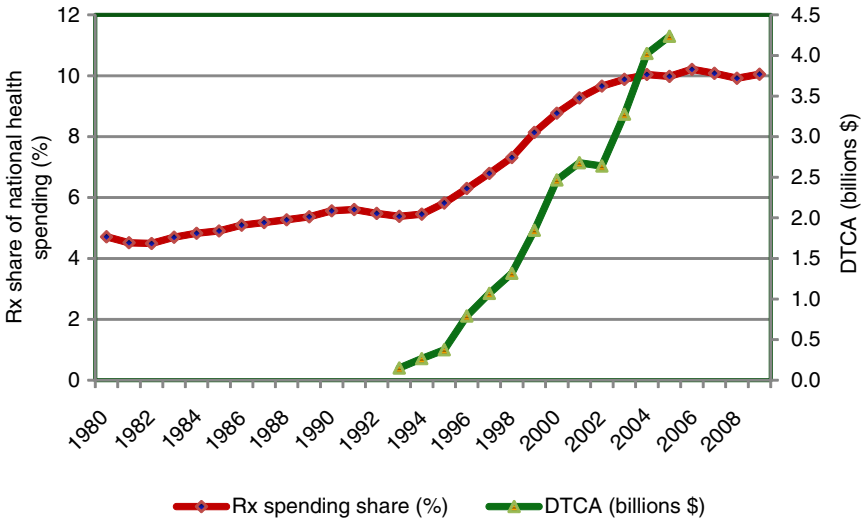


Figure 4 Rx spending share of national health expenditures and direct-to-consumer advertising. Based on data from CMS, Dave and Saffer (2012); Frank, R. G., Berndt, E. R., Donohue, J. M., Epstein, A. and Rosenthal, M. (2002). Trends in direct-to-consumer advertising of prescription drugs. Kaiser Family Foundation February. Available at: http://www.kff.org/rxdrugs/loader.cfmurl=/commonspot/security/getfile.cfm&PageID=14881; Donohue, J. M., Cevasco, M. and Rosenthal, M. B. (2007). A decade of direct-to-consumer advertising of prescription drugs. *New England Journal of Medicine* **357**(7), 673–681.

The growth in the share of prescription drug expenditures has coincided with the growth in pharmaceutical promotion, which increased from $11.4 billion in 1996 to $29.9 billion in 2005 (Donohue *et al.*, 2007). The promotion-to-sales ratio for the pharmaceutical industry is approximately 20%; this compares to an all-industry average of 4–5%. Pharmaceutical products tend to have experience attributes, a low price elasticity of demand (because of the presence of insurance and third-party payers), and a relatively high sales-advertising elasticity – all of which contribute to a high advertising and promotion intensity.

Promotion of prescription drugs is generally limited to patented drugs. It includes direct-to-consumer advertising (DTCA) on broadcast and print media as well as direct-to-physician promotion (DTPP) through visits by company representatives to physician offices (known as detailing), free samples provided to physicians and advertising in professional journals. Although DTPP still comprises most of the promotional budget, the largest relative increase in promotion between 1995 and 2005 resulted from the expansion of DTCA into broadcast media. The share of total promotional spending allocated to DTCA increased from less than 1% in the early 1990s to 8.6% in 1996 to 14.5% in 2003 (see **Figure 5**).

This expansion of DTCA was precipitated by the FDA's clarification of the rules governing broadcast advertising in 1997 and 1999, making it feasible for companies to promote via television and radio advertisements. For a number of years, the FDA had guidelines requiring the advertiser to provide detailed information on usage and risks that is contained in the drug's FDA-approved product label insert, thereby confining ads to print form. The new regulations now require broadcast advertisements to include only 'major statements' of the risks and benefits of the drug along with directions to alternate information sources for full disclosure. This clarification of what constitutes adequate disclosure removed a major barrier that had initially made TV and radio advertisements infeasible. Specifically there was no broadcast advertising in 1993, but it now comprises the primary form of DTCA – amounting to $2.55 billion in 2005.

These new regulations remain a controversial policy and are facing increased scrutiny from Congress and consumer groups. Currently only the US and New Zealand permit broadcast DTCA. At the heart of this debate is whether pharmaceutical promotion and advertising are welfare-promoting. The pharmaceutical industry claims that such advertising educates patients on potential treatment options, opens up lines of communication between the patient and the physician, and can even increase patient–physician contact or expand appropriate treatment for under treated conditions, consistent with the informative view of advertising. Congressional leaders have contended that DTCA raises prescription drug costs, consistent with brand differentiation and the persuasive view of advertising, and requested that the policy be revisited. Some consumer groups maintain that consumers may be harmed by misleading advertising and that the recent expansions in DTCA are responsible for the increases in expenditures on prescription drugs.

Growth in prescription drug spending is broadly driven by increases in utilization and price, and shifts in the composition of drugs being used, all of which may be impacted by DTCA. A comprehensive assessment regarding the welfare effects of pharmaceutical advertising and promotion requires information on three broad but related issues: (1) effects on primary versus selective demand; (2) effects on price; and (3) effects on competition. To inform on the first question, many prior studies gave focus on how DTCA and DTPP have affected pharmaceutical sales and patient adherence. Rosenthal *et al.* (2003) studied brands in five therapeutic classes using aggregated US monthly time-series data from August 1996 to December 1999. They employed an instrumental variables methodology to account for the endogeneity of DTCA and concluded that consumer advertising was primarily

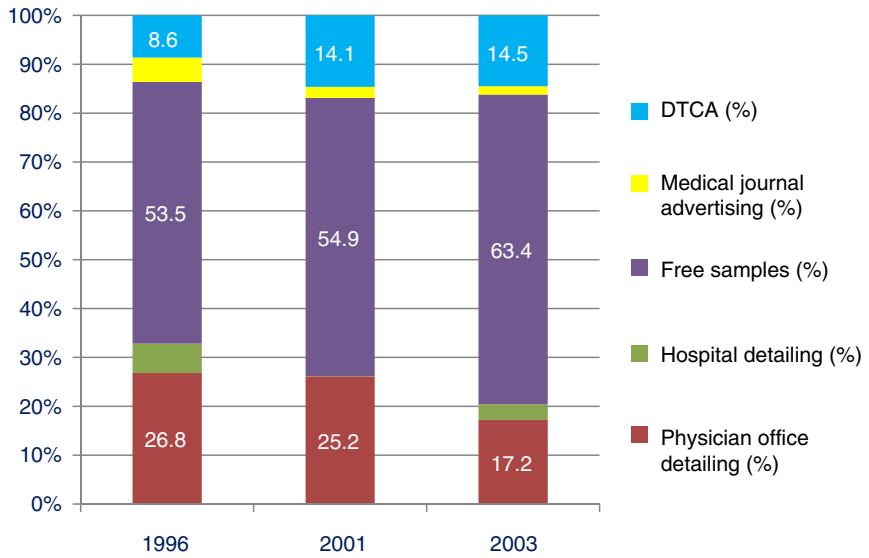


Figure 5 Components of pharmaceutical promotion. Based on data from Donohue, J. M., Cevasco, M. and Rosenthal, M. B. (2007). A decade of direct-to-consumer advertising of prescription drugs. *New England Journal of Medicine* **357**(7), 673–681, and authors' calculations from data used in Dave and Saffer (2012).

effective in raising sales for the entire therapeutic class. Other studies have also noted this market-expansion effect of DTCA, and suggested that DTCA may be more effective in increasing aggregate class demand than in increasing the demand for a particular drug (Iizuka and Jin, 2005, 2007).

These studies combine broadcast and nonbroadcast DTCA into a single aggregate measure, and utilize older data from a time-period when DTCA was just starting to take off and much of it still comprised nonbroadcast forms. This may obscure certain effects since the shift in FDA guidelines specifically applied only to broadcast DTCA; the composition of DTCA has increasingly shifted away from print and toward television and radio advertising as broadcast DTCA became more feasible as a form of promotion for the pharmaceutical industry. Second, both of these forms of DTCA may be expected to have differential effects on pharmaceutical prices and sales.

Dave and Saffer (2012) utilized monthly data on all prescription drugs in four major therapeutic classes from 1994 to 2005, thereby exploiting the period enveloping the FDA's shift in regulations as a natural experiment and exogenous shock to consumer advertising. They separately analyzed the effects of broadcast and nonbroadcast DTCA. Based on drug fixed effects models, they found that broadcast DTCA did impact own-sales with an elasticity of 0.10, and this response is higher relative to nonbroadcast DTCA. This study also found some evidence that class-level DTCA may raise sales for the non advertised drugs. Assuming that physicians are prescribing an equally effective drug, this may be a spillover benefit of DTCA in some cases because non-advertised drugs tend to be older and also cost less.

Directly bypassing the potential endogeneity of advertising, Kravitz *et al.* (2005) examined how DTCA impacts the prescribing behavior of antidepressants in a randomized control trial setting. Standardized patients, mostly professional actors, were assigned to visit physicians and make a specific brand request (referring to a DTC ad), a general drug request, or no request. Results pointed to the role of brand-specific DTCA in raising own-demand by leading to a prescription for that brand, as well as in raising overall class demand.

Additional evidence on the demand effects of DTCA is also provided by studies that examine patient adherence. For instance, Bradford *et al.* (2006), using patient-level data from 1998 to 2004 merged with DTCA information at the national and market levels, found that higher levels of DTC television advertising of statin treatment was significantly associated with improvements in the likelihood of attaining cholesterol management goals for at least some patients. Donohue *et al.* (2004) studied claims data for depressed patients between 1997 and 2000 matched with information on DTCA. They found that consumer advertising of antidepressants was associated with an increase in the number of people diagnosed with depression who initiated medication therapy and a small increase in the number of individuals treated with antidepressants who received the appropriate duration of therapy.

Studies have also examined the impact of advertising aimed at health-care providers, which historically has been the primary form of promotion used by the pharmaceutical industry. Berndt *et al.* (1995), for instance, considered the role of detailing, medical journal advertisements and DTCA in the market for antiulcer drugs before the shift in FDA guidelines. The DTCA examined in this study is very limited and confined only to print media because the study predated the FDA's shift in regulations that made broadcast DTCA feasible. They found the strongest demand effect for detailing and the smallest effect for DTCA. Many other studies also confirmed larger effects of physician-directed promotion relative to those for consumer-directed promotion.

Overall, most of these studies point to positive demand effects of DTCA and DTPP, and generally find that DTCA has stronger class-level effects whereas DTPP has stronger brand-specific effects. There is some suggestive evidence from studies utilizing newer data that DTCA may also have some brand-specific effects, particularly broadcast DTCA, though all studies point to DTPP being more effective relative to DTCA in raising sales. Some of the research also highlights a potential benefit of DTCA – that is, encouraging consumers to seek treatment and take their medications as prescribed.

With respect to the effects of advertising and promotion on price, the evidence is more limited. This paucity of research partly derives from the difficulty in obtaining salient measures of Rx drug prices because of the presence of third-party payers and unobserved rebates from drug manufacturers to third-party payers.

As underscored by the discussion on the three views of advertising, the potential effects on price primarily depend on the strength of scale economies in production and on the impact of advertising on the price elasticity of demand. Under the persuasive view of advertising where the shift in demand becomes relatively more inelastic, advertising raises price as long as there are no strong economies of scale in production to counteract the inelastic demand. Under the informative view of advertising, prices are predicted to decrease because demand would become relatively more elastic.

The few studies that have focused on advertising-induced price effects appear to be in accord with the persuasive view. Rizzo (1999), for instance, found that increased detailing efforts among antihypertensive drugs reduced the price elasticity. This reduction may consequently result in higher prices, though Rizzo did not examine the direct link between detailing and price. The study was based on pooled annual data from 1988 to 1993, which predates the DTCA policy shift, and only considers promotion to physicians. Law *et al.* (2009) examined pharmacy data for Plavix (an antiplatelet drug used to prevent stroke and heart attack in at-risk patients) from 27 Medicaid programs over the period 1999–2005. Plavix initiated DTCA in 2001. This study found that, although there was no change in the preexisting trend in demand, there was a sustained increase in cost per unit of $0.40 (11.8%) after the expansion in DTCA.

Dave and Saffer (2012), utilizing a larger sample of all Rx drugs in four therapeutic classes, also found that DTCA raised the average wholesale price, though the estimated elasticity was of a relatively small magnitude (0.04). Consistent with the positive impact on price, this study also found that the consumer price response became relatively more inelastic during the period when DTCA was expanding. Saffer and Dave presented simulations suggesting that expansions in broadcast DTCA over 1994–2005 accounted for 19% of the overall growth in prescription drug spending, with two-thirds of this

impact driven by an increase in demand and the remainder because of higher advertising-induced prices.

One challenge faced by these empirical studies concerns the simultaneity between advertising and pricing decisions. For instance, Bhattacharya and Vogt (2003) presented a model of joint price and promotion determination over the drug's life cycle. The dynamic profit maximizing strategy for the firm was to initially employ a relatively high level of promotion and to set a relatively low price. These levels would not only increase current quantity demanded, but also raise future demand because high promotion and low prices increased the physicians' and the consumers' stock of knowledge about the drug. In subsequent periods, promotion could be decreased to lower costs and price could be raised to increase revenue.

This trajectory of higher prices and lower advertising over the drug's life cycle is also consistent with the Dorfman-Steiner (1954) condition for optimal advertising discussed in Section 'Overview'; the optimal advertising-to-sales ratio is a positive function of the elasticity of sales with respect to advertising and is inversely related to the elasticity of sales with respect to price. Thus, the decline in advertising over the drug's life cycle is consistent with an age-related decline in the sales-advertising elasticity (Berndt, 2006). It is also consistent with an increase in the price elasticity as the drug ages and newer drugs enter the therapeutic class. A positive association between advertising and price inelasticity may thus reflect causality in both directions – for persuasive goods, advertising may make demand more inelastic, but *ceteris paribus* more inelastic demand also leads to a higher optimal level of advertising.

While both Rizzo (1999) and Dave and Saffer (2012) attempted to address this simultaneity through additional controls, the results should be interpreted in the context of the limitations noted. Nevertheless, these studies point to certain anticompetitive effects of Rx drug promotion. Further evidence is gleaned from studies that have investigated the effects of advertising on entry in the pharmaceutical markets. Scott Morton (2000) found that advertising by branded drugs before patent expiration and generic entry may have a very small deterrence effect on subsequent generic entry depending on the type of advertising, though this effect becomes insignificant in models which instrument for advertising. In a classic study, Benham (1972) found that eyeglass prices were substantially higher in states that prohibited all advertising relative to states with no restrictions. Prices were slightly higher in states that allowed only non price advertising than in states with no restrictions. This strand of the literature suggests that non price advertising by the Rx industry may exert some small upward pressure on prices and possibly have anticompetitive effects, though the evidence is far from conclusive and requires further study.

In summary, DTCA has emerged as a marketing force in the US healthcare system and is only expected to grow along with expenditures on prescription drugs. Although the debate surrounding DTCA is unlikely to be resolved anytime soon, DTCA should be evaluated both in terms of its costs as well as its benefits. The benefits derive from improved health because of increases in the number of individuals using prescription drugs and increased adherence with drug therapy. Detecting and treating health conditions at an earlier stage, through primary care, may also be more cost-effective relative to treatment at a later stage through acute care. Pointing to another potential benefit of promotion, Kwong and Norton (2007) found that detailing (but not other types of advertising) may have a significant positive effect on the number of new products entering into clinical development, with markets for chronic disease with high levels of detailing being more attractive to pharmaceutical firms.

Studies that show advertising-induced market expansion effects generally interpret these findings as welfare-improving. Although there was certainly an element of improved adherence and expanded treatment underlying the market expansion, David *et al.* (2010) showed that increased levels of promotion and advertising lead to increased reporting of adverse medical events for certain conditions. This suggests that promotion-driven market expansion could raise the risk that the drug is prescribed inappropriately. In addition to potential misuse, the costs of DTCA also result from increased drug prices and increased use of more expensive drugs in place of equally effective lower-priced drugs. Higher drug and health care expenditures in turn can raise insurance premiums and may lead to a larger prevalence of uninsured.

New Directions

Online Advertising

The Pew Research Center showed that Internet usage among Americans has increased by approximately 72% since 2000, with an estimated 46% of respondents using the Internet in 2000 and 79% in 2010 (Pew Research Center, 2011). Residential broadband subscribers increased from 5.2 million in 2000 to 70.1 million in 2008, a 1248% increase over 8 years. With more households having access to the Internet, online advertising, a form of 'interactive media' (which also includes mobile phones) has become more prevalent, as is evident in **Figure 6**. Online advertising was in existence in the early 1990s (Li and Leckenby, 2006), yet as **Figure 6** reveals, as recently as 2000 online media suppliers represented less than 5% of total suppliers, compared with 14% in 2009.

As Li and Leckenby (2006) pointed out, "the internet has capacities to extend the function of advertising far beyond what traditional media are able to accomplish... The expanded function of internet advertising comes from its horizontal integration of three key marketing channel capacities (communication, transaction and distribution) and vertical integration of marketing communications, including advertising, public relations, sales promotion and direct marketing" (p. 203).

Figure 7 shows the importance of control ownership by the advertiser or consumer in determining the effectiveness of Internet advertising (Li and Leckenby, 2006). This Interactive Advertising Model (IAM), developed by Rodgers and Thorson (2000), revealed the increased complexity of Internet advertising as compared with advertising in other media. Some ad formats are controversial; for example, interstitial ads, which include pop-ups and pop-unders, could be intrusive and irritating, particularly for individuals who were in 'search mode' rather than in 'surf mode' (Li and Leckenby, 2006). (Banner ads, by contrast, are usually viewed voluntarily.) New formats adopted by Internet advertisers included three-dimensional

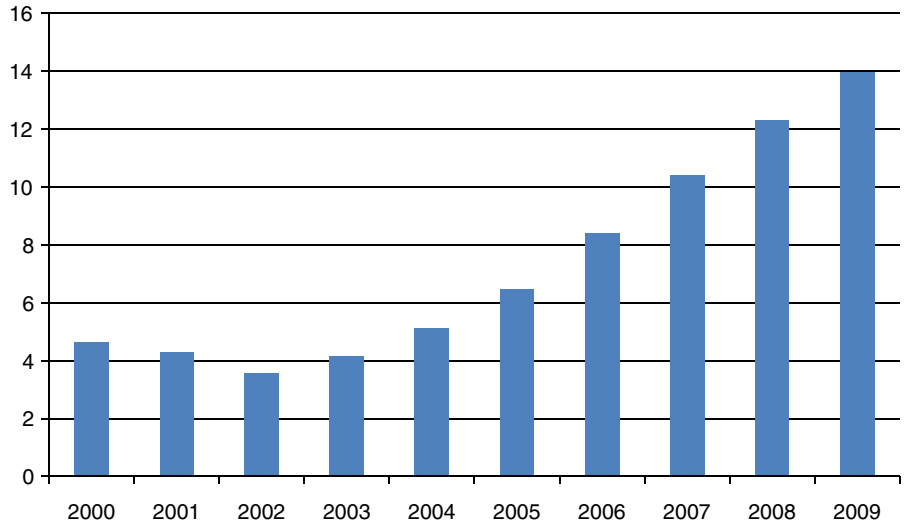


Figure 6 Percentage of online media supplier advertising revenues (out of total). *Note*: Authors' calculations based on data from the Statistical Abstract of the U.S. Online media suppliers include all online digital suppliers in direct, national, and local markets.

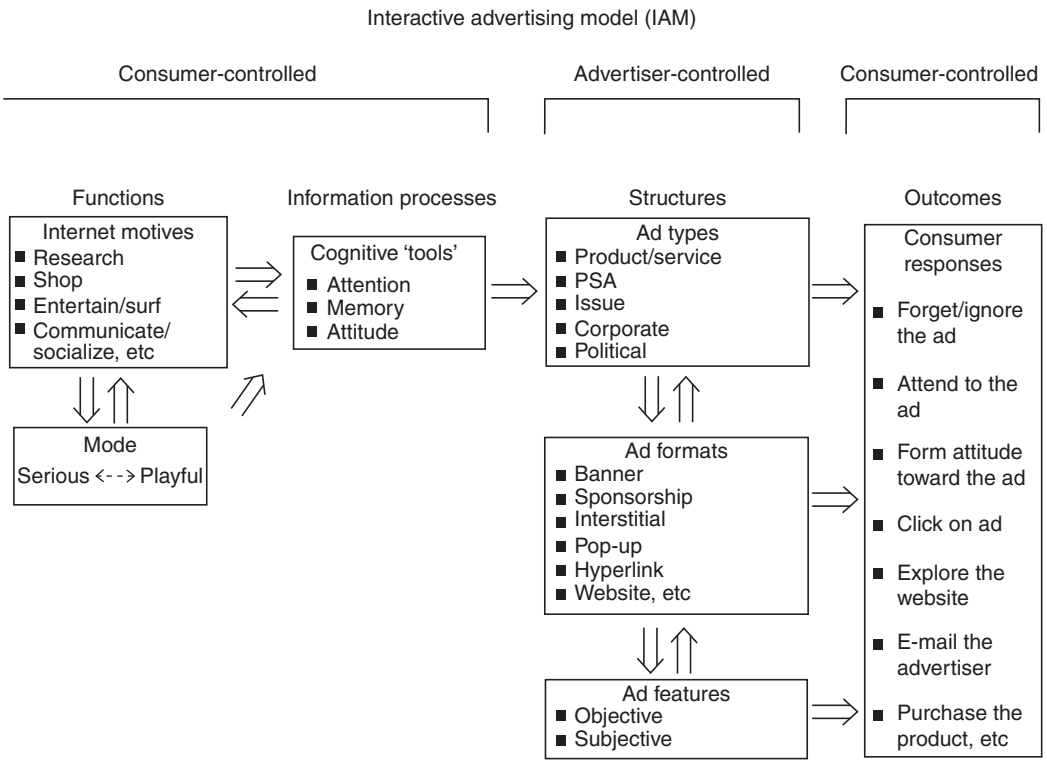


Figure 7 The interactive advertising model (IAM). Adapted from Rodgers, S. and Thorson, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising* **1**(1), 42–61.

visualization and product placement in online games (Li and Leckenby, 2006). There were also virtual worlds in which companies could pinpoint when avatars look at specific ads (Grewal and Levy, 2009).

The FTC warns advertisers that if they wish to advertise on the Internet, the same rules apply for electronic advertising as for other forms of advertising (Federal Trade Commission, 2000). Advertising must not mislead consumers or make claims that are unsubstantiated. The FTC summary also sets forth guidelines to protect consumer privacy, particularly relevant for online advertisers.

Neuroeconomic Framework

Economists have integrated insights from behavioral economics and neuroscience in a budding area of research known

as neuroeconomics. Bernheim and Rangel (2004, 2005), for instance, presented a theory of addiction based on a neuroeconomic framework of decision-making. This area of research provided a promising new direction for advertising studies on two fronts. First, the persuasive view of advertising posits that advertising impacts demand through potentially spurious brand differentiation, which in turn affects consumer preferences. However, as Bagwell (2007) noted, studies remain "agnostic as to the underlying mechanism through which advertising shifts tastes" (p. 1825). Assimilating insights from neurological research with regard to how decisions are made can help advance our understanding of the mechanisms underlying the response to advertising. Second, relevance for public health and policy requires not just knowing the average population response but also an understanding of how advertising particularly affects behaviors of at-risk individuals – that is, those who impose external costs on others and who are the targets of public policy. For instance, is advertising predicted to affect drinking behaviors of heavy alcohol users or affect junk food consumption habits among overweight/obese individuals? Neuroeconomic models of decision-making, particularly in the context of goods with addictive properties, have distinct predictions in this regard.

Consider the following neuroeconomic model based on Bernheim and Rangel (2004, 2005). Saffer (2011) provided a related discussion on alcohol advertising, and Ruhm (2012) provided a discussion of neuroeconomic models as they applied to overeating and obesity. Individuals have been found to rely on two neural systems to make decisions relating to addictive consumption. One system reflects a rational mechanism (RM), where choices and decisions are based on reasoning and rational cost–benefit calculus. When decisions are made according to the RM, the individual is in a 'cold state' and here the standard neoclassical demand model is applicable. The other neural system reflects a hedonic forecasting mechanism (HFM), where choices are based on 'cravings' and short-term rewards. The HFM does not involve higher reasoning, and is in control when decisions must be made very quickly. In this case, the individual is defined to be in a 'hot state.' The switching mechanism between cold and hot states depends on environmental cues such as advertising, the individual's addictive stock accumulated through past consumption experience, and other factors.

Under this neuroeconomic framework, advertising would increase primary demand and lead to overall market expansion effects, not just brand-switching effects. The model also indicates that the response to advertising is a learned behavior, and individuals with a higher addictive stock may be particularly susceptible to advertising-related cues. Individuals can also override evaluations of the HFM by exercising cognitive control and asserting dominance of the RM; this points to individual heterogeneity in the response to advertising based on factors that affect the costs of exercising cognitive control.

In summary, it is known from various studies conducted for healthcare markets that advertising can affect both selective and primary demand, can be persuasive and in turn affect tastes and preferences, and can have an average population response that may mask heterogeneous responses across individual characteristics, population subgroups, and along the consumption distribution. It is less clear *why* these responses are observed. Integrating insights from cognitive psychology, neuroscience, behavioral economics, and other disciplines provides a promising avenue for further understanding these responses.

Summary

This article has provided a conceptual and empirical framework through which to study the economics of advertising in the context of markets for health inputs. The Dorfman–Steiner model positively relates advertising intensity to the advertising-sales elasticity and negatively relates it to the price elasticity of demand. The competing informative and persuasive views of advertising are explored, in addition to the view of advertising simply as a complement to the advertised good. Search and experience goods are distinguished and briefly discussed. These attributes, combined with the product's price and advertising elasticities, generally determine the advertising intensity of the product.

An analysis of advertising in select health markets is covered, with a focus on selective versus primary demand effects and relevance for public health. Econometric studies typically find effects on consumption for tobacco, soft drinks, fast-food restaurants, and prescription drugs, which reflect an advertising-induced industry expansion effect. For the alcohol industry, there is some evidence of small positive overall demand effects for certain segments of the population such as problem drinkers and youth. More empirical research, however, needs to be conducted, particularly addressing the potential endogeneity of advertising. A key obstacle for researchers is the high price of acquiring detailed advertising data. Currently, advertising data are only provided by a few companies, including Nielsen and TNS (now part of Kantar Media).

Future research in this area will increasingly stress the roles of online advertising, which allows greater targeting of the product to the potential user, and neuroeconomics, which may yield insights on the pathways underlying the consumer response. The emerging research combining behavioral economics and neuroscience is timely, for instance, as online purchases made after exposure to advertising may have higher probabilities of being 'hot state,' impulsive purchases. Some thoughts are provided on new directions for research in these increasingly important topic areas.

See also: Advertising Health Care: Causes and Consequences. Pharmaceutical Marketing and Promotion

References

Adamy, J. (2010). Coming soon: Theaters, airplanes to post calories. *Wall Street Journal*. Available at: http://online.wsj.com/article/SB10001424052748704323704575462021475610064.htmlmod=WSJ_hps_MIDDLEForthNews (accessed 09.02.13).

Anderson, P., de Bruijin, A., Angus, K., Gordon, R. and Hastings, G. (2009). Impact of alcohol advertising and media exposure on adolescent alcohol use: A systematic review of longitudinal studies. *Alcohol and Alcoholism* **44**(3), 229–243.

Andreyeva, T., Kelly, I. R. and Harris, J. (2011). Exposure to food advertising on television: Associations with children's fast food and soft drink consumption and obesity. *Economics and Human Biology* **9**(3), 221–233.

Avery, R. J., Kenkel, D. S., Lillard, D. and Mathios, A. (2007). Private profits and public health: Does advertising smoking cessation products encourage smokers to quit? *Journal of Political Economy* **115**(3), 447–481.

Bagwell, K. (2007). The economic analysis of advertising. In Armstrong, M. and Porter, R. (eds.) *Handbook of Industrial Organization*, vol. III. North-Holland: Amsterdam.

Barr-Anderson, D. J., Larson, N. I., Nelson, M. C., Neumark-Sztainer, D. and Story, M. (2009). Does television viewing predict dietary intake five years later in high school students and young adults? *International Journal of Behavioral Nutrition and Physical Activity* **6**, 7.

Becker, G. S. and Murphy, K. M. (1993). A simple theory of advertising as a good or bad. *Quarterly Journal of Economics* **108**, 941–964.

Benham, L. (1972). The effect of advertising on the price of eyeglasses. *Journal of Law and Economics* **15**, 337–352.

Berndt, E., Bui, L., Reiley, D. and Urban, G. (1995). Information, marketing and pricing in the US antiulcer drug market. *American Economic Review* **85**(2), 100–105.

Berndt, E. R. (2006). The United States experience with direct-to-consumer advertising of prescription drugs: What have we learned? In Sloan, F. A. and Hsieh, C. R. (eds.) *Promoting and coping with pharmaceutical innovation: An international perspective*. New York: Cambridge University Press.

Bernheim, B. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review* **94**(5), 1558–1590.

Bernheim, B. and Rangel, A. (2005). From neuroscience to public policy: A new economic view of addiction. *Swedish Economic Policy Review* **12**, 99–144.

Bhattacharya, J. and Vogt, G. (2003). A simple model of pharmaceutical price dynamics. *Journal of Law and Economics* **46**(2), 599–626.

Bittlingmayer, G. (2008). Advertising. The concise encyclopedia of economics. Library of Economics and Liberty. Available at: http://www.econlib.org/library/Enc/Advertising.html (accessed 09.02.13).

Bollinger, B., Leslie, P. and Sorensen, A. (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy* **3**(1), 91–128.

Borden, N. H. (1942). *The economic effects of advertising*. Chicago: Richard D. Irwin, Inc.

Bradford, W. D., Kleit, A. N., Nietert, P. J., et al. (2006). Effects of direct-to-consumer advertising of hydroxymethylglutaryl coenzyme A reductase inhibitors on attainment of LDL-C goals. *Clinical Therapeutics* **28**(12), 2105–2118.

Brown, R. S. (1978). Estimating advantages to large-scale advertising. *Review of Economics and Statistics* **60**, 428–437.

Bureau of the Census, US Department of Commerce (2007). *Economic census*. Washington, DC: US Government Printing Office.

Cawley, J., Rosemary, A. and Matthew E. (2010). Effect of advertising and deceptive advertising on consumption: The case of over-the-counter weight loss products. Presented at the City University of New York Graduate Center, October 1.

Chaloupka, F. and Warner, K. (2000). Economics of smoking. In Newhouse, J. and Culyer, A. (eds.) *Handbook of health economics*, vol. IB. North-Holland: Amsterdam.

Chamberlin, E. (1933). *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Chou, S., Rashad, I. and Grossman, M. (2008). Fast-food restaurant advertising on television and its influence on childhood obesity. *Journal of Law and Economics* **51**, 599–618.

Council of Better Business Bureaus (2009). Children's food and beverage advertising initiative. Available at: http://us.bbb.org/WWWRoot/SitePage.aspx site=113&id=dba51fbb-9317-4f88-9bcb-3942d7336e87 (accessed 20.05.11).

Darby, M. R. and Karni, E. (1973). Free competition and the optimal amount of fraud. *Journal of Law and Economics* **16**(1), 67–88.

Dave, D. and Saffer, H. (2012). Impact of direct-to-consumer advertising on pharmaceutical prices and demand. *Southern Economic Journal* **79**(1), 97–126.

Dave, D. and Saffer, H. (2013). Demand for smokeless tobacco: Role of advertising. *Journal of Health Economics* **32**(4), 682–697.

David, G., Markowitz, S. and Richards-Shubik, S. (2010). The effects of pharmaceutical marketing and promotion on adverse drug events and regulation. *American Economic Journal: Economic Policy* **2**(4), 1–25.

Donohue, J. M., Berndt, E. R., Rosenthal, M., Epstein, A. M. and Frank, R. G. (2004). Effects of pharmaceutical promotion on adherence to the treatment guidelines for depression. *Medical Care* **42**(12), 1176–1185.

Donohue, J. M., Cevasco, M. and Rosenthal, M. B. (2007). A decade of direct-to-consumer advertising of prescription drugs. *New England Journal of Medicine* **357**(7), 673–681.

Dorfman, R. and Steiner, P. O. (1954). Optimal advertising and optimal quality. *American Economic Review* **44**, 826–836.

Doyle, P. (1968). Advertising expenditure and consumer demand. *Oxford Economic Papers* **20**(3), 394–415.

Eckard, Jr., E. W. (1991). Competition and the cigarette TV advertising ban. *Economic Inquiry* **29**, 119–133.

Elbel, B., Kersh, R., Brescoll, V. L. and Dixon, L. B. (2009). Calorie labeling and food choices: A first look at the effects on low-income people in New York city. *Health Affairs* **28**(6), w1110–w1121.

Emery, S., Wakefield, M. A., Terry-McElrath, Y., et al. (2005). Televised state-sponsored anti-tobacco advertising and youth smoking beliefs and behavior in the United States, 1999–2000. *Archives of Pediatrics and Adolescent Medicine* **159**, 639–645.

Epstein, L. H., Roemmich, J. N., Robinson, J. L., et al. (2008). A randomized trial of the effects of reducing television viewing and computer use on body mass index in young children. *Archives of Pediatrics and Adolescent Medicine* **162**, 239–245.

Federal Trade Commission (2000). Advertising and marketing on the internet: Rules of the road. Available at: http://business.ftc.gov/documents/bus28-advertising-and-marketing-internet-rules-road (accessed 15.07.13).

Fisher, J. C. and Cook, P. A. (1995). *Advertising, alcohol consumption, and mortality: An empirical investigation*. Westport, CT: Greenwood Press.

Gasmi, F., Laffont, J. J. and Vuong, Q. (1992). Econometric analysis of collusive behavior in a soft-drink market. *Journal of Economics and Management Strategy* **1**, 277–311.

Glazer, A. (1981). Advertising, information, and prices – A case study. *Economic Inquiry* **19**, 661–671.

Goel, R. K. and Morey, M. J. (1995). The interdependence of cigarette and liquor demand. *Southern Economic Journal* **62**(2), 451–459.

Goldman, L. K. and Glantz, S. A. (1998). Evaluation of antismoking advertising campaigns. *Journal of the American Medical Association* **279**(10), 772–777.

Grewal, D. and Levy, M. (2009). *Marketing*. 2nd ed, Irwin: McGraw-Hill.

Halford, J. C. G., Boyland, M. J., Hughes, G., Oliveira, L. P. and Dovey, T. M. (2007). Beyond-brand effect of television (TV) food advertisement/commercials on caloric intake and food choice of 5-7-year-old children. *Appetite* **49**, 263–267.

Halford, J. C. G., Gillespie, J., Brown, V., et al. (2004). Effect of television advertisements for foods on food consumption in children. *Appetite* **42**, 221–225.

Hamilton, W. L., Turner-Bowker, D. M., Celebucki, C. C. and Connolly, G. N. (2002). Cigarette advertising in magazines: The tobacco industry response to the master settlement agreement and to public pressure. *Tobacco Control* **11**, 54–58.

Harris, J. L., Pomeranz, J. L., Lobstein, T. and Brownell, K. D. (2009). A crisis in the marketplace: How food marketing contributes to childhood obesity and what can be done. *Annual Review of Public Health* **30**, 211–225.

Iizuka, T. and Jin, G. Z. (2005). The effect of prescription drug advertising on doctor visits. *Journal of Economics and Management Strategy* **14**(3), 701–727.

Iizuka, T. and Jin, G. Z. (2007). Direct to consumer advertising and prescription choice. *Journal of Industrial Economics* **55**(4), 771.

Institute of Medicine (2006). Food *Marketing to children and youth: Threat or opportunity?* Washington, DC: National Academy of Sciences, Committee on Food Marketing and the Diets of Children and Youth.

Ippolito, P. M. and Mathios, A. D. (1990). Information, advertising and health choices: A study of the cereal market. *RAND Journal of Economics* **21**, 459–480.

Jernigan, D. and O'Hara, J. (2004). Alcohol advertising and promotion. In Bonnie, R. J. and O'Connell, M. E. (eds.) *Reducing underage drinking: A collective responsibility*. Washington, DC: National Academies Press.

Kaiser Family Foundation (2004). *The role of media in childhood obesity*. Menlo Park, CA: Kaiser Family Foundation.

Kaldor, N. and Silverman, R. (1948). *A statistical analysis of advertising expenditure and of the revenue of the press*. Cambridge, UK: University Press.

Kaldor, N. V. (1950). The economic aspects of advertising. *Review of Economic Studies* **18**, 1–27.

Kravitz, R. L., Epstein, R. M., Feldman, M. D., et al. (2005). Influence of patients' requests for direct-to-consumer advertised antidepressants: A randomized controlled trial. *Journal of the American Medical Association* **293**(16),

1995–2002, (Erratum in: *Journal of the American Medical Association* **294**(19), 2436).
Kunkel, D., McKinley, C. and Wright, P. (2009). The impact of industry self-regulation on the nutritional quality of foods advertised on television to children. Available at: http://www.childrennow.org/uploads/documents/adstudy_2009.pdf (accessed 15.07.13).
Kwong, W. J. and Norton, E. C. (2007). The effect of advertising on pharmaceutical promotion. *Review of Industrial Organization* **31**, 221–236.
Law, M. R., Soumerai, S. B., Adams, A. S. and Majumdar, S. R. (2009). Costs and consequences of direct-to-consumer advertising for Clopidogrel in Medicaid. *Archives of Internal Medicine* **169**(21), 1969–1974.
Lewit, E. M., Coate, D. and Grossman, M. (1981). The effects of government regulation on teenage smoking. *Journal of Law and Economics* **24**(3), 545–569.
Li, H. and Leckenby, J. (2006). Internet advertising formats and effectiveness. In Schumann, D. and Thorson, E. (eds.) *Internet Advertising, Theory and Research*. Mahwah, NJ: Lawrence Erlbaum Associates.
Lipsitz, A., Brake, G., Vincent, E. J. and Winters, M. (1993). Another round for the brewers: Television ads and children's alcohol expectancies. *Journal of Applied Social Psychology* **23**(6), 439–450.
Merriam-Webster (2011). *Advertising*. Available at: http://www.merriam-webster.com/dictionary/advertising (accessed 20.05.2000).
Milyo, J. and Waldfogel, J. (1999). The effect of price advertising on prices: Evidence in the wake of 44 Liquormart. *American Economic Review* **89**(5), 1081–1096.
Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy* **78**, 311–329.
Nelson, P. (1974). Advertising as information. *Journal of Political Economy* **82**, 729–754.
Nelson, P. (1975). The economic consequences of advertising. *Journal of Business* **48**, 213–241.
Nelson, P. and Moran, J. R. (1995). Advertising and U.S. alcohol beverage demand: System-wide estimates. *Applied Economics* **27**(12), 1225–1236.
Ozga, S. A. (1960). Imperfect markets through lack of knowledge. *Quarterly Journal of Economics* **74**, 29–52.
Pew Research Center (2011). *Usage over time*. Available at: http://www.pewinternet.org/Static-Pages/Trend-Data/Usage-Over-Time.aspx (accessed 09.02.13).
Powell, L. M., Schermbeck, R. M., Szczypka, G., Chaloupka, F. J. and Braunschweig, C. L. (2011). Trends in the nutritional content of television food advertisements seen by children in the United States. *Archives of Pediatrics and Adolescent Medicine* **165**(12), 1078–1086.
Rizzo, J. (1999). Advertising and competition in the ethical pharmaceutical industry: The case of hypertensive drugs. *Journal of Law and Economics* **42**(1), 89–116.
Roberts, M. J. and Samuelson, L. (1988). An empirical analysis of dynamic, nonprice competition in an oligopolistic industry. *RAND Journal of Economics* **19**(2), 200–220.
Robinson, J. (1933). *Economics of Imperfect Competition*. London: MacMillan and Co.
Rodgers, S. and Thorson, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising* **1**(1), 42–61.
Rosenthal, M. B., Berndt, E. R., Donohue, J. M., Epstein, A. M. and Frank, R. G. (2003). Demand effects of recent changes in prescription drug promotion. In Cutler, D. M. and Garber, A. M. (eds.) *Frontiers in health policy research*, vol. 6. Cambridge, MA: MIT Press.
Ross, H. and Chaloupka, F. J. (2002). *Economics of tobacco control*. Chicago, IL: International Tobacco Evidence Network.
Ruhm, C. (2012). Understanding overeating and obesity. *Journal of Health Economics* **31**(6), 781–796.
Saffer, H. (1991). Alcohol advertising bans and alcohol abuse: An international perspective. *Journal of Health Economics* **10**, 65–79.
Saffer, H. (1997). Alcohol advertising and motor vehicle fatalities. *Review of Economics and Statistics* **79**(3), 431–442.
Saffer, H. (2000). Tobacco advertising and promotion. In Jha, P. and Chaloupka, F. (eds.) *Tobacco Control Policies in Developing Countries*. New York: Oxford University Press.
Saffer, H. (2011). New approaches to alcohol marketing research. *Addiction* **106**, 472–479.
Saffer, H. and Chaloupka, F. (2000). The effect of tobacco advertising bans on tobacco consumption. *Journal of Health Economics* **19**(6), 1117–1137.
Saffer, H. and Dave, D. (2006). Alcohol advertising and alcohol consumption by adolescents. *Health Economics* **15**, 617–637.
Sass, T. R. and Saurman, D. S. (1995). Advertising restrictions and concentration: The case of malt beverages. *Review of Economics and Statistics* **77**, 66–81.
Schmalensee, R. (1972). *The Economics of Advertising*. Amsterdam: North-Holland.
Schonfeld & Associates (2010). *Advertising ratios and budgets*. Libertyville, IL: Schonfeld & Associates, Inc., June 1.
Scott Morton, F. M. (2000). Barriers to entry, brand advertising, and generic entry in the U.S. pharmaceutical industry. *International Journal of Industrial Organization* **18**(7), 1085–1104.
Slater, M., Rouner, D., Domenech-Rodriquez, M., et al. (1997). Adolescent responses to TV beer ads and sports content/context: Gender and ethnic differences. *Journalism and Mass Communication Quarterly* **74**, 108–122.
Solving the Problem of Childhood Obesity within a Generation (2010). *White House Task Force on Childhood Obesity Report to the President*, Washington, DC. May.
Statistical Abstract of the United States (2009). Information & communications. Internet Publishing and Broadcasting and Internet Usage. Available at: http://www.census.gov/compendia/statab/cats/information_communications/internet_publishing_and_broadcasting_and_internet_usage.html (accessed 20.05.11).
Stigler, G. J. (1961). The economics of information. *Journal of Political Economy* **69**, 213–225.
Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *American Economic Review* **67**, 76–90.
Termini, R. B., Roberto, T. A. and Hostetter, S. G. (2011). Should congress pass legislation to regulate child-directed food advertising? *Food and Drug Policy Forum* **1**, 9), Available at http://www.foodanddrugpolicyforum.org/2011/05/vol-1-no-9-should-congress-pass.html.
Thomas, L. G. (1989). Advertising in consumer good industries: Durability, economies of scale, and heterogeneity. *Journal of Law and Economics* **32**, 164–194.
Tremblay, V. J. and Okuyama, K. (2001). Advertising restrictions, competition, and alcohol consumption. *Contemporary Economic Policy* **19**(3), 313–321.
Variyam, J.N. and Cawley J. 2006. *Nutrition labels and obesity*. National Bureau of Economic Research Working Paper No. 11956. Cambridge, MA: National Bureau of Economic Research.
Verma, V. K. (1980). A price theoretic approach to the specification and estimation of the sales-advertising function. *Journal of Business* **53**, S115–S137.
Wilcox, G. B., Sara Kamal and Gangadharbatla, H. (2009). Soft drink advertising and consumption in the United States 1984–2007. *International Journal of Advertising* **28**(2), 351–367.
Wilde, P. E. (2006). Federal communication about obesity in the dietary guidelines and checkoff programs. *Obesity* **14**(6), 967–973.
Wilde, P. (2007). Plowing through the politics of agriculture. *Tufts Nutrition* **9**(1), 15.

Further Reading

Baltagi, B. H. and Levin, D. (1986). Estimating dynamic demand for cigarettes using panel data: The effects of bootlegging, taxation, and advertising reconsidered. *Review of Economics and Statistics* **68**(1), 148B55.
Chaloupka, F. J., Grossman, M. and Saffer, H. (2002). The effects of price on alcohol consumption and alcohol-related problems. *Alcohol Research and Health* **26**, 22–34.
Duffy, M. (1996). Econometric studies of advertising, advertising restrictions, and cigarette demand: A survey. *International Journal of Advertising* **15**, 1–23.
Federal Trade Commission (2008). Marketing food to children and adolescents. a review of industry expenditures, activities, and self-regulation. A Report to Congress. Available at http://www.ftc.gov (accessed 20.09.08).
Frank, R. G. Berndt, E. R. Donohue, J. M. Epstein, A. and Rosenthal, M. (2002).*Trends in direct-to-consumer advertising of prescription drugs*. Kaiser Family Foundation. Available at: http://www.kff.org/rxdrugs/loader.cfmurl=/commonspot/security/getfile.cfm&PageID=14881 (accessed 26.08.09).
General Accounting Office (2002). *Prescription drugs: FDA oversight of direct-to-consumer advertising has limitations. Report to Congressional Reporters*. Washington, DC: U.S. General Accounting Office.
Grabowski, H. G. (1976). The effect of advertising on the inter-industry distribution of demand. *Explorations in Economic Research* **3**, 21–75.
Hamilton, J. L. (1972). Advertising, the health scare, and the cigarette advertising ban. *Review of Economics and Statistics* **54**, 401–411.

Kalyanaram, G. (2008). The order of entry effect in prescription (Rx) and over-the-counter (OTC) pharmaceutical drugs. *International Journal of Pharmaceutical and Healthcare Marketing* **2**(1), 35–46.

Kalyanaram, G. (2009). The endogenous modeling of the effect of direct advertising to consumers (DTCA) in prescription drugs. *International Journal of Pharmaceutical and Healthcare Marketing* **3**(2), 137–148.

National Institute on Alcohol Abuse and Alcoholism (2000). *Alcohol and Health, 10th Special Report to Congress*. Washington, DC: U.S. Department of Health and Human Services.

Pollay, R. W. (1994). Promises, promises: Self-regulation of the U.S. cigarette broadcast advertising in the 1960s. *Tobacco Control* **3**, 134–144.

Posner, R. (1973). *Regulation and advertising by the FTC*. Washington, DC: American Enterprise Institute.

Saffer, H. (1993). Alcohol advertising bans and alcohol abuse: Reply. *Journal of Health Economics* **12**(2), 229–234.

Saffer, H. (1998). Economic issues in cigarette and alcohol advertising. *Journal of Drug Issues* **28**(3), 781–793.

Saffer, H. and Dave, D. (2002). Alcohol consumption and alcohol advertising bans. *Applied Economics* **34**, 1325–1334.

Seldon, B. J. and Doroodian, K. (1989). A simultaneous model of cigarette advertising: Effects on demand and industry response to public policy. *Review of Economics and Statistics* **71**, 673B7.

Thomas, L. A. (1999). Incumbent firms' response to entry: Price, advertising and new product introduction. *International Journal of Industrial Organization* **17**, 527–555.

Wilcox, G. B. and Vacker, B. (1992). Cigarette advertising and consumption in the United States. *International Journal of Advertising* **11**, 269–278.

Young, D. (1993). Alcohol advertising bans and alcohol abuse: Comment. *Journal of Health Economics* **12**, 213–228.

Advertising Health Care: Causes and Consequences

OR Straume, University of Minho, Braga, Portugal

Introduction

This overview starts by giving a brief introduction to the economic theory of advertising, including a short presentation of the two main models of advertising and discussion of how advertising affects market outcomes in light of these two models. These theoretical underpinnings are then used to discuss the causes and potential effects of advertising in health care markets, with tentative implications for the social desirability of such advertising. A distinction is made between advertising of health care providers (hospitals or physicians) and advertising of prescription drugs, which are treated separately in this overview. Not only does drug advertising constitute the main bulk of total advertising expenditures in health care markets but it also involves some particular issues (and controversies) that demand separate attention.

The Economics of Advertising

Advertising is a widespread feature of economic life and has been a major topic for economic research since the early twentieth century. This research has led to the emergence of two distinct and competing views about what advertising is, with very different implications about the effects – positive as well as normative – of advertising. We can think of these as two different models of advertising.

Informative versus Persuasive Advertising

The informative advertising model takes as a starting point that most markets are characterized by asymmetric information, where consumers are ex ante imperfectly informed and need to search for information about products offered in the market. Because this search is costly, too few consumers will learn about the existence, price, and quality of products, causing market inefficiencies. According to the informative advertising model, advertising is a means to convey product information to consumers, which reduces consumers' search costs and thus reduces the inefficiencies caused by asymmetric information.

The persuasive advertising model, however, has a very different starting point. According to this view, the main purpose and effect of advertising is to change consumers' tastes and perceptions about the advertised product. Advertising is therefore a means to create 'artificial' product differentiation and brand loyalty, thereby increasing consumers' willingness to pay for the product. Whereas informative advertising has a positive effect in terms of reducing information imperfections, the persuasive view arguably implies that advertising is socially wasteful because its main effect is to distort the 'true' preferences of consumers.

Effects of Advertising

The informative and persuasive models of advertising predict very different effects of advertising, particularly with respect to competition and prices. Because real-life advertising rarely conforms to either of the two stylized models, but usually includes both persuasive and informative elements, an assessment of how advertising affects market outcomes, with corresponding implications for the social desirability of advertising, is a challenging exercise.

The main purpose of advertising is to increase demand for the advertised product. However, there are two sources of demand increases. Advertising could induce consumers to switch from a similar product offered by a competing firm toward the advertised one, or it could induce demand from new consumers who did not previously purchase any product from the market in question. The former is commonly referred to as business-stealing, whereas the latter is referred to as market expansion.

Advertising generally affects market prices. Theoretically, the price effects depend crucially on whether advertising is predominantly informative or persuasive. Informative advertising results in more consumers becoming aware of the existence and objective characteristics (including price) of available products. This makes demand more elastic (more price sensitive) and intensifies competition between competing brands, leading to lower prices in the market. However, persuasive advertising creates artificial product differentiation and brand loyalty, making consumers less willing to substitute between competing brands. This makes demand less elastic and allows firms to charge a higher price.

In many markets, with health care being a prime example, quality (rather than price) is a key characteristic of the products and services offered. Compared with the price effects, the effects of advertising on quality is theoretically less well established. If quality is observable and firms compete mainly on quality, informational advertising should lead to higher quality through increased competition.

However, quality is often not easily observable and it is therefore harder to assess to which extent advertising contains truthful information about quality. An important distinction can be made between search goods and experience goods. The quality of search goods can be ascertained before the purchase of the good, whereas the quality of experience goods can only be confirmed after the good is consumed. This suggests that producers of experience goods may have stronger incentives to advertise untruthfully about quality. However, in markets where consumers generally make repeated purchases, advertising in itself could function as a signal of high quality. Under the assumption that high-quality goods will be subject to more repeat purchases, producers of such goods will have incentives to advertise more to attract more first-time customers. This argument does not depend on the truthfulness of the advertising. Thus, seemingly persuasive advertising could

have an informational value as a signal of high quality. However, the empirical evidence of a positive relationship between advertising and quality is mixed.

Advertising might also affect entry of new firms/products into the market. The potential entry-deterring effect of advertising is a much-researched topic. From a theoretical viewpoint, it is possible that persuasive advertising might deter entry by creating brand loyalty to incumbent firms' products, implying that potential entrants would have to advertise more to capture these brand-loyal consumers, thereby increasing entry costs. However, there are also theoretical arguments suggesting that the optimal entry-deterring strategy is to underinvest in advertising. The key argument for this seemingly paradoxical result is that reducing the number of loyal consumers through lower advertising levels is a way for incumbent firms to commit themselves to higher output levels (or lower prices) in case of entry. Thus, incumbent firms might be able to deter entry by credibly committing themselves, through low pre-entry advertising levels, to become tough competitors post-entry. In either case, the empirical evidence of entry deterrence through advertising remains ambiguous.

Advertising of Health Care Providers

Below, the basic economics concepts and theories of advertising outlined above are used to discuss advertising in health care markets specifically. The present section deals with advertising of health care providers (hospitals or physicians) while the subsequent section deals with drug advertising.

Why Do Health Care Providers Advertise?

Because advertising is a means to increase demand, health care providers have incentives to advertise only as long as they can increase demand. Thus, incentives for advertising essentially require that providers' revenues are positively correlated with demand and that patients are able to choose their preferred provider. It is therefore no coincidence that health care advertising is mainly done by private health care providers. Traditionally, health care advertising has been more prevalent in the US, which has experienced health care advertising since the 1970s, particularly from for-profit providers. However, the introduction of market-based reforms in several European countries has made advertising relevant also for public (government-funded) health care providers, resulting, for example, in the lifting of the advertising ban on UK hospitals in 2008. In general, more competition in health care markets have been accompanied by a huge increase in advertising by health care providers (both hospitals and physicians) over the past couple of decades, although the advertising intensity in the health care sector remains as a relatively low fraction of total spending.

Is Heath Care Advertising Informative or Persuasive?

An important characteristic of health care markets is the high degree of asymmetric information, in which providers have generally much more information than patients about the quality of the services offered. This makes informative advertising potentially more valuable as a means to reduce informational market imperfections. However, because health services are often complex and highly nonstandard products that make information harder to assess and compare, this arguably also increases the scope for persuasive advertising (for example, by using celebrities to endorse products or services). The slower consumers revise their beliefs about quality, the stronger the incentive to mislead consumers through persuasive advertising. However, as previously argued, even purely persuasive advertising might have informational value if it functions as a signal of quality. This argument is clearly applicable to health care services, which are better characterized as experience goods rather than search goods.

Although the empirical evidence is scant, there exists research indicating that physician advertising leads to higher prices, which suggests that such advertising is predominantly persuasive. However, this is clearly an under-researched topic in the health economics literature.

Direct-to-Consumer Advertising and the Role of Physicians

A distinguishing feature of health care markets is that demand for health care is often a result of the interaction between patients and physicians, where, in most health care systems, general practitioners (GPs) act as gatekeepers to secondary health care (hospitals) through their referral decisions. Consequently, the effects of direct-to-consumer advertising (DTCA) must be analyzed and understood in the context of the physician–patient relationship.

DTCA can in principle have two different effects on demand; it can increase the number of patients seeking treatment for a particular condition and affect the choice of health care provider for patients seeking treatment. In health care systems that practice GP gatekeeping, the latter effect is determined by the patient–physician relationship. In a gatekeeping system, GPs provide information and affect patient choices. However, a more educated population arguably implies that patients play a more active role (vis-à-vis the GP) in the process of choosing health care providers, and DTCA is an alternative source of information for patients.

If the GPs are well-informed perfect agents for patients, there is little or no role for positive effects of DTCA. In this case, the patient will only disagree with the GP's recommendation if he is being misled by false advertising. However, GPs may not be perfect agents for their patients, either because GPs are not perfectly informed about available treatments or the two parties have different preferences with respect to the type of information they value. For example, GPs may care less about price information than patients do. Thus, to the extent that DTCA conveys accurate and relevant information to the patient, it may have positive effects in terms of reducing provider–patient mismatches if GPs are not perfectly informed or they do not always act in the best interest of the patient.

The above discussion ignores the potential effect of DTCA on the number of physician visits. A more thorough discussion of DTCA will be given in the context of drug advertising in the Section Advertising of Prescription Drugs, in which this is a more contentious issue.

Is Health Care Advertising Socially Wasteful?

If total demand for health care is relatively advertising-inelastic, the effect of advertising is mainly business-stealing. This could improve the matching between patients and providers, but it could also imply a waste of resources, as a form of 'medical arms race.' This depends on the extent to which advertising works as an instrument to reduce information imperfections in the health care market. Informational advertising can also have a positive welfare effect if it lowers prices (or raises quality) through increased competition.

If advertising leads to a demand expansion, this could still be socially wasteful if this expansion is 'artificially' created by persuasive advertising, leading to overconsumption of health care. However, even persuasive advertising can have positive welfare effects if such advertising works as a reliable signal of quality, as discussed in the Section The Economics of Advertising.

Advertising of Prescription Drugs

In contrast to health care providers (physician or hospitals), pharmaceutical companies spend a considerable share of revenues on advertising, often exceeding the share spent on research and development of new drugs.

With respect to advertising, there is a key distinction between prescription drugs and so-called over-the-counter (OTC) drugs. Because OTCs may be sold directly to consumers without a physician's prescription, the natural advertising target is therefore consumers. For prescription drugs, by contrast, there are potentially two different advertising targets: consumers and physicians. Therefore, prescription drugs are advertised through two different channels: DTCA (if allowed) and physician detailing. In the following, the two different channels of prescription drug marketing will be discussed and compared.

Direct-to-Consumer Drug Advertising

In contrast to advertising of OTC drugs, DTCA of prescription drugs is currently banned in all developed countries, except in the USA and New Zealand, although steps towards liberalization have been taken in several countries. In the US, DTCA has been allowed since the 1980s, though subject to regulation. New and more liberal guidelines were adopted in 1997.

What are the main effects of direct-to-consumer drug advertising? There is little doubt that DTCA results in an increased total number of drug prescriptions, the most important contributing factor being that DTCA increases demand for physician consultations. Thus, in addition to direct advertising costs, there are considerable indirect costs of DTCA because of a higher number of physician consultations and more drug prescriptions. The extent to which these costs are outweighed by higher patient benefits depend on whether advertising-induced consultations are necessary or unnecessary, and whether advertising-induced prescriptions are cost-effective or not.

Like advertising of health care providers, direct-to-consumer drug advertising could also affect competition between pharmaceutical companies and thus drug prices. Drug advertising does not normally contain price information, but increased information about the existence of competing drug therapies may increase competition and lead to lower prices. Although DTCA is mainly undertaken by patent-holding firms, these are seldom pure monopolies due to the existence of therapeutic substitutes in many submarkets.

The contentious nature of DTCA of prescription drugs, reflected in the widespread ban on such activities, requires a more thorough discussion of the relevant arguments.

A main argument in favor of DTCA is that it contributes to consumer education by increasing awareness of alternative drug treatments. This is the standard informative advertising viewpoint, and the validity of this argument clearly relies on the informational content of DTCA. However, another important side-effect of DTCA is that information about alternative drug treatments may also increase consumer awareness about the underlying medical conditions, thus increasing the likelihood that potentially serious diseases are detected at an earlier stage.

Besides the potential for reducing informational inefficiencies, DTCA arguably also promotes greater patient autonomy by motivating patients to play a more active role in their treatment. One could also argue that DTCA works to counterbalance the effect of physician detailing. If persuasive drug detailing towards physicians leads to a distortion of prescription choices, this could partly be corrected by making patients better informed about alternative drug treatments through DTCA.

However, several arguments have been put forward against allowing DTCA of prescription drugs. Although DTCA has the potential to reduce inefficiencies caused by imperfect information on the demand side of the market, this requires that consumers are equipped with sufficient background knowledge to understand and properly evaluate the information given by DTCA. If this is not the case, DTCA might lead consumers to demand drug treatment against medical conditions that are either nonexistent or better left untreated. Thus, DTCA might induce overconsumption of drugs and encourage the use of unnecessary medication. Similarly, DTCA might also contribute to overmedication by creating a bias in favor of drug treatment instead of nonpharmacological interventions, such as lifestyle changes.

Although DTCA can have positive effects in terms of promoting greater patient autonomy, there is also a potential flip side. If DTCA has mainly a persuasive, rather than an informative effect, this might introduce more costs and strains in the physician–patient relationship, in which physicians have to spend more time correcting misinformed views because of DTCA. Physicians might also face increased pressure from patients to prescribe new and less-well tested drugs.

DTCA versus Physician Detailing

Although DTCA is banned in most countries, advertising targeted towards physicians – so called detailing – is generally allowed (though regulated). Indeed, physician detailing constitutes the main share of total drug marketing expenditures.

This form of drug marketing includes visits by sales representatives to physicians, as well as advertising in medical journals. Because face-to-face advertising is more costly, the likely impact on prescription choices is also higher.

Like DTCA, physician detailing can, in principle, have both market-expanding and business-stealing effects. It has a market-expanding effect if it increases physicians' propensity to choose drug treatment over nonpharmacological treatments, and it has a business-stealing effect if it affects physicians' propensity to prescribe drug treatment A over drug treatment B. Like other types of advertising, detailing can reduce informational inefficiencies and improve the matches between medical conditions and drug treatments, if the informational content of this type of marketing is sufficiently high. However, it would be naive to disregard the possibility that there is also a substantial persuasive element to physician detailing. In fact, empirical studies showing that detailing reduces the price elasticity of demand suggest the existence of a significant persuasive effect.

An interesting question is whether DTCA and physician detailing are complement or substitute marketing strategies for pharmaceutical companies. Although detailing clearly affects prescription choices, empirical evidence suggests that DTCA has a larger effect on physician visits than on prescription choices, implying that DTCA mainly has a market-expanding effect. If the effect of detailing is mainly business stealing, while the effect of DTCA is mainly market expansion, this suggests that detailing and DTCA are complement strategies: More DTCA leads to a higher number of physician visits, which increases the profitability of spending resources on physician detailing to influence prescription choices.

Thus, if DTCA and physician detailing are complement strategies, an unintended side-effect of allowing DTCA is that it would lead to increased levels of physician detailing as well.

Drug Advertising and Generic Competition

A major concern for policy makers and regulators of pharmaceutical markets is to ensure that competition in the off-patent market is sufficiently stimulated. An important question in this respect is how advertising affects competition in the off-patent market. More specifically, how does advertising affect the probability of generic entry and how does it affect price competition between brand name and generic drugs?

A robust empirical regularity in the off-patent market for prescription drugs is that brand name drugs are consistently priced higher than their generic versions. Some early empirical studies even found that brand name prices tended to increase after generic entry. From an economics perspective, the persistent positive price difference between brand name and generic drugs is somewhat puzzling, as competition between homogeneous products (brand names and their generic versions) should be expected to lead to fierce price competition with uniformly low drug prices as a result. This strongly suggests that brand name and generic drugs are vertically differentiated in the eyes of consumers (or prescribing physicians), where brand name drugs are somehow perceived to be of higher quality. The most prominent theoretical explanation for the price difference between brand names and generics is that it is a result of persuasive advertising of the brand name drug during the patent period, creating a brand-loyalty that allows for brand-name drugs to be charged a higher price than its generic alternatives after patent expiry. Given that brand-name and generic drug versions have, by definition, identical active chemical ingredients and absorption rates, the observed price difference is a strong indicator of a significant persuasive element in drug advertising, which is usually considered to be detrimental for welfare.

The vertical differentiation created by brand-name drug advertising relaxes price competition in the off-patent market and allows for higher prices, not only of the brand-name drugs but also of the generic competitors. This suggests that brand-name drug advertising has potentially two counteracting effects on generic entry. On the one hand, persuasive advertising creates brand-loyalty that, all else being equal, reduces demand for generics and makes generic entry less profitable. However, such advertising creates 'artificial' vertical differentiation and relaxes price competition, which, all else being equal, makes generic entry more profitable. The second effect is more likely to dominate if advertising also has a market-expanding effect, which allows for generally higher drug prices in the market.

Whether advertising stimulates or deters generic entry (i.e., which of the two mentioned effects dominates) depends crucially on the strictness of price regulation in the off-patent market. If price regulation is very strict, advertising leads to brand-loyalty without a corresponding increase in prices. In this case, advertising is likely to have an entry-deterring effect. However, the price competition effect might dominate if price regulation in the off-patent market is absent or sufficiently lax.

The above reasoning implies that even purely persuasive advertising might have positive welfare effects if it induces generic entry after patent expiration. Persuasive advertising relaxes price competition by creating artificial vertical differentiation, but this might also induce generic entry that would otherwise have been deterred because of strong price competition (in the absence of advertising).

Notice that the above discussion implies that, to the extent that brand-name drug producers can deter entry through advertising, the nature of the optimal entry-deterring strategy is a priori ambiguous. Patent-holding firms might overinvest in advertising in order to build up brand-loyalty and thereby make generic entry less profitable. However, because advertising may partly benefit generic entrants, through market expansion and relaxed price competition, the optimal entry-deterring strategy might instead be to underinvest in advertising.

Finally, although the results are somewhat mixed and inconclusive, the empirical literature on strategic entry deterrence in pharmaceutical markets does not seem to produce very strong evidence that brand name advertising deters entry.

See also: Advertising as a Determinant of Health in the USA. Competition on the Hospital Sector. Pharmaceutical Marketing and Promotion. Physician-Induced Demand

Further Reading

Bagwell, K. (2007). The economic analysis of advertising. In Armstrong, M. and Porter, R. (eds.) *Handbook of industrial organization*, Vol. 3, pp. 1701–1844. Amsterdam: Elsevier.

Brekke, K. R. and Kuhn, M. (2006). Direct to consumer advertising in pharmaceutical markets. *Journal of Health Economics* **25**, 102–130.

Ellison, G. and Ellison, S. F. (2011). Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. *American Economic Journal: Microeconomics* **3**, 1–36.

Iizuka, T. (2004). What explains the use of direct-to-consumer advertising of prescription drugs? *Journal of Industrial Economics* **52**, 349–379.

Königbauer, I. (2007). Advertising and generic market entry. *Journal of Health Economics* **26**, 286–305.

Rizzo, J. A. and Zeckhauser, R. J. (1989). Advertising and the price, quantity, and quality of primary care physician services. *Journal of Human Resources* **27**, 381–421.

Scott Morton, F. M. (2000). Barriers to entry, brand advertising, and generic entry in the us pharmaceutical industry. *International Journal of Industrial Organization* **18**, 1085–1104.

Wilkes, M. S., Bell, R. A. and Kravitz, R. L. (2000). Direct-to-consumer prescription drug advertising: Trends, impact, and implications. *Health Affairs* **19**, 110–128.

Aging: Health at Advanced Ages

GJ van den Berg, University of Mannheim, Mannheim, Germany; IFAU Uppsala; VU University Amsterdam, and IZA
M Lindeboom, VU University Amsterdam, HV Amsterdam, The Netherlands

Introduction

This article examines how health and mortality at advanced ages evolves from conditions early in life. Here, the authors summarize the findings, examine econometric strategies to identify causal effects, and discuss the implications of the findings for public policies aimed at improving population health.

The larger part of health care that individuals consume during their life course is concentrated in the final few years of their life. Proximity to death may be the driving factor of these costs, but age may also have an additional effect on healthcare spending. The latter view is in line with a simple health capital model and implies that in the context of the trend toward aging, increases in healthcare costs are to be expected. More in general, healthcare costs across cohorts vary if mortality and morbidity rates differ across age cohorts. A second empirical observation is that health is known to be very unevenly distributed at advanced ages.

Socioeconomic differences are important determinants of late-life health variation across individuals. There is a strong connection all over the industrialized world between an individual's current socioeconomic status (SES) and his/her current health (the association between income and health is commonly denoted as 'the gradient'). The magnitude of this gradient differs across countries, and SES-related inequality in health has increased over the past decades. Clearly, the statistical relation between SES and health can also be explained by a reverse causality from health to SES, or by a mutual dependence of SES and health on common determinants such as genetic characteristics, education, or conditions early in life. This naturally leads to a dynamic view in which causal pathways between various factors may create associations between SES and health at different stages of life.

Recent evidence suggests that much of the association between SES and health during middle age and old age is driven by a causal effect of health on SES, rather than the other way. Furthermore, already at relatively young ages, substantial health differences exist between different SES groups. Recent papers in this area (see Van den Berg and Lindeboom, 2007, for a survey) suggest that the determinants of health and SES-related differences in health may originate earlier in life. Heckman *et al*. (2006) show that "early intervention programs targeted to disadvantaged children have had their biggest effect on noncognitive skills: motivation, self-control, and time preference," and that these noncognitive skills are powerful predictors of educational attainment, lifestyle, and health behaviors. Their work also shows that for severely disadvantaged children early-childhood interventions are important and can have a long-lasting effect on cognitive and noncognitive functioning.

Motivated by the above, the authors therefore start with a discussion of the relationships between conditions early in childhood and later-life health. Section Causal Effects of Early-Life Conditions reviews the epidemiological and economic literature in this field, presents evidence of the importance of early-childhood conditions for later-life outcomes, discusses the methodological problems in this area when researchers have to rely on observational data, and proposes appropriate research designs that allow one to assess the causal effect of early-childhood conditions on health and mortality later in life. Section Indirect Effects: Causal Pathways from Early Childhood by Way of Education to Later-Life Morbidity and Mortality discusses mechanisms that may underlie the causal effect of early-childhood conditions, focusing on the role of education. Section Summary and Implications for Health Policy concludes and addresses policy implications.

Causal Effects of Early-Life Conditions

Empirical Approaches and Empirical Findings

For expositional reasons, this section begins with a subsection on the methodological approaches used in the empirical literature to detect long-run effects of early-life conditions. This includes a discussion of empirical findings that capture the overall causal effect. The overall effect can be a direct causal effect or it can be the result of a causal pathway that involves intermediate events during life. Section Direct and Indirect Long-Run Effects discusses the difference between direct and indirect effects in more detail. Section Indirect Effects: Causal Pathways from Early Childhood by Way of Education to Later-Life Morbidity and Mortality discusses empirical studies of indirect effects that include data information on events occurring along the pathway of interest.

A natural starting point to analyze whether early-life conditions are important is to compare health and mortality outcomes among elderly individuals who faced different living conditions early in life. Empirical studies have shown that adverse socioeconomic conditions early in life are associated with susceptibility to a wide range of health problems later in life. Similarly, medical studies have shown that individuals with a low birth weight (sometimes adjusted for gestation time) are more likely to suffer from health problems later in life.

Observed associations do not necessarily imply the presence of causal effects of early-life conditions. Individual socioeconomic and medical conditions during early childhood and health outcomes later in life may be jointly affected by unobserved heterogeneity. For example, certain genes may simultaneously influence the average level of the parents' income, the birth weight, and the health outcomes later in life. To be able to detect causal effects, one needs to observe exogenous variation in the early-life conditions, and relate this to outcomes later in life. In all fairness, it should be noted that

 doi:10.1016/B978-0-12-375678-7.00302-3

even if descriptive studies do not capture causal effects, they are still useful from an intervention point of view. Markers for unfavorable future health outcomes can be used as a flag for monitoring or initiating interventions to mitigate such outcomes.

A recent approach has recently become popular to detect causal effects, by using data on indicators Z of individual conditions X early in life with the following property: the only way in which the indicator Z can plausibly affect high-age morbidity or mortality Y is by way of the individual early-life conditions X. (An extreme example is where Z is the outcome of a lottery in which individuals with a baby may win some money. More common examples are given below.) By analogy to the econometrics literature, such indicators Z may be called instrumental variables. Typically, these are not unique characteristics of the newborn individual, his/her family, or household, but rather temporary characteristics of the macro-environment into which the child is born. In that case they are also called contextual variables. Indicators Z with the above 'exclusion restriction' property do not give rise to endogeneity and simultaneity biases, because they are exogenous from the individual's point of view. Moreover, they do not have direct causal effects on health later in life except through early-life conditions. If one observes an association between such an indicator Z and the health outcome Y later in life, then one can conclude that there is a causal effect of early-life conditions X on that health outcome Y.

In the current context, three types of such 'instrumental variables' Z may be distinguished. First is the season of birth. The idea is that the month of birth has no other effect on health outcomes later in life than by way of the early-life conditions of the child. Note that this requires that the composition of newborns is not systematically different across seasons, in terms of unobserved characteristics of the newborns. The literature has typically found significant effects of the season of birth on the mortality rate later in life, with an order of magnitude of a few months of extra lifetime if one is born in the fall, as compared with the late spring. In the southern hemisphere, these effects are mirror-imaged, in the sense that the effect of a month of birth is similar to the effect of the month half a year earlier or later in the other hemisphere. In equatorial areas, seasonal effects are in accordance to what constitutes the rainy (monsoon) and the dry season.

A second type of exogenous variation is provided by epidemics, wars, famines, and other disastrous events. Lumey *et al.* (2011) provide an excellent overview. For a recent example, see Lindeboom *et al.* (2010), who examine whether exposure to nutritional shocks early in life affects later-life mortality. They use historical data that include the period of 1845–48, which includes the Dutch potato famine. During this period, potato crops failed due to the Potato Blight disease and bad weather conditions. They found strong evidence for long-run effects of exposure to the Potato famine. The results were stronger for boys than girls and lower social classes appeared to be more affected than higher social classes. Studies based on the Dutch 'hunger winter' under German occupation at the end of World War II and on China's great famine indicated significant long-run effects on adult morbidity, but not on adult mortality. These studies confirmed that malnutrition has a separate effect on adult morbidity (and sometimes) mortality. Experimental animal research has also provided support for the theory that there are long-run effects of malnutrition during pregnancy.

Almond (2002) examines individuals born around the time of the 1918 influenza epidemic. He finds significant effects on the mortality rate later in life, and this finding has been confirmed by subsequent studies using epidemics. Similar to many of these studies, Almond investigates primarily the sign and significance of the mortality-rate differences between birth cohorts, and not the exact size of the effect. This is because the interest ultimately is not in the size of the effect of the indicator Z on the mortality rate, but in the issue of whether there is a causal effect from early-life conditions X on the mortality rate. Long-run effects may, of course, be nonlinear in terms of early-life conditions. In that case, the relevance of long-run effects of disastrous conditions may be limited, and may not lead to a full understanding of the effects of less spectacular variation in early-life conditions.

A third approach was pioneered by Bengtsson and Lindström (2000). They use the transitory component (or deviation) in the price of rye around the time of birth as an indicator of food accessibility early in life – any observed relation between this indicator and the mortality rate later in life signifies the existence of a long-run causal effect of food accessibility on mortality later in life. Similarly, the transitory component in the local infant mortality rate was used as an indicator of exposure to diseases early in life. This study uses data from a relatively small area in Sweden from the eighteenth and nineteenth centuries. The results indicate that individuals born in years with epidemics lived on average a few years less than otherwise, conditional on surviving the epidemic itself. Van den Berg *et al.* (2006) use the state of the business cycle at early ages as a determinant of individual mortality. Cyclical macroeconomic conditions during the pregnancy of the mother and childhood might affect mortality later in life because they are unanticipated and affect household income. In a recession, the provision of sufficient nutrients and good living conditions for children and pregnant women may be hampered. Van den Berg *et al.* (2006) find that the average lifetime duration in the Netherlands in the nineteenth century was reduced by approximately 1–3 years if the individual is born in a recession, as compared with having been born in a boom (under otherwise identical conditions during life, and conditional on surviving early childhood). Van den Berg *et al.* (2011) find analogous effects on cardiovascular mortality, using Danish data.

One important requirement for the analysis of causal long-run effects of early-life conditions is that the individual data cover a sufficiently long time span. After all, the dates of birth and death (or high-age health) must be observed for a substantial number of individuals. An implication of this requirement is that the existing studies have necessarily considered cohorts of individuals who were born a long time ago. In this sense, the most recent evidence comes from studies of individuals born in the Dutch hunger winter (1944–45) and from studies of more recent birth cohorts from developing countries. One way to circumvent this restriction would be to focus on adult health proxies such as adult height (see the upcoming sections).

Direct and Indirect Long-Run Effects

Empirical Approaches and Empirical Findings listed studies that use exogenous variation in the environment to show that there are causal effects from early childhood on later-life morbidity and mortality. The present subsection briefly sets out the main mechanisms underlying these long-term causal effects. Although there are many ways in which early-life conditions may affect outcomes later in life, it can be distinguished roughly between two main views.

First, adverse prenatal and postneonatal (from birth to 12 months) conditions can have a direct effect on later-life morbidity and mortality. The main idea is that the development of vital organs and the immune system is programmed when the body is exposed prenatally or just after birth to adverse conditions. According to the 'developmental programming' or 'fetal origins' hypothesis), this may lead to increased vulnerability to chronic diseases in later life. The most commonly mentioned factors mentioned in the literature are malnutrition and exposure to infectious diseases. Other factors are increased stress in the household and lower income to cover housing accommodation costs. Most of the empirical studies mentioned in this section are consistent with a direct effect. As it can be seen, in order to detect long-run effects, it is natural to focus on temporary shocks around the birth date. Any long-run effect found in this way could be a direct effect. Moreover, the estimated size of the mortality effects is usually moderate and in line with the medical evidence. The type of shock is informative regarding whether the effect concerns malnutrition, disease exposure, other adverse conditions, or just bad conditions in general.

Exposure to infectious diseases and malnutrition is likely to be less relevant for the developed world today than it was in the past. However, Bozzoli *et al*. (2009) recently examined the effect of income and disease exposure on adult height in populations, where height is used as a proxy for lifetime health. They use postneonatal mortality as a measure for nutrition and disease load in early childhood and examine their effect on height for cohorts born from 1950 to 1980 in the US and 11 European countries. They find a strong negative relationship between adult height and the burden of disease and malnutrition.

According to the second main view, adverse conditions early in life have indirect effects in that they may be the start of a causal chain of events or pathways during life that leads to worse health later in life. For instance, poor early-life conditions may lead to poor health early in life and later in childhood, which may affect educational outcomes and subsequently social status and health in adulthood. Or, more generally, a poor start may affect an individual's life career, which may ultimately lead to higher mortality rates. The authors discuss this view in more detail below, but before that it is good to note that some studies have stressed that it is the interaction with social factors later in life that determines whether people who are exposed to adverse early-childhood conditions will be more vulnerable to ill health in later life. For example, among individuals born in recessions, the decline in mental fitness after experiencing a negative life event at high ages (such as a stroke, surgery, illness, or death of a family member) is worse. Among women, marriage leads to increased mortality in child-bearing ages, but this increase is smaller if the woman was born under favorable economic conditions around birth, as captured by the business cycle early in life. In a similar vein, the body accommodates to stress, and that it is repeated stress that leads to higher risks of chronic diseases.

Indirect Effects: Causal Pathways from Early Childhood by Way of Education to Later-Life Morbidity and Mortality

Figure 1 shows the main causal pathways that are considered. Note that compared to all the previous sections, the setting has been expanded: it is not restricted to the pathways that can be tracked down to the causes early in life, but also other possible determinants of later health are considered. The direct effect that links infant health to later-life morbidity and mortality is not discussed explicitly here (see Section Causal Effects of Early-Life Conditions).

Note that the methodological complications in the case of indirect effects are even larger than in the case of direct effects. In the former case, most studies typically restrict attention to just one of the arrows in the diagram, conditioning on the individual position at the starting point of the arrow. In general, this starting position can be endogenously affected by earlier events in the life of the individual or by unobserved determinants that also have a causal effect on the outcome.

The Effect of Child Health on Educational Attainment

Quite a few studies in the development literature study the effect of child health or child nutrition on schooling outcomes. Ordinary least squares (OLS) estimates generally suggest a strong association between child health or nutrition and educational attainment. Several studies have tried to assess the causal effect of child health via Instrumental Variable approaches and sibling fixed-effect approaches. These studies seem to confirm the naïve OLS estimates, but the size of the effect is generally larger. Miguel and Kremer (2004) used a

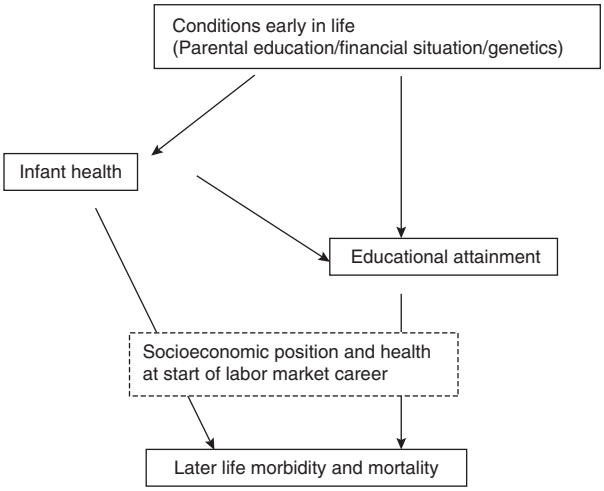


Figure 1 A graphic representation of the indirect effects of early-childhood conditions.

randomized experiment to evaluate a program of a school-based treatment with a deworming drug in Kenya and found that absenteeism in treatment schools was substantially lower than in comparison schools, and that deworming increased schooling by approximately 1 month per pupil treated.

The literature for developed countries is small. Case *et al.* (2005) used British data from the Child Development Study to look at (among other things) the effect of childhood health on educational attainment. They found a strong association between childhood health and later educational attainment. It appears that the presence of chronic conditions in childhood has a stronger impact on educational attainment than does health at puberty. Their conclusion: the negative effect of bad health is cumulative in its effect on education. These results are based on observational data that follow a single cohort, which makes it difficult to make causal statements. Case and Paxson (2006) use adult height as a measure for childhood conditions and childhood health, and find that the height premium in adulthood (i.e., better labor market outcomes for taller people) can be explained by childhood scores on cognitive tests and by the fact that taller children selected into occupations that have higher cognitive skill requirements. Currie and Stabile (2003) examine the relationship between several common health disorders, such as attention deficit hyperactivity disorder (ADHD), depression, anxiety, and aggression, on future educational outcomes. They conclude that early-childhood mental health problems affect educational outcomes and that there is little evidence that income protects against the negative effects of mental health. A recent and innovative approach of Ding *et al.* (2006) focuses on a specific set of conditions (ADHD, depression, and obesity), and uses genetic markers that strongly predict these conditions as instruments. They find strong effects of these health conditions on student grade point averages. The larger part of this effect seems to be driven by the effect for females; for males they find no effect.

The Effect of Education on Later Health and Mortality

Since Cutler and Lleras–Muney (2007) recently provided an excellent review of the literature on education and health, there is no need to fully review the papers discussed in their study, and can be drawn from their findings. Cutler and Lleras Muney performed some analyses of their own that confirm the strong association between education and (later-life) health. There is evidence for a causal effect of education on health. The most convincing evidence comes from studies that use changes in minimum schooling laws. This implies that one can make statements about the effect of additional schooling only regarding those who are at the bottom of the schooling distribution. Identifying which mechanisms generate these causal impacts remains speculative. The better educated have the better jobs and higher incomes, which may lead to better health and lower mortality rates at later ages. Case and Deaton (2003) find that people in manual occupations have worse self-reported health, and that there is a greater rate of health declines in these occupations. Their argument: much of the differences in health are driven by health-related absence from the labor force. Smith (2005) found that current and lagged financial measures of SES have no effect on future health, but that education does. This holds for older and for younger workers, thereby suggesting a potentially important role for factors such as the rank in the social distribution, the ability to process information and health behaviors.

The Whitehall studies of British civil servants show that morbidity and mortality fall with increases in social class. A low position in the social distribution leads to low control and high (job) demands, which in turn lead to stress, which puts workers at risk for cardiovascular disease. There is a strong relation between a measure for control and cardiovascular disease risk and this relationship also holds for non-civil servants. Cutler and Lleras–Muney (2007) argue that social position cannot be the main determinant of SES-related health differences. Life expectancy has increased in the developed world over the past three decades, although income inequality and crime have increased and social networks generally have become smaller. Also, some studies have shown that there are gradients in diseases that are not related to stress.

Schooling provides individuals with skills that help them acquire and process information, which helps them make better decisions. For example, consumer health information has been shown to increase the demand for medical services. More information increases the probability of care use, but conditional on care use, the quantity of care use is not related to information. Apparently, poorly informed consumers tend to underestimate the productivity of medical care in treating disease. However, differences in knowledge by SES create only modest differences in health behaviors by SES. Indeed, as noted by Cutler and Lleras–Muney (2007), although both educated and uneducated people today are well aware of the dangers involved with smoking, smoking is still more prevalent among the uneducated. Of interest is whether this association between smoking and schooling is causal, and if so, what mechanisms drive this effect. This can be addressed using Vietnam draft-avoidance behavior as an instrument for college attendance. The cohort of males born between 1945 and 1950 could avoid the Vietnam draft by enrolling into college, and this can be used as an instrument for college enrolment. The female cohort born between 1945 and 1950 can be used as a control group. It turns out that the level of education does causally affect smoking, and that those who initiated smoking are more likely to stop once they enter college. Peer effects or endogenous time preferences are likely to be important determinants. Improved information-processing capabilities due to increased schooling do not seem to be important. Subjective time discount rates are not related to smoking, but more general measures of time preference and self-control, such as impulsivity and financial planning, are related to smoking.

Summary and Implications for Health Policy

The literature suggests that long-run effects of early-childhood conditions are important for morbidity and mortality later in life. There are roughly two channels: direct long-run effects due to 'programming,' and indirect effects via education, health, and SES at different points in the life course.

Direct effects are likely to be quantitatively relevant for developing countries, where exposure to extreme conditions is more common, and where behavior later in life may be less successful in mitigating early-life effects. There are, however, some other studies that point toward the relevance of environmental insults, disease exposure and malnutrition for cohorts born in the twentieth century in developed countries. Of importance for *healthcare* policy is that this suggests that one can expect mortality differentials across different cohorts and that the younger cohorts do not necessarily live longer in better health. Also, policies focused on vulnerable families (those living in poor circumstances, exposed to stress, and employing bad health behaviors) can be effective in improving the health of the next generation.

Childhood conditions may affect child health, and this may persist into adulthood. The evidence on the effect of family income is mixed, at least for developed countries – although any effect that might be found is expected to be modest. Most studies point at a potentially strong role for the family-specific environment. This includes parenting skills, health behaviors, and maternal and paternal health. Maternal health is probably the most important determinant for child health. This does not mean that there is no role for health policies. Policies aimed at improving the health of young adolescents can be effective in improving the health of the next generation. These interventions may reverse the impact of a poor start early in life and improve health in adolescence and beyond.

Education is undoubtedly one of the strongest determinants of health in later life. Education increases income and labor market opportunities and positively affects health-enhancing behavior. The effect of education on health behavior is causal and likely to be of core importance for health later in life. Policies focusing on educational outcomes should intervene at early ages. Recent work Heckman *et al.* (2006) shows that early intervention programs targeted to disadvantaged children have their biggest impact on noncognitive skills such as motivation, self-control, and time preference. Studies cited in The Effect of Education on Later Health and Mortality show the importance of these factors for health behaviors. Heckman *et al.* (2006) show that these noncognitive skills strongly influence schooling decisions and later wages.

In sum, with new cohorts one should focus on early health and education interventions. It would be useful to screen babies and young children at their household circumstances, to determine whether nutrition, heating, stress levels, and other indicators are at acceptable levels. Programs targeted to children of disadvantaged households should be implemented at an early age. Among existing cohorts, it is useful to screen individuals born in particularly adverse conditions, to verify whether they are susceptible to cardiovascular disease and other diseases thought to be programmed early in life.

It is important to emphasize that even if early-life conditions have a small overall effect on the per-period morbidity or mortality rate later in life, it may nevertheless be very important from a policy point of view to intervene in the lives of individuals with an adverse starting position. After all, the benefits of such interventions will be reaped over a very long time period, and intervention is facilitated by the fact that there is a time interval in between a particular cause and the moment its effect materializes. This is quite different from the instantaneous effects of current events on the health of elderly individuals, like a summer with unusually high temperatures. Such instantaneous effects may be large, but they may be relevant only over a short period, and policy makers would have to react very quickly to prevent the negative health implications.

See also: Education and Health. Fetal Origins of Lifetime Health

References

Almond, D. V. (2002) *Cohort differences in health: a duration analysis using the national longitudinal mortality study.* Working Paper, University of Chicago, Chicago.

Bengtsson, T. and Lindström, M. (2000). Childhood misery and disease in later life: The effects on mortality in old age of hazards experienced in early life, Southern Sweden, 1760–1894. *Population Studies* **54**, 263–277.

Bozzoli, C., Deaton, A and Quintana-Domeque, C. (2009). Adult height and childhood disease. *Demography* **46**(4), 647–669.

Case, A. C. and Deaton, A. (2003). Broken down by work and sex: How our health declines. *NBER Working Papers 9821*. National Bureau of Economic Research, Inc.

Case, A., Fertig, A. and Paxson, C. (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics* **24**, 365–389.

Case A., C. Paxson (2006) *Stature and status: Height, ability and labor market outcomes*. NBER working paper 12466.

Currie, J. and Stabile, M. (2003). Socioeconomic status and child health: Why is the relationship stronger for older children? *American Economic Review* **93**(5), 1813–1823.

Cutler, D. and A. Lleras–Muney (2007) *Education and Health: Evaluating Theories and Evidence*, NBER Working Paper 12352. Cambridge, MA.

Ding, W., Lehrer, S. F., Rosenquist, J. N. and Audrain-McGovern, J. (2006). The impact of poor health on education: New evidence using genetic markers. *NBER Working Papers 12304*. National Bureau of Economic Research, Inc.

Heckman, J. J., Stixrud, J. and Urzua, S. (2006). The effect of cognitive and non cognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24**(**3**), 411–482.

Lindeboom, M., Portrait, F. and van den Berg, G. J. (2010). Long-run effects on longevity of a nutritional shock early in life: The Dutch Potato Famine of 1846–1847. *Journal of Health Economics* **29**(5), 617–629.

Lumey, L. H., Stein, A. D. and Susser, E. (2011). Prenatal famine and adult health. *Annual Review of Public Health* **32**, 24.1–24.26.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**(1), 159–217.

Smith, J. P. (2005) *The Impact of SES on Health over the Life Course*. RAND working paper.

Van den Berg, G. J., Doblhammer, G. and Christensen, K. (2011). Being born under adverse economic conditions leads to a higher cardiovascular mortality rate later in life: Evidence based on individuals born at different stages of the business cycle. *Demography* **48**, 507–530.

Van den Berg, G. J., Lindeboom, M. and Portrait, F. (2006). Economic conditions early in life and individual mortality. *American Economic Review* **96**, 290–302.

Van den Berg,, G.J. and Lindeboom, M. (2007) *Birth is the messenger of death – but policy may help to postpone the bad news. New evidence on the importance of conditions early in life for health and mortality at advanced ages*. Netspar Panel Paper 3, Tilburg University, Tilburg.

Further Reading

Ravelli, A. C., van der Meulen, J. H., Michels, R. P., Osmonds, C. and Barker, D. J. (1998). Glucose tolerance in adults after prenatal exposure to famine. *Lancet* **351**, 173–177.

Alcohol

C Carpenter, Vanderbilt University, Nashville, TN, USA

Introduction

Alcohol is extremely prevalent in contemporary society. According to the World Health Organization, in 2005 the per capita alcohol consumption totaled 6.13 l of pure alcohol for every person age 15 and older worldwide. More than a quarter of this consumption is estimated to be from illegal or homemade production and thus not likely to be reflected in standard statistics on alcohol sales. People in the developed world drink much more heavily than people in less developed places such as sub-Saharan Africa. Populations with strong religious prohibitions on drinking (e.g., the Islamic faith) also exhibit much lower drinking rates. Beverage type varies substantially throughout the world: In many European and South American countries, wine is the primary alcoholic drink consumed. In the Western Hemisphere, Northern Europe, and Australia, beer is the most widely consumed alcoholic beverage. For example, in the US a little more than half of total alcohol sales is attributable to beer, approximately a third is spirits, and the remainder is wine. Worldwide, however, nearly half of the total consumption is attributable to neither beer nor wine but rather spirits (which is more common in southeast Asia).

Alcohol consumption has remained relatively stable throughout the world since 1990. With respect to demographic patterns worldwide, men are much more likely to drink and to drink more heavily than women, although it is notable that almost half of all men and two-thirds of all women in the world did not consume alcohol in the past 12 months. Heavy episodic drinking varies substantially across the world in complex ways. For example, it is not always the case that high per capita consumption is associated with higher rates of heavy episodic drinking: Many Western European countries, for example, have very high per capita consumption rates despite having low heavy episodic drinking rates, suggesting that patterns of drinking in those countries are more moderate. Moreover, it is not always the case that higher income countries have higher rates of heavy episodic drinking within a broad geographic area: In Europe and the Americas, for example, heavy episodic drinking is more prevalent in the lower income countries, whereas in Africa and southeast Asia, the relationship is reversed.

Alcohol consumption has both positive and negative aspects. The positives derive from the fact that people enjoy consuming alcohol, moderate alcohol has been suggested to have some health benefits, and there is ample evidence that drinkers earn more than abstainers in developed countries. The most commonly cited negatives include the problem that some of the social aspects of drinking and the direct pharmacological effects of alcohol can lead to a variety of adverse outcomes such as premature death and illness, crime, risky sexual activity, and alcohol dependence. Economists and economics have played an important role in informing the policy and academic debate about alcohol use and alcohol control by providing a conceptual framework for evaluating not only the costs but also the benefits of alcohol use when thinking about optimal alcohol control and by measuring and testing the relationships among alcohol use, alcohol control policies, and outcomes. This article discusses the economics of alcohol use and alcohol control policies and provides a very broad summary of what is known about the causes and consequences of alcohol consumption.

Alcohol's Pharmacological Profile

A substantial portion of the economics research on alcohol addresses whether and to what extent alcohol causes adverse outcomes such as premature death and morbidity. The most prominent channel through which these adverse events are thought to occur is biological. People's 'blood alcohol concentration' (BAC) from drinking affects their level of impairment. The most important determinant of impairment is the size of the dose. The number of drinks consumed, the speed with which they are consumed, and the alcohol content of the drinks are the major determinants of the dose. Dose size is moderated by numerous individual characteristics. Heavier and more muscular individuals have more water mass and as a consequence will reach a lower BAC than a smaller, less muscular individual who has consumed the same amount of alcohol. Individuals also differ substantially in the rate at which the liver metabolizes alcohol. For example, there is evidence that older individuals metabolize alcohol more slowly than younger individuals and that chronic drinkers metabolize alcohol more rapidly than less frequent drinkers.

Generally speaking, a 160 lb man will reach a BAC of 0.02% (or 2 g per 100 mm of blood) after one standard-sized drink (roughly one shot (1–1.5 oz) of liquor, one 12 oz beer, or one 5 oz glass of wine). That same man will reach a BAC of 0.05%, 0.07%, 0.09%, and 0.12% after two, three, four, and five drinks, respectively, and will accordingly reach increasingly higher BACs with successive drinks (assuming no time between drinks). A similarly sized woman will, on average, reach a higher BAC after the same number of drinks due to sex-specific differences in body composition.

Though the exact level of impairment at a given BAC varies from person to person, intoxication due to alcohol usually follows several stages associated with different BAC levels. At low BACs (below 0.05%), alcohol can induce enjoyment, happiness, and euphoria characterized by increased sociability and talkativeness. Loss of inhibitions and reduced attention are also characteristic of this level of intoxication. At higher BACs (0.06–0.10%), disinhibition is more apparent, as are impairments in judgment, coordination, concentration, reflexes, depth perception, distance acuity, and peripheral vision. Because these impairments can be dangerous in certain environments, many countries set the BAC at which a driver is considered legally impaired at approximately 0.05% or 0.08%

(and often lower for younger or less experienced drivers). In the range 0.11–0.30% BAC, individuals experience exaggerated emotional states, including anger and sadness; they may also have a higher pain threshold, reduced reaction time, loss of balance, slurred speech, and moderate-to-severe motor impairment. At extremely high BACs (above 0.35%), individuals are likely to suffer from incontinence or impaired respiration, or they may lose consciousness and even die from respiratory arrest. For lower levels of BAC, many of the effects have been documented in controlled laboratory settings, particularly impairments of driving-related skills and tasks, as well as aggression.

Alcohol's pharmacological profile is distinct from that of other commonly consumed drugs. Probably the closest to alcohol in its pharmacological effects is cocaine, which has similarly been shown to increase aggression, reduce self-control, and increase irritability. Amphetamines can also produce an increase in aggression; however, unlike the aggression induced by alcohol, it is sometimes accompanied by a paranoid psychotic state that may independently contribute to violent acts. In contrast, marijuana has generally been found to inhibit (rather than promote) aggressive behavior in humans, mice, and primates. Similarly, opiates have been shown to decrease aggressive behavior and hostility in animals and humans, though the period of opiate withdrawal is usually characterized as increasing risk for aggressive behaviors. Thus, alcohol has a pharmacological profile that is significantly different from that of the most commonly consumed illicit drugs.

The differential pharmacological effects of alcohol and other drugs on human behavior raise a potentially important issue regarding the economics of alcohol regulation. Specifically, it is possible that alcohol use is fundamentally linked to the use of other drugs. If alcohol and other drugs are complements in consumption, then an increase in the price of alcohol (through, e.g., stricter regulations) will reduce not only drinking (through the own-price effect) but also the use of other drugs (through a cross-price effect). In contrast, if alcohol and other drugs are substitutes in consumption, then an increase in the price of alcohol will reduce drinking but will lead to an increase in the use of other drugs. Existing research is mixed on this question, but these relationships are important to consider when designing optimal alcohol control policies because the effects of those policies on the use of other drugs – and the independent effects of other drug use on outcomes – need to be acknowledged.

Economics Perspectives on Alcohol Use and Alcohol Regulation: Distinguishing Factors

Economics may not be the first discipline that comes to mind as relevant for studying alcohol. As such, it is useful to clarify the distinguishing characteristics of the economics way of thinking that are relevant for understanding this important topic. Arguably one of the most important distinctions is that economists put value not only on the costs of alcohol consumption in terms of productivity losses, health impairments, and criminality but also on the benefits of alcohol consumption. That is, a great deal of alcohol consumption is utility increasing, and these benefits of drinking must be taken into account when considering tighter restrictions on alcohol availability. The public health tradition, in contrast, generally calls for stricter alcohol control to reduce alcohol-related harms without consideration for the benefits of drinking that accrue to most moderate drinkers. Economics recognizes that adoption of stricter alcohol control policies for the purposes of harm reduction imposes deadweight loss on moderate, responsible consumers. Higher taxes, for example, may reduce alcohol consumption by people whose drinking causes them to be at risk for adverse health events or to commit crime but may also reduce the consumption by law-abiding drinkers. Because a large share of the population consumes alcohol and does so in a responsible way, the foregone value of alcohol consumption by this group cannot be easily dismissed.

This does not mean that economists oppose any move to tighten alcohol restrictions. But the discipline does provide a unified framework for thinking about the conditions under which government intervention in the form of alcohol control may be justified. Specifically, if drinkers impose costs on other members of society (e.g., an alcohol-involved driver may kill or injure someone, or a drinker may commit a crime against someone), it is said that the marginal social costs of alcohol are greater than the private costs (i.e., there is a negative externality), leading unregulated private markets to result in too much alcohol consumption and resulting in alcohol-related harms. In this case, economics theory justifies correcting this behavior in a variety of ways. Next, a host of alcohol control regulations are described that have been proposed and adopted across many places and that deal with the negative externality problem in very different ways. It is important to remember, however, that because economists value both the benefits of drinking and the harms, the socially optimal level of alcohol consumption and alcohol-related harms will be lower than in a completely unregulated environment but will be strictly positive.

A final distinguishing feature of the economics tradition with respect to research on alcohol use and alcohol control is that the discipline of economics has been a leader in the social and public health sciences in advancing methodologies regarding causal inference. In many cases, including alcohol consumption, researchers are faced with the problem that observed associations between a treatment (here, drinking) and an outcome (e.g., death, illness, productivity, crime, etc.) may be simultaneously determined. That is, factors that affect the treatment may independently affect the outcome. In the case of drinking and adverse health outcomes, for example, one might worry about population heterogeneity in risk attitudes and discount rates (i.e., how much people trade off utility today against utility at a later date). It could be that heavily discounting the future causes people to both consume alcohol and engage in other risky behavior that puts them at risk of an adverse health event. If so, one might observe that people who drink are at an increased risk for adverse health events even if there is no direct causal effect of alcohol. Put differently, those same people might have experienced the adverse health event even in the absence of their drinking; alcohol consumption and adverse health events may both simply reflect their high discount rate. To see the importance of disentangling correlation from causation, note that alcohol

availability can be (and is) regulated by local, state, and federal governments. If the correlations between alcohol use and adverse events are not causal, then tighter alcohol control will not be an effective means to improve population health; if, in contrast, alcohol use does cause adverse events, then stricter alcohol policies can be expected to reduce not only drinking but also subsequent adverse outcomes. The relative importance of distinguishing correlation from causation varies dramatically across disciplines, with economics very much at the end of the spectrum that cares deeply about this distinction. Public health, health services research, and sociology do not place as much of a premium on this component of research; in these traditions, detailed descriptive analyses of associations between alcohol use and individual-level factors are more common.

How do economists deal with the evaluation problem (sometimes referred to as 'omitted variables bias,' 'unobserved heterogeneity bias,' 'endogeneity bias,' 'simultaneity bias,' and others) when treatment assignment is nonrandom? First, note that the ideal solution to nonrandom treatment assignment commonly used in the natural sciences is to randomize treatment and compare outcomes between the treated and untreated; because the treatment assignment was manipulated to be random, the difference in outcomes can be causally attributable to the treatment. In the real world, however, researchers cannot randomize alcohol consumption, and so social scientists have had to take different approaches. One is to try to control for as many of these omitted factors as possible in regression models either directly or through the use of single indices such as propensity scores; these approaches are common in health services and some economics research. In the past few decades, however, economists have pushed for stronger research designs that mimic the experimental variation in the natural sciences. This class of methods, commonly referred to as 'quasi-experimental' approaches, includes difference-in-differences (DID), instrumental variables (IV), and regression discontinuity (RD) approaches, among others. When applied appropriately, each of these designs isolates variation in the treatment that is thought to be 'exogenous to outcomes' or to create variation in treatment that is 'as good as random' for some subpopulation of interest, thus overcoming the omitted variables bias problem. An example with respect to alcohol availability, alcohol consumption, and outcomes is research that has capitalized on labor strikes for workers at government-run liquor stores in Scandinavia (where the government owns a liquor monopoly), which exogenously reduced alcohol availability, alcohol consumption, and subsequent alcohol-related problems. These rigorous standards for identification of treatment effects also distinguish the economics approach to studying alcohol consumption and alcohol control from other disciplinary traditions.

Alcohol Control Policies and Alcohol Consumption

A great deal of economics research on alcohol use has focused on estimating the effects of alcohol control policies on alcohol consumption, both because this type of policy evaluation is independently interesting and because many policy-induced changes in alcohol consumption can be used to identify causal effects of alcohol use on outcomes (e.g., mortality and morbidity). Research on alcohol control policies is particularly appealing to economists because of the fundamental tenet in economics that demand curves slope downward. That is, the price of a commodity and the quantity demanded of that commodity are inversely related. In the context of alcohol consumption, this means that policies and practices that raise the full price of drinking either directly (e.g., through alcohol taxes, which are passed through to consumers in the form of higher alcohol prices) or indirectly (e.g., through other types of availability restrictions) should reduce the quantity of alcohol consumed. Although alcohol taxes are probably the most widely studied alcohol control policies in the economics literature (and have been summarized in multiple recent meta-analyses), many others have also received scholarly attention, including the presence of government liquor monopolies; age-based alcohol availability restrictions (e.g., minimum legal drinking ages (MLDAs)); drunk driving laws (e.g., BAC limits, driver license suspensions, random breath tests, and sanctions/penalties for driving under the influence); spatial restrictions on alcohol availability (e.g., liquor license restrictions); temporal restrictions on alcohol availability (e.g., Sunday alcohol sales bans or bar/pub closing hours); advertising and sponsorship restrictions (including health warnings); other 'circumstance' regulations such as prohibitions on alcohol sales at sporting events; and legal liability for bartenders and bar owners for serving intoxicated persons, among others.

The most common approach taken in this literature to test whether the demand curve for alcohol slopes downward has been to relate drinking rates (using either individual-level survey data on alcohol consumption or aggregate data on alcohol sales) to variation across places in the alcohol control environment at a point in time. This approach is made possible by the fact that places (e.g., localities, states, provinces, countries, etc.) vary substantially in their chosen menu of alcohol control policies. For example, some places have higher alcohol taxes and/or severe penalties for alcohol-involved driving compared to other places. A finding that drinking rates are lower in places where alcohol is more difficult to obtain (i.e., where individuals face higher full prices to drink) is usually taken as evidence that demand curves slope downward, or that drinking is negatively related to price.

One weakness of the type of approach described in the previous paragraph is that the types of designs that rely on variation across places at any one point in time may suffer from the omitted variables biases. For example, if places that are very religious are the ones that are more likely to have high alcohol taxes and strict availability restrictions, then the inverse relationship might be observed between the full price of obtaining alcohol and drinking rates that is due to the religious attitudes of people in that area, not due to the policies and prices themselves. This criticism has in the past decade led economists to incorporate other types of research designs commonly found in other applied microeconomics disciplines (most notably labor, public, and development economics). As such, the more commonly accepted standard for evaluation research on alcohol control policies and alcohol consumption is to compare changes in drinking rates coincident with changes in alcohol control policies (e.g., alcohol

excise tax increases or tightening of availability rules). The advantage of this 'changes on changes' or 'DID' type of specification is that, because areas usually adopt different alcohol control policies at different times, researchers can use the staggered timing of adoption to rule out the possibility that permanent unobserved differences about individual places are driving the observed relationships between alcohol prices (broadly defined) and alcohol consumption. In practice, this amounts to including dummy variables, or fixed effects, for each area in multivariate regression models of drinking that include controls for area-specific alcohol policies.

Results from this and other types of quasi-experimental approaches have been somewhat less conclusive about the role of alcohol taxes in determining alcohol consumption behaviors, in that they have not uniformly returned evidence of significant relationships between alcohol excise tax increases and alcohol consumption decreases, particularly for research on youths and for research focusing on the US. Part of the lack of clarity around the effects of alcohol taxes on consumption is that in the US there have been relatively few large alcohol tax increases in the past three decades; by construction, this makes estimating difference or change-based models more difficult. (The lack of alcohol tax change variation is notably different from the case of tobacco.) Similarly, studies of spatial, temporal, and other 'circumstance'-type regulations of alcohol availability have not produced overwhelming evidence that these policies seriously affect overall alcohol consumption, which is perhaps not surprising because it is not particularly costly to undo the effects of these types of restrictions (e.g., purchasing alcohol on Saturday can undo the effects of a Sunday alcohol sales prohibition). There is, however, ample evidence from these types of stronger designs that age-based alcohol restrictions (such as MLDAs) causally reduce alcohol use. For example, research in the US has shown that state experimentation with lower drinking ages in the 1970s and early 1980s led to higher drinking rates among youths who were newly legal to drink, and similarly state increases in drinking ages back to age 21 (the current MLDA in the US) led to lower drinking rates among youths who were no longer legally allowed to drink. Moreover, research has also shown that alcohol use increases sharply and discretely exactly at a country's MLDA, even when other policies do not change discontinuously at the same threshold. This further bolsters the idea that minimum drinking age policies causally reduce alcohol consumption. Because drinking ages affect the total price of obtaining alcohol through time and convenience costs for youths who are too young to legally consume alcohol, studies of minimum drinking ages have played an important role in confirming that demand curves do, indeed, slope downward for alcohol.

Finally, it can be noted that research exploiting changes in place-specific alcohol control regulations to identify the effects of higher effective prices for obtaining alcohol on drinking rates – with improvements over comparisons of drinking rates across areas at a point in time – are not a panacea. Specifically, these studies must also contend with the fact that alcohol policy changes may themselves likely be the result of unobserved population preferences, because in democratic societies voters elect officials who make or change policy. If sharp changes in attitudes toward alcohol underlie the changes in alcohol control policies, then studies using DID can still be biased. In this situation, other strategies that are less prone to these criticisms, such as RD or IV, can be useful alternatives.

Causal Effects of Alcohol Consumption on Outcomes

The other area where economists have contributed substantially to the literature on alcohol is in estimating causal effects of alcohol consumption on outcomes. Adverse health events such as mortality, crime, and risky sexual behavior are the most widely studied outcomes, and the pharmacological profile of alcohol consumption makes a causal role for alcohol in determining each of these outcomes eminently plausible. Of course, extreme alcohol consumption can directly lead to respiratory failure and death. But there are many other pharmacological mechanisms as well. By reducing reaction time and peripheral vision, alcohol-involved driving can directly increase motor vehicle fatality risk. By altering perceptions of right and wrong and compromising a person's ability to reason through the consequences of one's choices, alcohol consumption can increase risk-taking that could lead to many other types of nonvehicle-related accidents and to the commission of several types of crime. By increasing aggression and exaggerating emotional state, alcohol consumption can increase the likelihood individuals will commit a violent crime. By incapacitating a person, alcohol consumption can increase criminal victimization risk. And the social aspects of drinking can put people in situations that independently increase their risk of an unwanted physical or sexual encounter. All of these channels make it plausible that alcohol use can cause adverse events.

The plurality of research studies in economics examining the effects of alcohol have examined mortality as the outcome of interest. Although mortality is rare, it is very well measured and is an unambiguously negative outcome. Mortality also has the advantage that certain types of deaths are more likely to be alcohol related than others, for example, motor vehicle fatalities are far more likely to be attributable to alcohol than cancer deaths, and studies of the blood alcohol levels of decedents show that very high proportions of deaths from suicide, falls, drowns, burnings, and other 'external' causes are alcohol involved. This means that a relationship between alcohol prices and policies and deaths that are more commonly thought to be alcohol related can provide stronger evidence of a causal role for alcohol use in mortality events. Motor vehicle fatalities are by far the most commonly studied mortality outcome; in the US these data provide the additional advantage that accident characteristics such as time (e.g., nighttime vs. daytime) and day (weekend vs. weekday) can strongly correlate with the likely involvement of alcohol as a contributing factor. Morbidity and nonfatal injury share many of these same benefits (to researchers) as mortality, but availability of comparable large-scale morbidity data spanning multiple places and time periods has been much sparser in the past three decades (with a few exceptions such as occupational and workplace injuries, which are tracked administratively).

Many economics studies report that areas with higher alcohol taxes or stricter alcohol availability regimes have lower motor vehicle fatality rates, though as with the alcohol

consumption evidence, studies in this literature have not uniformly shown that alcohol excise tax increases lead to significant motor vehicle mortality decreases. Other quasi-experimental approaches, however, have strongly demonstrated that higher full alcohol prices reduce mortality. For example, economics research has used DID approaches to demonstrate that higher (lower) drinking ages reduce (increase) motor vehicle fatalities in the age groups newly illegal (legal) to drink that are likely to have involved alcohol. More recently, RD approaches have also shown that mortality rates for motor vehicle deaths and suicides increase discretely at the MLDA, suggesting a causal role for alcohol in these mortality events. Perhaps not surprisingly, drunk driving laws such as state movements to lower legal blood alcohol-content thresholds have also been shown to directly and significantly reduce motor vehicle deaths likely to have involved alcohol.

Of the other adverse outcomes associated with (and possibly caused by) alcohol consumption, crime and risky sexual behavior have received the most attention from economists. Both of these outcomes have the advantage over mortality that they are very common events routinely associated with alcohol. Indeed, vast public health literatures show that individuals who consume alcohol are more likely to commit crime, more likely to have been arrested for a crime, more likely to be victims of crime, more likely to have engaged in sexual activity, more likely to have engaged in sexual activity at an earlier age, more likely to have had unprotected sex, more likely to have had an unplanned pregnancy, and more likely to have had a complicated birth. To what extent are these relationships causal effects of alcohol use?

Several studies have used the money price of alcohol in an IV framework to try to disentangle alcohol's causal role in crime and violence. These studies generally find that individuals in places with low alcohol taxes are more likely to drink, more likely to commit intrahousehold violence, more likely to get into physical fights, and more likely to carry weapons, though concern about omitted variables biases from using cross-sectional variation in alcohol taxes and prices to identify these effects is a serious issue. However, multiple economics studies have used DID methods to examine whether alcohol price increases lead to crime decreases, and these studies have found evidence supporting a causal effect of alcohol availability on certain types of crime – especially violent crime. Studies of drinking ages using the similar approach of relying on state policy changes have also provided evidence that alcohol availability is causally related to crime, and more recent research also using the minimum drinking age in an RD framework has shown that arrests increase discretely at the MLDA – further evidence for a causal effect of alcohol use on the commission of crime.

Economists have also studied alcohol's causal role in sexual activity using quasi-experimental approaches and have found some evidence that alcohol taxes are negatively related to the probability of sexual intercourse and are positively related to the likelihood of using condoms during intercourse. Other economics research has documented a negative relationship between the full price of alcohol and both teen birthrates and rates of sexually transmitted infections such as gonorrhea and syphilis, including in models that rely on changes in alcohol prices and policies for identification of alcohol's effects. Arguably stronger evidence for such a relationship comes from research designs based on drinking ages, as these studies have shown that youths exposed to relatively more lenient drinking ages were more likely to have births than otherwise similar youths who came of age in the same state but just a few years before or after and who were exposed to relatively less lenient drinking ages. Because these youths are likely to be very similar on observed and unobserved dimensions, omitted variables bias concerns are mitigated.

In summary, much of the economics literature addressing the causal effects of alcohol use on adverse outcomes has used a variety of quasi-experimental approaches to try to overcome the potentially severe omitted variables bias concerns. These studies have had mixed success in relying on tax-induced variation in alcohol consumption, in part because large alcohol tax changes have historically been rare (at least in the US); often this has translated into precision challenges for research designs that rely on alcohol tax variation. Studies employing alternative alcohol control policies such as drinking ages have produced stronger evidence in this respect, both because there are many policy changes to work with and because multiple age-based designs can be used (e.g., DID and RD approaches). Of course, drinking-age-based designs do not necessarily tell much about the effects of alcohol at higher points in the age distribution, so more research is needed on these important questions.

Finally, it is important to note that alcohol may also have causal effects that are positive, not negative. For example, drinkers earn more than abstainers, and part of this may reflect a causal effect of drinking (plausibly related to social interactions in certain types of occupations). Similarly, very large observational studies in public health have shown that moderate alcohol consumption is associated with reduced risk of heart disease mortality, giving rise to the oft-cited benefits of a glass of wine per day. This too may reflect a causal beneficial effect of alcohol on health (biological mechanisms include the possibility that alcohol reduces plaque deposits in the arteries and reduces the risk of blood clots). Economics research on these plausible benefits of drinking is much less complete than on the costs of drinking, in part perhaps because the types of designs that can provide relatively compelling evidence on causality are better suited to well-measured acute events such as deaths and arrests (as opposed to longevity or earnings, which are more likely the product of a series of important decisions and outcomes). Understanding whether and to what extent alcohol has causal effects on beneficial outcomes is an important area for research.

Conclusion

Economists have contributed greatly to the study of alcohol availability, alcohol consumption, and alcohol regulation. Key to the economics framework is a complete accounting of both the costs and the benefits of drinking, which has important implications for government intervention to correct negative externalities associated with alcohol consumption. Economists have also distinguished themselves among the social and public health sciences by advancing methodological rigor with respect to causal inference. Arguably the strongest consistent finding in

the broad economics literature on alcohol is that demand curves for alcohol slope downward: increases in the price of alcohol (broadly defined to include increases in both monetary prices and other nonmonetary costs of drinking) are negatively associated with the probability and frequency of drinking and with the quantity of alcohol consumed. Research has also credibly demonstrated that alcohol availability and alcohol consumption are causally related to increased risk of premature death, and there is growing evidence that drinking also causes individuals to be at increased risk for nonfatal injury, crime, and risky sexual behavior. More work is needed to understand whether and to what extent alcohol may have causal effects of improving (rather than harming) some health and social outcomes, as well as to understand the extent and nature of heterogeneity in the effects of alcohol control policies on drinking and health outcomes.

See also: Illegal Drug Use, Health Effects of. Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap. Peer Effects in Health Behaviors. Smoking, Economics of

Further Reading

Becker, G. S., Grossman, M. and Murphy, K. M. (1991). Rational addiction and the effect of price on consumption. *American Economic Review* **81**, 237–241.

Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy* **96**, 675–700.

Bonnie, R. J. and O'Connell, M. E. (eds.) (2004). *Reducing underage drinking: A collective responsibility.* Washington, DC: The National Academies Press.

Carpenter, C. and Dobkin, C. (2011a). The minimum legal drinking age and public health. *Journal of Economic Perspectives* **25**, 133–156.

Carpenter, C. and Dobkin, C. (2011b). Alcohol regulation and crime. In Cook, P., Ludwig, J. and McCrary, J. (eds.) *Controlling crime: Strategies and tradeoffs*, pp. 291–329. Chicago: University of Chicago Press.

Chaloupka, F. J., Grossman, M. and Saffer, H. (2002). The effects of price on alcohol consumption and alcohol-related problems. NIAAA publication. Available at: http://pubs.niaaa.nih.gov/publications/arh26-1/22-34.htm (accessed 03.06.13).

Cook, P. J. (2010). *Paying the tab: The costs and benefits of alcohol control.* Princeton: Princeton University Press.

Cook, P. J. and Moore, M. J. (2000). Alcohol. In Cuyler, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1b, pp. 1629–1673. USA: Elsevier Science and Technology and North Holland.

Cook, P. J. and Moore, M. J. (2002). The economics of alcohol abuse and alcohol-control policies. *Health Affairs* **21**, 120–133.

Dee, T. S. (1999). State alcohol policies, teen drinking and traffic fatalities. *Journal of Public Economics* **72**, 289–315.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy* **80**, 223–255.

Grossman, M. (2005). Individual behaviors and substance use: The role of price. In Lindgren, B. and Grossman, M. (eds.) *Substance use: Individual behavior, social interaction, markets and politics*, pp. 15–39. Amsterdam: JAI, an Imprint of Elsevier Ltd.

Manning, W. G., Keller, E. B., Newhouse, J. P., Sloss, E. M. and Wasserman, J. (1989). The taxes of sin: Do smokers and drinkers pay their way? *Journal of the American Medical Association* **261**, 1604–1609.

Wagenaar, A. C., Salois, M. J. and Komro, K. A. (2009). Effects of beverage alcohol price and tax levels on drinking: A meta-analysis of 1003 estimates from 112 studies. *Addiction* **104**, 179–190.

World Health Organization (2011). *Global status report on alcohol and health.* Geneva: World Health Organization Press.

Relevant Website

http://www.niaaa.nih.gov/Resources/DatabaseResources/QuickFacts/Pages/default.aspx

National Institute of Alcohol Abuse and Alcoholism.

Ambulance and Patient Transport Services

Elizabeth T Wilde, Columbia University, New York, NY, USA

Introduction

Ambulance and Patient Transport Services include Emergency Medical Services (EMS) and private ambulance services, which supply emergency prehospital care, including basic medical support and roadside transport to hospitals for patients experiencing medical emergencies. In recent years, a number of economists have written thoughtful and careful papers on EMS; this article will summarize their work and the work of others who write on EMS topics of interest to economists. Sections Taxonomy of Ambulance and Patient Transport Services and US Emergency Medical Services contain an introduction to EMS. Section Private Provision of Emergency Medical Services, describes the work analyzing the decision to outsource EMS. Section Factors Affecting Quality of Care, summarizes the existing evidence on supply side factors affecting the quality of care. Section Quality of Care and Health Outcomes describes research on the relationship between quality of care and health outcomes. Section Demand for Emergency Medical Services explores factors predicting demand. Section Cost-Effectiveness/Cost–Benefit Analyses includes a description of cost-effectiveness and cost–benefit analyses. Section Conclusion concludes.

Taxonomy of Ambulance and Patient Transport Services

EMS are rival and excludable – only one patient can use an ambulance at a time and patients can be barred from service. In practice, however, access to EMS is frequently available to all, not only to those with the ability to pay. This makes EMS an impure local public good. EMS could also be considered an option good; patients frequently pay through taxes for the option of having EMS available when needed. As EMS systems are built to address urgent, unpredictable needs, there may be excess capacity most of the time.

EMS systems vary tremendously throughout the world. In Japan, EMS are provided by Emergency Life Support Technicians who have limited roles – they can provide cardiopulmonary resuscitation (CPR), defibrillate patients, and insert an airway, but are prohibited from distributing drugs. In Germany, EMS is regulated and organized at the 'Lander' or state level; the German population has the right and guarantee of prehospital emergency medical care either through a physician available through 24 h house call or via EMS. A person calling for EMS in Germany would reach a central dispatcher and then, most likely, would be served by a two-tiered system including a physician-staffed ALS. In 1998, EMS in Russia was two-tiered and staffed by physicians, with nurses dispatching ambulances from a central location and treatment initiated in the field; many patients are treated by physicians and then not transported to the hospital, unlike in the US system, for example, where transport is required for compensation. In general, European prehospital care is more likely to include care from a physician or a nurse in addition to a paramedic compared with American ambulances that do not have personnel with more than paramedic training; for a description, by country, of prehospital care arrangements (organization of EMS system, ambulance staffing, and helicopter availability), see Lethbridge (2009).

In low- and middle-income countries, prehospital care is frequently unavailable; if it exists, it is concentrated in urban areas, likely to be privately provided (and only available to those with the ability to pay), of uneven quality and largely unregulated, even though trauma, particularly as a result of car accidents, represents an increasingly significant and growing source of disability and mortality in developing countries. In Islamabad, Pakistan, police officers as well as physicians staff ambulances provided through a public–private partnership; members of the community, including Non Government Organizations subsidize physician salaries, equipment, and ongoing operational costs other than police salaries. In Turkey in the late 1990s, no personnel or equipment standards existed for prehospital care; in a typical city, Izmir, ambulances were staffed with a medical doctor with limited training and a driver without medical expertise, and it was unusual when the ambulance arrived before the patient had been transported by other means to the hospital. In 1997, Vietnam had no organized prehospital system; ambulances may be used for transport, but most often prehospital care relied on bystanders' transporting patients. In 1998, consistent with many other developing countries, there was no centralized prehospital care system in Thailand; approximately 30 pick up trucks staffed by volunteers picked up residents around Bangkok. Drivers have limited first-aid training. A water rescue boat must travel first from the hospital to the river, decreasing its usefulness significantly. Despite a large and increasing number of traffic accidents, prehospital care in India is largely nonexistent; with no centralized regulating body and the ambulance services only provided in only a few large cities where they are largely privately funded, most Indians lack access to trauma care of any kind. What is provided is of uneven quality; few programs exist to train paramedics and Emergency Medical Technicians (EMTs), and no certification or accreditation exists for professionals or programs. These characteristics define prehospital care throughout Southeast Asia (Bangladesh, India, Nepal, Pakistan, Bhutan, Maldives, and Sri Lanka); disproportionately concentrated in urban areas, serving those of higher socioeconomic status, frequently privately provided, without regulation or certification requirements, and limited in capabilities.

US Emergency Medical Services

In a typical EMS call in the US, a patient calls 911. A dispatcher at a local call center asks the patient a series of questions, evaluating the situation and eliminating false calls. The

dispatcher may also give the patient medical instructions over the phone while simultaneously activating the local EMS response. In urban areas, first responders typically arrive first at the scene. A first responder captures vital signs, determines the patient's medical history, and provides CPR. Meanwhile, the EMS response team composed of basic or intermediate EMTs or paramedics (advanced EMTs) travels to the scene by helicopter or by ground ambulance. Although the particular responsibilities of each type of personnel differ by state, EMTs supply more advanced care to patients than first-aid trained first responders. After arriving at the scene, assessing the situation, and providing initial care, the EMT or paramedic loads the patient into an ambulance or a helicopter and takes the patient to a hospital. In some cases, a medical director instructs and authorizes treatments en route. After transferring the patient to the care of physicians within the hospital, the EMS personnel collect billing information and fill out a call log with demographic and incident characteristics. In rural areas, the ambulance would likely be staffed by volunteers capable of delivering Basic Life Support services.

Most large cities in the US publicly provide EMS; in nearly half of all communities, EMS are organized and delivered through the fire department. Although first responders are almost always employed by a local government, either public or private ambulance or helicopter services may transfer patients. Many communities outsource emergency transport to for-profit ambulance agencies (more than 3000 in the US) or to hospital-based companies (approximately 7% of systems). In a hospital-based EMS system, the ambulance might park at the hospital in between calls and might be encouraged to bring patients to the affiliated hospital. With a private agency (hospital-based or other), the provider would likely own the infrastructure including the ambulance.

Revenues collected from private and public insurance for patient transports provide the majority of funding for EMS, potentially encouraging agencies to transport patients for whom the trip to the hospital is unnecessary. State and local taxes frequently supplement fees collected through insurance, along with grants from the state and the federal government. A variety of mechanisms, including government grants, fundraising, and donations, fund volunteer ambulance services.

Rather than being transported by ground ambulance, some patients may travel by medical helicopters. As of 2006, more than 650 medical helicopters operated within the US, run by private for-profit providers, hospitals, government agencies, or the military. More expensive to operate than traditional ambulances, helicopters may be no faster than ground ambulances, except in rural areas far from hospitals or in places where a ground ambulance cannot travel. Many patients transported by helicopter could have safely been transported by ground ambulance at considerably less expense without any survival loss. Using a helicopter may also limit the set of hospitals that a patient can be transported to.

Private Provision of Emergency Medical Services

When do some communities choose to outsource patient transport? In a 2009 paper, Holian hypothesizes that a vote maximizing politician will outsource patient transport when it will increase her votes. In his model of private provision, as the proportion of the elderly, who consume a disproportionate amount of EMS rises, service levels change. Empirical work suggests an inverted U-shaped relationship between the proportion of the voting population which is elderly and the proportion of privately provided ambulance services.

Communities might outsource their EMS for many reasons. In 2009, David and Chiang found that although fire departments may have lower EMS transportation costs because they can take advantage of the existing firehouse infrastructure to get closer to patients, it may be cheaper for private agencies, which can spread costs across multiple communities – to introduce technology which improves the quality of care (such as Geographic Information System). Arguably, then, the decision to privatize depends on several factors including the distance to other cities, the population, and the number of hospitals in the city (all but the former negatively associated with private provision). Among the ten largest and ten smallest cities in the US, larger cities, with older, less healthy populations, a higher chance of disasters, more crime, less geographically dispersed fire stations and trauma centers, and strong unions, tend to be less likely to contract with private providers.

A related question not yet evaluated empirically is whether public or private agencies are better providers. Some hypothesize that private ambulances may provide EMS care more efficiently than public ambulances, because private paramedics frequently earn lower salaries than paramedics employed directly by state and local governments, even as they appear to have more sophisticated equipment and greater flexibility.

Factors Affecting Quality of Care

Unfortunately, there are no nationally or internationally agreed-upon measures of EMS quality. However, response time, defined as the difference between the time of the initial call and the time of arrival at the scene, is one commonly used metric. Other metrics commonly used include total call time. Such metrics have not been systematically used by communities or states in the US to assess the quality of their EMS because a large proportion of states do not systematically collect response time data.

Many factors appear to be correlated with response times. In one southern state, Mississippi, whites appear to have higher response times than blacks, but these differences are eliminated after controlling for a county-level measure of population density. Others have found that distance, evening rush hour, patient being of Native American or Pacific Islander race, and gender predict longer total response times and that these factors plus bypass, neighborhood population density and percentage of white population are associated with delays of more than 15 min. Other factors including population density, the age of the housing stock, per-capita income, and first responders per square mile seem to be negatively correlated with mean response times, with area being positively correlated with mean response times.

It appears that incentives also affect response times – or at least the reported response times. One program in England profiled by Bevan and Hamblin (2009) publicly rewarded

agencies meeting response time targets with gold stars. After the program was introduced, the proportion of agencies meeting performance targets increased, but the gains were illusory – response times were systematically shaved and calls recategorized as less severe to satisfy requirements.

Worker fatigue, experience, human capital depreciation, and turnover also affect response times. In a 2009 article, David and Brachet used the detailed call level data from Mississippi to measure the relationship between experience and time out of hospital or at the scene. They construct person-specific and firm-specific measures of experience, and control for individual fixed effects and a lengthy set of covariates. A one standard deviation increase in the number of trauma runs conducted by an individual in a given quarter is associated with a reduction of 35 s in out-of-hospital time and 10 s on scene. Brachet *et al.* (2010) compare the performance of paramedics working late at night in 24 h shifts with those same paramedics working late at night on 12 h shifts. They observed that paramedics on 24 h shifts have significantly longer response times and take longer to transport patients to the hospital and perform fewer procedures. David and Brachet's (2011) article uses incident level data to measure the impact of human capital depreciation and turnover on time out of hospital. Turnover among EMS personnel is a significant problem for all EMS agencies, both paid staff and volunteers; one estimate puts the annual turnover among EMS personnel as high as 10%, with a median cost to agencies of over US$70 000. Partitioning experience into the human capital of those who work at the firm, those who have left the firm, and those who are joining the firm; David and Brachet derived an expression for firm-level experience and construct a measure of the relative contribution of turnover and human capital depreciation to organizational forgetting. Their reduced form estimates of organizational forgetting suggest that a quarter of the stock of experience existing at the beginning of the year survives to the end. When experience is separated into human capital accrued by individuals in the firm and those who have left the firm, they find the turnover to be a larger source of organizational forgetting (twice as large) than human capital depreciation.

Quality of Care and Health Outcomes

How do factors which affect response time affect health? Although there are many studies that look at factors that affect the quality of EMS care, few evaluate the relationship between quality of care and health outcomes largely because of the challenges in linking prehospital records to mortality and hospital records and in finding a credible nonexperimental identification strategies in a context where experiments may not be feasible.

Athey and Stern's (2002) work uses a differences-in-differences approach to determine the impact of the introduction of the new 911 technology on health outcomes. They model health as a function of response time and initial incident severity; they find that the introduction of Enhanced 911 in Pennsylvania improves the intermediate health measures for patients suffering from cardiac emergencies, as well as improving mortality measured 6 and 48 h after the initial incident. Enhanced 911 also reduces hospital costs for cardiac emergency patients. Wilde takes a different approach in her 2008 paper; she uses distance to the closest EMS agency as an instrument for EMS response time to account for the potential endogeneity of response time to patient severity. She finds that response time matters for mortality, but not health care utilization.

Shen and Hsia investigated the impact of bypass or diversion by EMS providers on mortality after acute myocardial infarction in a 2011 JAMA paper – an event which is arguably unrelated to the characteristics of the patient. Diversion may affect outcomes by affecting EMS response times (when the nearest hospital is on diversion, patients must be transported to hospitals that are farther away); it may mean that patients are transported to poorer quality hospitals or hospitals less capable of providing adequate care; it may also be an indicator for the quality of care for patients within the hospital experiencing the diversion (more crowded hospitals may provide worse care). Patients whose closest emergency department is on diversion for more than 12 h on the day of the incident experience higher mortality 30 days, 90 days, 9 months, and 1 year after the initial incident.

An example of work that explores a key policy question in EMS without a natural experiment or randomized controlled trial is that of a 2008 work by Concannon *et al.* who conducted a simulation of different EMS treatment choices for patients with acute ST-segment elevation myocardial infarctions. Patients can either be transported to the closest available hospital, transported only to hospitals with the capability of providing primary percutaneous coronary intervention (PCI) and treated with PCI or thrombolytic therapy, or be evaluated by EMS or by personnel at the local thrombolytic therapy-only hospital and then transported for PCI. Concannon *et al.* observed that selecting high-benefit patients for transport to PCI-capable hospitals reduces mortality without major shifts in hospital volumes.

Demand for Emergency Medical Services

What affects the use of EMS? There appears to be distinct EMS usage patterns by day (more calls between 10.00 a.m. and 8.00 p.m.) and the day of week (more calls on Friday and Saturday). Age and race/ethnicity also predict usage: people over the age of 85 years call more than 3 times the rate of those between 45 and 64 years of age and are transported at more than 4 times the rate of patients between 45 and 64 years of age. African Americans also call at a much higher rate than non-Hispanic whites.

In an intriguing analysis, Ringburg *et al.* conducted a discrete choice experiment in the Netherlands and found that households were willing to pay much higher amounts than would actually be necessary to provide 24 h helicopter emergency medical service as described in a 2009 paper. It appears that even if helicopter services are not cost-effective, households are willing to pay for them.

Many researchers in the field of operations research and applied mathematics have tackled questions regarding the optimal design of EMS systems, including identifying the optimal ambulance and helicopter station location and the

optimal response time threshold for performance measurement, in addition to building models to forecast demand. That research is beyond the scope of this work.

Cost-Effectiveness/Cost–Benefit Analyses

Most existing cost analyses compare the costs and benefits of particular intervention or mode of care. For example, in their 2002 paper Athey and Stern calculate the costs and benefits from introducing Enhanced 911, a service that helps dispatchers to identify caller locations. They find that improvements in outcomes for cardiac issues cover 85% of the costs of implementing Enhanced 911, making the policy almost certain to be beneficial. Wilde conducts a cost–benefit analysis of a reduction in response times caused by eliminating mutual aid – a policy whereby communities share resources to cover excess demand – and finds that the per life year cost of a 9.5 s reduction in response times would be considerably less than US$50 000.

Evidence on the cost-effectiveness of air transport is mixed. One study that determined the costs of operating a local air ambulance service, supplemented with hospital costs for trauma survivors, estimated the cost of air transport per life year saved as US$2454. Another study collected microlevel costs, surveyed patients two years after their initial trauma incident, and estimated the incremental cost per quality-adjusted life-year (QALY) of helicopter use at more than 28 000 Euros. Several other studies looked retrospectively at patient records and concluded that there were few benefits for patients from air transport, and considerable costs to the health care system. Unfortunately, many of these studies fail to identify the perspective (societal or other), the year the costs were gathered in, fail to include comprehensive costs, and are inconsistent in their assessment of effectiveness making it difficult to draw concrete conclusions (QALY or mortality).

Conclusion

In recent years, there has been an increase in the literature written by or for economists on EMS. Nevertheless, many key clinical and policy questions remain unanswered, providing scope for further research. Economists have much to offer in the field of EMS: by asking different types of questions (i.e., on private vs. public provision, or cost-effectiveness) and using different techniques. Given the growing recognition of EMS as an essential part of emergency care, such research should only increase in the coming years.

See also: Health Care Demand, Empirical Determinants of. Healthcare Safety Net in the US. Waiting Times

References

Athey, S. and Stern, S. (2002). The impact of information technology on emergency health care outcomes. *RAND Journal of Economics* **33**(3), 399–432.

Bevan, G. and Hamblin, R. (2009). Hitting and missing targets by ambulance services for emergency calls: Effects of different systems of performance measurement within the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(1), 161–190.

Brachet, T., David, G. and Duseja, R. (2010). The effect of shift structure on performance: The role of fatigue for paramedics. *NBER Working Paper 16418*. Available at: http://www.nber.org/papers/w16418 (accessed 11.06.13).

David, G. and Brachet, T. (2011). On the determinants of organizational forgetting. *American Economic Journal: Microeconomics* **3**(3), 100–123.

Lethbridge, J. (2009). Privatisation of ambulance, emergency and firefighting services in Europe – A growing threat? pp. 1–21. *Report Commissioned by European Federation of Public Service Unions*. Available at: http://www.psiru.org./ (accessed 05.06.10).

Further Reading

Concannon, T. W., Griffith, J. L., Kent, D. M., et al. (2009). Elapsed time in emergency medical services for patients with cardiac complaints are some patients at greater risk for delay? *Circulation: Cardiovascular Quality and Outcomes* **2**(1), 9–15.

Concannon, T. W., Kent, D. M., Normand, S. L., et al. (2008). A geospatial analysis of emergency transport and inter-hospital transfer in ST-segment elevation myocardial infarction. *American Journal of Cardiology* **101**(1), 69–74.

David, G. and Brachet, T. (2009). Retention, learning by doing, and performance in emergency medical services. *Health Services Research* **44**(3), 902–925.

David, G. and Chiang, A. J. (2009). The determinants of public versus private provision of emergency medical services. *International Journal of Industrial Organization* **27**(2), 312–319.

David, G. and Harrington, S. (2010). Population density and racial differences in the performance of Emergency Medical Services. *Journal of Health Economics* **29**(4), 603–615.

Holian, M. J. (2009). Outsourcing in US cities, ambulances and elderly voters. *Public Choice* **141**(3–4), 421–445.

Institute of Medicine (US). Committee on the Future of Emergency Care in the United States Health System (2007). *Emergency medical services at the crossroads*. Washington, DC: National Academies Press.

McConnel, C. E. and Wilson, R. W. (1998). The demand for prehospital emergency services in an aging society. *Social Science & Medicine* **46**(8), 1027–1031.

Ringburg, A. N., Buljac, M., Stolk, E. A., et al. (2009). Willingness to pay for lives saved by helicopter emergency medical services. *Prehospital Emergency Care* **13**(1), 37–43.

Shen, Y.-C. and Hsia, R. Y. (2011). Association between ambulance diversion and survival among patients with acute myocardial infarction. *Journal of the American Medical Association* **305**(23), 2440–2447.

Wilde, E. (2008). Do response times matter? The impact of EMS response times on health outcomes. *Princeton University Industrial Relations Section Working Paper 527*, pp. 1–78. Available at: http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01b2660cw26d/1/527.pdf (accessed 30.03.13).

Analysing Heterogeneity to Support Decision Making

MA Espinoza, Pontificia Universidad Católica de Chile, Santiago, Chile, and Institute of Public Health of Chile, Santiago, Chile
MJ Sculpher and A Manca, University of York, York, UK
A Basu, University of Washington, Seattle, WA, USA

Glossary

Complete information Refers to the knowledge of the set of covariates that explain differences in outcomes between all individuals.
Essential heterogeneity It corresponds to the unobserved heterogeneity when the selection of treatment depends on these unobserved characteristics.
Ex-ante choices Decisions that a data analyst expects the patients to make based on some of the observed patient characteristics but without access to other relevant information.
Ex-post choices Decision resulting from the interaction of the patient with health professionals, relatives, and other sources of information that are relevant for the treatment selection, but were unobserved to the data analyst trying to predict these choices.
Expected value of individualized care It represents the expected cost of omitting information about individuals when making decisions based on the average estimates.
Nonessential heterogeneity It corresponds to the unobserved heterogeneity when the selection of treatment does not depend on these unobserved characteristics.
Observed heterogeneity Proportion of the variability that can be explained by a set of observed (known) characteristics.
Perfect information Refers to the knowledge of the true mean effect of a particular covariate on the health outcome. It implies 100% precision and/or 100% accuracy, and therefore, no remaining uncertainty.
Preferences In the context of cost-effectiveness analysis (CEA), preferences refer to the rational element of judgement that guides individuals in their task of ranking health states in a particular order to reveal the relative value of such states.
Stratification Process whereby individuals are categorized in different subgroups.
Subgroups Subset of patients whose membership is defined by one or more individual characteristics.
Variability Differences in outcomes between individuals, which can be explained by observed and unobserved characteristics.

Introduction

The flow of new medical technologies is a response to several factors including an ageing population, changes in environmental conditions creating new epidemiological profiles and scientific development. This impacts on health care systems which, to satisfy increased demand for medical technologies, are faced with the need to increase expenditure on healthcare or to disinvest in other services to release resources. Regardless of the type of healthcare system, the problem of deciding which new technologies to fund is unavoidable. As policy makers are increasingly held accountable for these decisions, many are adopting explicit and evidence-based approaches to the allocation of limited resources. This needs, at the very least, information about which interventions work and the value of such technologies.

In a growing proportion of jurisdictions 'value' is defined in terms of cost-effectiveness, where the incremental cost of a new technology per additional health outcome relative to alternative interventions for a given patient group is assessed. This incremental cost-effectiveness ratio is then compared with a maximum (or threshold) value of a unit of health gain which is based either on an estimate of the health forgone as a result of displacing existing services to fund the new technology, administrative rule of thumb or an estimate of society's willingness to forgo consumption in exchange for health improvement. Both effectiveness and cost-effectiveness are usually considered as average estimates relating to a target population. This approach has been largely justified by the fact that it is impossible to observe the effect of alternative treatments in the same individual at the same time, a problem known as the 'fundamental problem of causal inference'. Average treatment effects derived from randomized controlled clinical trials are unbiased estimates when the groups being compared have, on average, similar characteristics, so that the differences in the outcomes are attributable to the treatment received by patients in each group. This causal statement is possible because randomization is expected to balance observed or unobserved confounding factors.

Although the focus on average treatment effects is widespread, this is essentially pragmatic given the challenges in estimating individual treatment effects. The promise of genetic testing is that patient management can more appropriately be tailored to the characteristics of the individual – a technological approach to understanding between-patient heterogeneity in treatment effects. However, in jurisdictions using formal cost-effectiveness analysis to inform resource allocation decisions, as well as those that are unwilling or unable to consider costs explicitly (e.g., comparative effectiveness research in the US), there is a recognized need to understand heterogeneity using existing data on predictors of patients' outcomes following alternative interventions. The focus on the average patient leads to dichotomous decisions – accept or reject a given intervention for all patients in a given population. In contrast, understanding of heterogeneity in costs, effects and cost-effectiveness between patients within the

population facilitates decisions which may guide the use of the intervention toward those patients in whom it is (cost-) effective. This targeted, rather than general, funding of interventions frees-up resources for more (cost-) effective alternatives, leading to an improvement in the overall population health from a given budget allocation. In principal, a full understanding of heterogeneity allows decisions to reflect the characteristics of the individual patient, so the gains from reflecting heterogeneity are maximized.

Interest in heterogeneity for decision-making takes various forms. From a biomedical perspective, reflecting heterogeneity in decisions has been promoted as a means of achieving personalized medicine, which requires the identification of measurable parameters (e.g., based on molecular biomarkers) that allow doctors to prescribe treatments according to specific individual characteristics. Even without such testing, many clinical specialties use existing clinical individual level information to maximize a patient's absolute benefit from treatment compared to its potential harms. An example of this is the use of easily accessible prognostic models for decisions about the choice of adjuvant chemotherapy in breast cancer. Those healthcare systems that use cost-effectiveness analysis to inform decisions increasingly take a step further in seeking to identify the groups of patients for whom absolute health benefit gains justify the relevant cost. Furthermore, despite being financed through taxation, social insurance or private insurance, many collectively funded jurisdictions have recognized the role for individual patient choice in healthcare decisions. There are several reasons for such a policy, and one of these is the potential role for patient choice as a vehicle for characterising unobserved heterogeneity in the costs and benefits of medical interventions.

This article reviews the key elements of the discussion about how heterogeneity should be examined, exploited and analysed for the purposes of decision-making about healthcare interventions. In terms of the methods for economic analysis, it focuses on the role of understanding heterogeneity as a source of value to achieve greater health. The remaining of the article is in four sections. The first section seeks to review standard approaches to the assessment of heterogeneity. The next explores methods developed to represent the value of considering heterogeneity in healthcare decision-making. The third describes the role of patients' choices and preferences in understanding heterogeneity. Finally, the authors conclude by summarizing the key messages of the article highlighting the opportunities for further research.

Standard Approaches to Assess Heterogeneity in Evaluation of Healthcare Technologies

The term 'variability' is used to express the differences in outcomes between individuals. They can be explained by both observed and unobserved characteristics. 'Heterogeneity' has been defined as the proportion of the variability that can be explained by a set of observed (known) characteristics at the time of analysis. In general terms, the set of characteristics that explain the total variability can be further divided in the knowable and the unknowable. In practice, only a portion of the knowable factors can be identified and observed, mainly because of the lack of data and limits on the conduct of further research (e.g., funding and human resources). These knowable only in principle characteristics go with the other unknowable characteristics in the general category of unobserved characteristics. In this article the authors consider unobserved variability or unobserved heterogeneity synonymous. This unobserved part is also referred as stochastic uncertainty or first-order uncertainty.

'Complete information' refers to knowledge of the set of covariates that are able to explain differences in outcomes between all individuals in the population (total variability or total heterogeneity). This is a theoretical concept that is reached when all the covariates needed to explain differences between individuals are revealed. 'Perfect information' refers to the knowledge of the true mean effect of a particular covariate (and its correlation with others) on the health outcome. Likewise, perfect information also refers to the knowledge of the true value of a particular covariate in one individual (e.g., the presence of a genetic characteristic with 100% accuracy). From a decision-making point of view, the main challenge is to take into account as much information about individual-level characteristics as possible. The aim for health researchers is, therefore, to achieve a full characterization of total heterogeneity, i.e., not only to convert the knowable characteristics into observed measurable variables, but also to make some prediction of the expected individual outcomes considering unobserved heterogeneity.

The literature in different areas provides alternative nomenclatures in the study of heterogeneity. For example, epidemiology and biostatistics emphasize the importance of distinguishing between moderators, mediators or nonspecific predictors of treatment outcomes. Variables considered as moderators inform for whom and under which conditions the treatment works. Mediators, in contrast, indicate potential mechanisms that explain the causal effect. Nonspecific predictors are variables that show an effect on the outcome without interacting with the treatment. These distinctions are relevant in understanding the underlying causal model of the health problem. In the context of the evaluation problem in econometrics, unobserved heterogeneity has been termed 'nonessential heterogeneity' when the selection of treatment does not depend on these unobserved characteristics. When treatment selection depends on the unobserved expected gains, this is called 'essential heterogeneity'. In the context of epidemiology and biostatistics, essential heterogeneity indicates that there are knowable moderators of treatment effect that are unobserved in the data. More generally, terms such as observable or measurable heterogeneity are broadly used across the sciences. Figure 1 synthesizes these terms, making a parallel correspondence between them. For example, observable heterogeneity includes, on one side, mediators, moderators and nonspecific predictors. However, it includes known and knowable heterogeneity. Unobserved heterogeneity, also called first order uncertainty or stochastic uncertainty, includes part of the observable heterogeneity (the part that has yet to be revealed) and the unobservable (or unknowable) heterogeneity.

In clinical epidemiology and economic evaluation, exploration of heterogeneity has classically been driven by subgroup analysis. Usually, the dimensions explored correspond

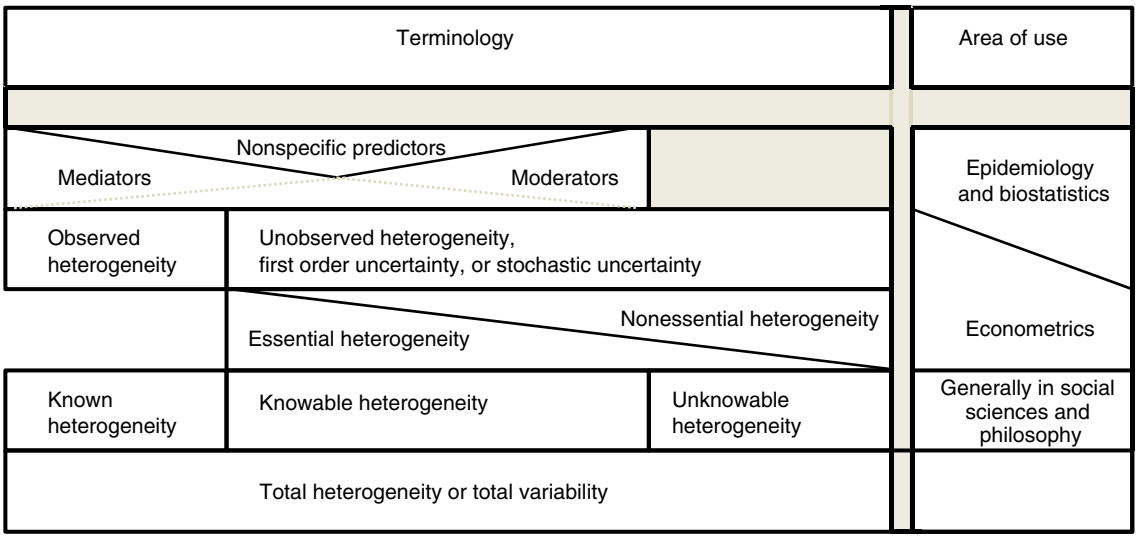


Figure 1 Terminology in the study of heterogeneity. Relationship between different terms and the field where it is used.

to baseline (or underlying) risk and treatment effect heterogeneity. Heterogeneity in baseline risk refers to the set of characteristics that predict a particular a priori probability of presenting the health outcome under standard care or without intervention (natural history). This probability may influence the effect of a new intervention relative to standard care, where the relative treatment effect might be expressed as, for example, a relative risk, odds ratio, or hazard ratio. However, even in the case where the relative treatment effect is the same across individuals, the absolute value of the health outcome can vary across patients if they are expected to have different baseline risk profiles.

Treatment effect heterogeneity, however, exists when a set of patient characteristics predict different relative treatment effects among a population of patients. In statistical terms, this corresponds to the interaction between the treatment effect and the covariate that defines the individual's membership of a particular subgroup. Treatment effect heterogeneity can be categorized as a quantitative interaction (differences between subgroups are in the same direction but they vary in terms of their magnitude), effect concentration (the treatment effect is only seen in one subgroup) and qualitative interaction (the treatment effect varies not only in magnitude but also in direction between subgroups). It is important to stress that both baseline risk and relative treatment effect heterogeneity are defined on the basis of one or more observed characteristics at baseline, assessed on the basis of health outcome(s).

Dealing with heterogeneity in economic evaluation may also relate to costs and preferences. Heterogeneity in costs typically takes the form of a set of patient characteristics predicting differences in the use of healthcare resources. For example, age might be expected to explain a large proportion of the variation in length of hospital stay for common procedures such as hip and knee replacement and heart failure. Heterogeneity in preferences is considered in detail below.

So far the discussion has focussed on heterogeneity at the level of the individual patient. Geographical variation has also been a matter for attention, particularly in cost-effectiveness analysis. This has been explored mostly in the context of countries, although this type of heterogeneity could also be important between localities or jurisdictions within a country, with specific characteristics that affect, for example, the incidence or prevalence of a particular condition. These differences can be explained by several elements of the health system, clinicians, patients or wider socioeconomic factors. For example, the relative prices of resources may vary across jurisdictions as well as the opportunity cost imposed on health outcome through additional costs falling on the system. Similarly, it is known that teaching and specialized hospitals incur higher expenditure than general hospitals, with marked differences within the same jurisdiction. Further, better trained health professionals might generate better clinical results and incur fewer costs as a result of more efficient care (e.g., quicker diagnostics and lower complication rates in surgical procedures).

Despite the growing interest in considering heterogeneity as part of decision-making in healthcare, researchers face some constraints in using these methods due to the orthodox adherence to classical methods of statistical inference. The first of these follows from the fact that most clinical trials are designed to find statistically significant average treatment effects and their sample sizes are determined accordingly, any attempt to make inference on subsets of the sample faces the problem of loss of power (i.e., increase in type-2 error). It can be shown, however, that using prespecified (baseline) covariates in a regression framework increases statistical power, something that can be explained by the magnitude of the prognostic effect of the covariate on the outcome. A second concern relates to the fact that, when additional testing is performed on the same data, there is a higher probability of finding statistically significant differences between groups explained by chance, a problem known as multiplicity (i.e., leading to greater false positives or an increase in type-1 error).

A third problem concerns the requirement of an interaction test to prove treatment effect heterogeneity in clinical studies. If heterogeneity in a treatment effect is shown to be

statistically significant, authors usually report both baseline and treatment effect heterogeneity. In contrast, if there is no statistical significance, information about (significant) baseline risk heterogeneity might be omitted. Although from a clinical point of view only treatment effect heterogeneity might be considered important, systematic variation between patients in baseline risk is also a relevant source of heterogeneity from a decision-making perspective. Indeed, between patient heterogeneity in baseline risk – even in the presence of a homogeneous relative treatment effect – yields heterogeneous absolute treatment effects, which interests policy makers because it has implications both for budget impact and equity concerns.

A further issue is that, although these tests reflect a genuine interest in achieving reliable and precise estimates of treatment effect differences in subgroups, they have been demonstrated to have low power and a high rate of false negatives. Finally, loss of balance between arms of a trial has also been raised as a concern in estimating treatment effects for subgroups.

All these concerns are relevant for clinical studies and they do not necessarily apply in a similar way to cost-effectiveness analysis. Although inference about treatment effects is mainly based on the magnitude of probability of error (errors type-1 and -2), decision rules should also consider the consequences of those errors. Thus, economic analysis in healthcare is focused on the correct characterization of uncertainty rather than inferential decision rules (e.g., taking *p*-value equal to .05 as a rule of thumb). However, even in the case of decisions that follow these principles, there are some constraints on the study of heterogeneity. For example, characteristics used to explain differences in (cost)-effectiveness between individuals may be constrained by equity considerations. The National Institute for Health and Clinical Excellence (NICE) for England and Wales, for instance, states that subgroup analysis based purely on differences in treatment costs is not relevant to their decisions. Furthermore, transaction costs involved in the operationalization of decisions at an individual level or in different subgroups need to be explicitly considered in the analysis.

Value of Heterogeneity

The consideration of heterogeneity has value for the healthcare system because greater population health can be achieved from a finite budget by conditioning treatment decisions on those factors responsible for such between-patient heterogeneity. Subgroup analysis has been the most common approach to explore heterogeneity in the context of health technology assessment. Coyle *et al.* (2003) represented the value of considering subgroups as the incremental net benefits (INB) that can be gained from a 'stratified' analysis for the case where two interventions are compared. If policy makers restrict the adoption of technologies to those subgroups with positive INB, then the gain derived from making different decisions for different subgroups is the difference between the sum of the positive INB, also termed TINB_S (total INB considering subgroups) and the total INB (TINB, including positive and negative INB). In other words, it is the absolute value of the sum of the INB in those subgroups where the INB is negative. Using an alternative notation, the value of stratification can be expressed as Δ_STINB:

$$\Delta_s TINB = INB_S - TINB = -\sum_{s=1}^{S} INB_S w_S, \quad \forall_S \text{ where } INB_S < 0$$

where $w_s \in (0,1)$ is a weight indicating the proportion of the total population represented by subgroup s and $\sum_{S=1}^{S} w_s = 1$.

Basu and Meltzer (2007) developed a framework for estimating the value of eliciting information at patient level to make individualized decisions. They introduced the concept of expected value of individualized care (EVIC), a metric that reflects the population net benefits (NBs) forgone because of the ignorance of heterogeneity in preferences when decisions are made based on the average estimates. EVIC is calculated as the difference between the average of the maximum NBs in each patient (individual NBs (*i*NBs)) and the maximum of the average NBs of the alternative treatments across patients. This formulation of EVIC has been termed 'with cost-internalization,' in the sense that the decision takes into account the opportunity cost of an alternative resource allocation. According to the original definition, EVIC can be expressed as:

$$EVIC = \int_{\theta \in \Theta} \left\{ \max_j NB(\theta) p(\theta) d\theta \right\} - \max_j \int_{\theta \in \Theta} NB(\theta) p(\theta) d\theta$$

The authors point out that, although EVIC was initially estimated for patient preferences, it can also be estimated for any other (set of) parameter(s) of interest in the decision model. Indeed, a total EVIC captures all parameters of interest and should be interpreted as the expected gains that could be attained if individual information about every patient is considered when estimating the outcome of interest.

EVIC can also be expressed as 'without cost-internalization.' In this case, the decision at individual level follows the rule of maximising expected health benefits instead of net health benefits (i.e., without accounting for opportunity costs). In their first application of EVIC to real data, the authors demonstrate how the value of individualized information can be affected by the decision rule applied. Using an illustrative example of alternative treatments for prostate cancer, the estimated EVIC with cost-internalization was greater than US$70 million, this value fell to US$0.9 million without cost-internalization, suggesting that efforts to elicit individualized information is much more valuable if doctors (and patients) internalize costs when making their decisions. Basu and Meltzer also presented parameter-specific EVIC ($\text{EVIC}_{\theta i}$), which is analogous to the expected value of perfect information for parameters. An advantage of this metric is that by ranking parameters according to $\text{EVIC}_{\theta i}$ the most valuable information for individualized decisions can be identified.

These recent methodological developments provide an adequate representation of the potential health that can be gained if heterogeneity is taken into account in decision-making. It is important to highlight that EVIC (total and for specific parameters) is conditional to the structure of and evidence within the decision model. Thus, if the model fails to capture an important source of heterogeneity, the estimate of EVIC may be unreliable. EVIC can be estimated from individual patient data or from aggregate data.

Current approaches to express the value of heterogeneity estimate the expected value of the health that could be gained by considering heterogeneity. However, sampling uncertainty must also be considered as part of the same characterization. For example, if EVIC for the parameter 'polymorphism A' represents the value of conducting a pharmacogenetic test to reveal whether the patient has such a polymorphism, then the estimate of EVIC implicitly assumes that the effect of having the polymorphism on the outcome is known with total precision, and also that the test is 100% accurate. Thus, EVIC provides an estimate only of the potential value of making different decision for patients with and without the polymorphism, but it does not provide any information about the probability that such alternative decisions are wrong. Consequently, an important issue that needs to be addressed is the role of decision uncertainty when heterogeneity is taken into account.

Preferences and Choice as Sources of Heterogeneity

Preferences and choices are concepts with important implications for the study of heterogeneity across individuals.

Preferences as a Source of Heterogeneity

Preferences have been central to how health outcomes have been valued in CEA, where the primary objective is to maximize health gain subject to a budget constraint. CEA often uses quality adjusted life years (QALYs) as a measure of health gain. Although the QALY can only be assumed to accord with individual preferences under very strong assumptions, quality of life weights are generally taken as reflecting the preferences of the relevant group of responders (typically patients or the public). Indeed, some methods used to elicit quality of life weights for QALYs have a strong basis in preference theory (e.g., the standard gamble method is derived from expected utility theory). These methods estimate a relative value of descriptive health states, which are a representation of a particular level of health related quality of life (HRQoL).

Although heterogeneity in preferences was an important part of the development of the concept of EVIC by Basu and Meltzer, relatively few studies have addressed the idea of considering heterogeneity in patients' preferences. Nease and Owens (1994) introduced the idea of estimating individualized expected health benefits to realize the value of a guideline that considers individual patients' preferences. Using a decision model for mild hypertension, they showed that decisions guided on the basis of individualized preference assessment should be considered cost-effective compared to average preference estimates. Sculpher (1998) compared different preference-based approaches to treatment allocation (based on expected individual health, expected individual cost-effectiveness and free treatment choice by the patient), revealing that decisions based on expected individual QALYs and net QALYs are not well correlated with treatment choice. This probably reflects the limited link between QALYs and individual preferences. In other words, patients were basing their treatment choices on criteria not reflected in the derivation of the QALY.

Choice as a Source of Heterogeneity

An optimal (treatment) choice for an individual patient is one that maximizes the individual's welfare, utility or health depending on the elements in his/her objective function. In the context of healthcare, ex-ante choices are the decisions that a data analyst expects the patients to make based on some of the observed patient characteristics but without access to other relevant information and points of views that patients may face while making actual decisions. This view contrasts with the notion of treatment selection (or revealed choices or ex-post choices) which is the individual's decision resulting from the interaction of the patient with health professionals, relatives and other sources of information that are relevant for the decision, but were unobserved to the data analyst trying to predict these choices. This can be operationalized in the context of, for example, a shared decision-making model, where patients and health professionals share information about alternative diagnostic and treatment options as well as outcome preferences with the aim of making the best choice among the alternative courses of action. Ex-post choices can also be driven by anticipated gains and losses. To the extent that these anticipations are not completely unfounded and they deviate from the average gain and loss from a treatment, ex-ante prediction of choices can be substantially different from ex-post choices. This has implications for policy making.

A policy concern for many healthcare systems is that patients' preferences and choices should be taken into consideration in the decision-making process. NICE, for example, recognizes this argument as part of its social value statement, but it also highlights the importance of making adequate judgments to ensure good use of the limited resources. Less clear, however, is the extent to which patients' unconstrained treatment choices can be consistent with the social objective of maximising health gain subject to finite resources. One possibility is that patients' choices can provide some information about the expected potential health gains from a particular treatment. In other words, choices provide information on the extent to which a patient expects (or is expected) to benefit from an intervention.

In the clinical trials literature it has been reported that when patients are allocated to their preferred treatments, their outcomes are affected positively without effect on attrition rates. This might indicate that treatment works better in patients who would choose it, irrespective of the causes that explain loss in follow-up. If ex-ante choices can be used to predict outcomes, then they could help select treatment as a form of subgroup analysis. However, findings indicating that ex-ante choices are not good predictors have also been reported. Although the role of ex-ante choices as predictors is not clear, this might not be the case for ex-post choices. Given the process needed to reveal those choices, they are likely to be more predictive of health outcomes than ex-ante choices. If so, revealed choices might correlate strongly with many other unobserved covariates that explain variability in health outcomes. Thus, by using appropriate statistical techniques,

individual treatment effects could be estimated and their heterogeneity at individual level characterized, producing a better understanding of the joint distribution of potential health outcomes (potential outcomes are defined, according to the Rubin's causality model, as the observed consequences (Y) of alternative treatments ($t=0,1$) in one particular individual (i), i.e., the outcome observed *de facto* and the counterfactual (unobservable), which defines the joint distribution as $G[Y_{0i}, Y_{1i}]$). A research agenda for understanding heterogeneity should include new approaches to reveal individual choices and their role in explaining variability in health outcomes. This should address alternative study designs and analytical techniques.

Some governments and health systems value providing patients with (at least some) unconstrained choices over the healthcare they receive regardless of the impact on their ultimate health outcome. This principle of patient autonomy may, however, clash with an efficiency objective of maximising health across population from available resources. That is, owing to resource limitations, one patient's choice can be another patient's health loss. To the extent that social decision makers have a more complex objective function which includes population health and patient autonomy, then economic evaluation will need to establish how one objective is valued against the other.

Conclusions

In conclusion, heterogeneity in decision-making is occupying an important place in the health research agenda, not only because there is an intrinsic value for individualization of care but also because it is consistent with the objectives of maximizing health under limited budgets. Important conceptual and methods contributions have made in the past few years; however, there are still several gaps that require further research. Future investigation should examine the need to produce a more systematic approach to exploring heterogeneity (e.g., through subgroup analysis), the incorporation of parameter uncertainty in a more integrative framework with heterogeneity and the exploration of the role of patient choices in explaining variation in health outcomes.

References

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making* **27**(2), 112–127.

Coyle, D., Buxton, M. J. and O'Brien, B. J. (2003). Stratified cost-effectiveness analysis: A framework for establishing efficient limited use criteria. *Health Economics* **12**, 421–427.

Nease, Jr, R. F. and Owens, D. K. (1994). A method for estimating the cost-effectiveness of incorporating patient preferences into practice guidelines. *Medical Decision Making* **14**, 382–392.

Further Reading

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology* **51**, 1173–1182.

Basu, A. (2011). Economics of individualization in comparative effectiveness research and a basis for a patient-centered healthcare. *Journal of Health Economics* **30**(3), 549–559.

Briggs, A., Sculpher, M. J. and Claxton, K. (eds.) (2006). *Decision modelling for health economic evaluation*. Gosport, Hampshire: Oxford University Press.

Conti, R., Veenstra, D. L., Armstrong, K., Lesko, L. J. and Grosse, S. D. (2010). Personalized medicine and genomics: Challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities. *Medical Decision-making* **30**, 328–340.

Hamburg, M. and Collins, F. (2010). The path to personalized medicine. *New England Journal of Medicine* **363**, 301–304.

Heckman, J. J., Clements, N. and Smith, J. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in program impacts. *Review of Economic Studies* **64**, 487–535.

Heckman, J. J., Urzua, S. and Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* **88**, 389–432.

Kravitz, R., Duan, N. and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q* **82**, 661–687.

Manca, A., Rice, N., Sculpher, M. J. and Briggs, A. H. (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis: The use of multilevel models. *Health Economics* **14**, 471–485.

National Institute for Health and Clinical Excellence. (2005). Social value judgments: Principles for the development of NICE guidelines, 2nd ed. London: NICE. Available at: www.nice.org.uk (accessed 30.05.12).

National Institute for Health and Clinical Excellence. (2008). *Guide to the Methods of Technology Appraisal*. Available at: www.nice.org.uk (accessed 30.05.12).

Nease, R. F., Kneeland, T., O'Connor, G. T., et al. (1995). Variation in patient utilities for outcomes of the management of chronic stable angina. *Journal of the American Medical Association* **273**, 1185–1190.

Oxman, A. and Guyatt, G. (1992). A consumer's guide to subgroup analyses. *Annals of Internal Medicine* **116**, 78–84.

Sculpher, M. J. (2008). Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics* **26**, 799–806.

Sculpher, M., Pang, F., Manca, A., et al. (2004). Generalisability in economic evaluation studies in healthcare: A review and case studies. *Health Technology Assessment* **8**, 49.

Stinnett, A. A. and Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision-making* **18**, S68–S80.

Biopharmaceutical and Medical Equipment Industries, Economics of

PM Danzon, University of Pennsylvania, Philadelphia, PA, USA

Introduction

The biopharmaceutical industry (including small molecule drugs, biologics, and vaccines) and the medical equipment industry (including implantable medical devices, diagnostic imaging, and other diagnostics) have been major contributors to both rising healthcare spending and improved quality and quantity of life globally over the past four decades. Global spending on biopharmaceuticals reached one trillion dollars in 2012. Biopharmaceuticals account for between 10% and 20% of healthcare spending in most Organization for Economic Cooperation and Development countries, and often a higher share in developing countries that spend relatively less on hospital and physician services. The medical equipment sector is both conceptually less precisely defined and empirically harder to measure. Industry revenues are estimated at $332 billion (Ernst and Young, 2012), or roughly one-third of biopharmaceutical industry revenues.

The US remains by far the largest single market for these industries. For biopharmaceuticals, the US share of global sales was 34% in 2011, down from 45% in 2000 (Table 1). Over the past decade growth of biopharmaceutical sales has slowed to low, single-digit annual growth rates in North America and Europe, due to patent expiries and genericization of many major drugs and slower growth of new drugs. This contrasts with double-digit growth of biopharmaceutical spending in many emerging markets, particularly China, Brazil, India, and some other countries of Asia, Africa, and Latin America, reflecting their rising incomes and increased spending on health care. For medical equipment, the US share is roughly 45% of global sales.

The economics literature has focused much more heavily on biopharmaceuticals than on medical devices and diagnostics, reflecting both the greater expenditure share of biopharmaceuticals and the greater availability of data. Economic analysis focuses on features that differentiate these industries from other health services or consumer goods industries, in particular: high research and development (R&D) intensity; heavy regulation of all business functions, including R&D, market access, pricing and marketing; and complex market environments due to physicians and payers being major customers, in addition to patients. Economic analysis has taken both a social welfare/policy perspective and a firm or industry perspective. From the policy perspective, key issues related to biopharmaceuticals are the design of intellectual property (IP) rights, regulatory and reimbursement systems to provide appropriate incentives for R&D, and to assure appropriate utilization and prices for drugs, devices, and diagnostics, such that they deliver value for money. From the firm or industry perspective, key issues include understanding the causes of declining R&D productivity and optimal strategic responses; measurement and demonstration of incremental value of new compounds to regulators and payers; and development of effective entry and sales strategies for emerging markets. Because regulation of market access, pricing, and reimbursement are decided by each country separately, global policy and strategy must consider the interaction of policies adopted in different countries, in particular, the many challenges related to segmentation and differential pricing when selling global products in markets that differ vastly in regulation, IP, and ability and willingness to pay.

This overview article on the economics of these industries lays out the theoretical issues and major empirical findings, focusing first on issues related to R&D and then turning to markets, reimbursement and pricing, promotion, and specific issues related to vaccines, personalized medicine, and biosimilars. Although this article focuses on biopharmaceuticals, reflecting the much larger literature, it also describes ways in

Table 1 World pharmaceutical markets

Region	*Pharmaceutical sales (US$ billion)*			*Percentage of worldwide sales (%)*		
	2006	*2011*	*2016 (estimate)*	*2006*	*2011*	*2016 (estimate)*
US	269.78	325.04	368.9	41	34	31
Canada	13.16	19.12	23.8	2	2	2
EU5	125.02	162.52	154.7	19	17	13
Rest of Europe	46.06	66.92	59.5	7	7	5
Japan	65.8	114.72	119	10	12	10
Pharmerging	92.12	191.2	357	14	20	30
Rest of world	46.06	76.48	107.1	7	8	9
Total	658.00	956.00	1190.00	100	100	100

Notes: Spending in US$ with variable exchange rates. Pharmerging countries are defined as those with > $1 billion absolute spending growth over 2012–16 and which have GDP per capita of less than $25 000 at purchasing power parity. Pharmerging markets include China, Brazil, India, Russia, Mexico, Turkey, Poland, Venezuela, Argentina, Indonesia, South Africa, Thailand, Romania, Egypt, Ukraine, Pakistan, and Vietnam. Rest of Europe excludes Russia, Turkey, Poland, Romania, Ukraine, which are included in the pharmerging markets.
Source: Reproduced with permission from Market Prognosis (2012). *Report of the IMS Institute of Healthcare Informatics*. Available at: www.imshealth.com (accessed 20.03.13).

which medical equipment is similar and different. Other articles in this volume provide greater depth on various issues.

R&D: Costs, Regulation, and IP

R&D Costs and Regulation

The biopharmaceutical industry is unusually research intensive. The US research-based industry invests approximately 15% of its sales in R&D, compared with approximately 4% for US industry in general and 8% for the US-based medical device industry. The R&D cost of bringing a new medical entity (NME) to market is currently estimated to be approximately $1.5 billion (Mestre-Ferrandiz *et al.*, 2012) and take 5–12 years from discovery through development, clinical trials, and regulatory approval. New drugs must meet stringent standards of safety, efficacy, and manufacturing quality before receiving market access approval. Large and lengthy clinical trials to demonstrate safety and efficacy, with high failure rates, are major drivers of the high cost per approved NME. Throughout the 1970s, 1980s, and 1990s, the cost per approved new drug increased by seven to eight percentage points per year above general price inflation. Factors contributing to rising cost per NME include not only rising clinical trial costs but also, more recently, higher failure rates. The evidence suggests that of drugs entering human clinical trials, only one in seven or eight reaches approval, compared to one in five in the 1990s. Rising failure rates reflect both safety, efficacy, and economic factors. Recent scientific advances have enabled development of novel therapies, but predictability remains imperfect. Further, because good treatments already exist for easier diseases, new drugs must now either provide significant incremental value relative to existing drugs that are available as low-priced generics, or tackle diseases that pose tougher scientific challenges, such as Alzheimer's disease and cancer, or target diseases that were previously ignored due to small populations. Most recently approved drugs target either specialty conditions (complex, relatively uncommon diseases treated by specialists) or even small orphan indications (defined in the US as affecting less than 200 000 patients per year). In the US in 2010 and 2011, one-third of new active substances approved had orphan designation. This reflects the intended incentives provided by the Orphan Drug Act, which provides special tax credits and market exclusivities for drugs that receive orphan status, as well as the very high prices realized by some orphan drugs, now more than $400 000 per patient per year for some drugs. It also reflects the granting of orphan status for small indications for drugs that may subsequently be approved for other, larger indications – for example, many cancer drugs serve both orphan and nonorphan indications.

The cost of developing a new drug includes the out-of-pocket expenses incurred by firms from discovery through first approval on the successful compound and related failures, because failures are an unavoidable part of the process. The full, capitalized cost per approved NME also includes the opportunity cost of capital invested, because investors must recoup their opportunity cost in order to continue investing in R&D. This cost of capital is about half the total cost (Di Masi and Grabowski, 2007). Although the mean cost is estimated at US$1.5 billion (Mestre-Ferrandiz *et al.*, 2012), there is significant variation with lower costs for rare diseases that necessarily have smaller trials, and relatively high costs for drugs to treat high-volume, chronic diseases that require large and long trials.

R&D expense for medical devices is much lower than that for drugs. Devices are classified into classes I through III, based on risk to patients and device novelty. The US Food and Drug Administration (FDA) has oversight over device safety, efficacy, and quality, but clinical trials are usually required only for novel devices classified as class III. Most devices are incremental modifications of existing products and can be approved by showing 'substantial similarity' to an existing device, without clinical trials. The EU's CE mark system authorizes either state or private oversight bodies to review safety and quality, and proof of efficacy is not required. Devices are therefore often launched earlier in the EU than the US, in contrast to drugs for which EU launch is often delayed by reimbursement requirements.

Safety: Benefits and Costs

Market access regulation that requires demonstration of safety and efficacy entails costs as well as benefits. The appropriate extent and structure of this regulation has been debated in the academic and policy literatures. The main economic focus has been whether the current regulatory approach to drug approval provides an optimal trade-off between safety and delay. The benefits of regulation include preventing unsafe and ineffective drugs from being sold and requiring the production of unbiased information about drug outcomes, including risks, benefits, and contraindications as demonstrated in controlled trials. The statistically significant findings from clinical trials form the basis for the product label and approved promotional messages. By revealing the true expected benefits and risks from drugs before launch, such information reduces the risk of adverse outcomes and drug withdrawals for safety reasons.

The costs of market access regulation include increased development costs, which may keep some potential drugs off the market, and delay in consumer access to new drugs. The FDA User Fees (which fund the hiring of additional reviewers) and the Fast Track and Priority Review regulatory initiatives have accelerated the review process of new drugs and provided mechanisms for approval based on surrogate endpoints, with postlaunch follow-up. Despite some mixed evidence that more rapid reviews have resulted in more postlaunch adverse events and drug withdrawals, on balance the evidence from pharmaceuticals suggests that these initiatives have increased consumer welfare. For medical devices, the appropriate structure and requirements for review are still under debate in the US. Delays in approval relative to the EU are a concern, but so is the number of recalls of devices approved through the accelerated process. Future economic research is needed on the optimal structure of market access regulation for medical devices.

Patents, Exclusivities, and Other Research and Development Incentives

The high cost of R&D for biopharmaceuticals (and, to a lesser extent, medical devices) implies a cost structure with high fixed costs that can benefit consumers globally but are sunk at launch, with low marginal cost per pill. Investment in the costly and risky process of pharmaceutical R&D therefore requires some mechanism to assure a return on successful investments for originator firms. The standard approach is patents which grant the innovator a monopoly for the duration of the patent by barring identical copies. Defining appropriate patent terms and criteria for postpatent generic entry are critical policy issues. All countries that are members of the World Trade Organization must recognize 20-year product patents, running from date of filing, for all products that meet requirements of novelty and utility, not just pharmaceuticals.

In addition to this basic patent protection that applies to all types of goods, the US and many other countries have added regulatory provisions that define certain exclusivity protections for qualifying originator pharmaceuticals, partially make-up for patent term lost before launch due to the lengthy R&D process, and also define entry conditions for generics. In the US, the 1984 Hatch–Waxman Patent Restoration and Generic Competition Act extended patent terms and defined regulatory exclusivities for originators, and eased entry requirements for generic versions of small molecule drugs. Specifically, Hatch–Waxman provided originator drugs with up to 5 years of patent restoration to compensate for patent life lost during R&D and regulatory review, and 5 years of exclusivity for originator data before generics can reference the data. For generics, Hatch–Waxman provided an Abbreviated New Drug Approval (ANDA) pathway that enables generics to be approved without doing new safety and efficacy trials, provided they can show bioequivalence to the originator drug and reference the originator safety and efficacy data. Paragraph IV provides a 180-day market exclusivity for the first ANDA generic that successfully challenges originator patents, to incentivize challenge to dubious patents.

The ANDA provisions greatly reduced the regulatory costs of approval for generics and facilitated the growth of generics in the US. The 180-day exclusivity period has led to successful challenges of many patents, and hence speeded generic entry. Generics now account for more than 80% of all prescriptions dispensed in the US, and a higher percentage for compounds for which generics are available. Unsurprisingly, because patentability requires that an invention be new, useful, and nonobvious, original composition-of-matter patents that apply to new molecules have generally withstood generic challenge in the US, whereas additional patents filed later on ancillary features or new delivery systems have more frequently been successfully challenged for failing to meet requirements of novelty and nonobviousness. The requirements for proof of novelty and nonobviousness differ across countries. This has led to some products that are patented in the US being denied patents in countries such as India.

Given the experience of patent litigation and uncertainty under the Hatch–Waxman Act, the 2010 Affordable Care Act (ACA) provisions for a new regulatory approval pathway for follow-on biologics (biosimilars) has focused on the regulatory exclusivity period for originator data. This is currently set at 12 years from the first licensing of the referenced biologic, in contrast to 5-year data exclusivity for chemical drugs in the US. Whether this much longer exclusivity period, combined with more favorable reimbursement for biologics, potentially distorts R&D choices toward biologics, despite their lower convenience and higher cost for consumers, is an important topic for future research. In contrast to these discrepant US data exclusivity periods, the EU grants 10 years of data exclusivity for both chemical and biologic drugs.

More generally, regulatory exclusivities offer more flexibility of duration and more certainty of enforcement, compared to patents that must run for 20 years from filing but may be challenged. However, this flexibility may make regulatory exclusivities more subject to manipulation by special interests. Given the vastly different costs involved in different types of biopharmaceutical and medical technology R&D, use of both patents and the more flexible exclusivities seems optimal.

For medical devices, patents are important but in general create weaker and less durable market power than for pharmaceuticals, because it is relatively easy to invent around a medical device patent using a slightly different product design. Moreover, entry of incrementally improved, follow-on devices renders the original design obsolete within a few years, even if the 20-year patent nominally remains valid.

Although patents are in some respects an efficient and effective mechanism to incentivize R&D, patents have other disadvantages besides the inflexible term and uncertain validity already mentioned. In particular, patents operate by limiting competition and enabling innovator firms to charge prices above marginal cost, which can lead to suboptimal use of drugs in the absence of insurance. High price–marginal cost margins also create strong incentives for promotion. Several alternatives to patents have been proposed for pharmaceuticals, including both 'push' programs that provide subsidies to reduce the cost of R&D and 'pull' programs that increase and/or guarantee revenues for companies that bring new drugs to market, including prizes, patent buyouts, and advance market commitments. Some of these alternatives have been applied to R&D for 'neglected' diseases with prevalence predominantly in low-income countries, including the advance market commitment for the pneumococcal vaccine.

Further research is needed on the optimal mix of IP alternatives, including patents, exclusivities, and others, for specific R&D contexts related to drugs, devices, and other technologies, in order to appropriately reward innovation without granting inefficient barriers to entry. Such research should consider how the optimal mix of protections might differ across countries at different levels of development. Because the goal of IP or other protections is to provide an appropriate financial reward to innovators, the optimal type and duration of IP should ideally also consider the pricing and reimbursement environment, which determines the prices and revenues that can be earned during the protection period. More on this below.

Mergers, Alliances, and Organization of R&D

The basic and translational science underlying many new drugs is developed in academic institutions, often supported by government research grants. The traditional mechanism for developing and commercializing such technologies has been the creation of start-up companies, usually with venture capital funding, taking advantage of the Bayh–Dole Act that encourages private commercialization of publicly funded research. Over the past two decades, thousands of start-up firms have been formed, many have been acquired by larger, established firms, some have failed, and a few have grown to become fully integrated biotechnology companies. Over time, the share of new approved drugs that originated with small firms has grown.

As large pharmaceutical firms have experienced declining returns on their internal R&D, they are increasingly using product licensing alliances and outright acquisition of small firms to source new compounds externally. For the small firms, such alliances with established biopharmaceutical firms provide an important source of R&D financing, as well as regulatory and commercial experience and expertise. The terms of these alliances and acquisitions are structured to align incentives and share risk, through payments that are triggered only if the product achieves certain goals. These contingent payments include R&D milestone payments, tiered sales royalties, opt-in options for the licensee in alliances, and contingent valuation rights linked to sales in acquisitions.

The theoretical literature has hypothesized that formation of product development alliances may be hampered by asymmetric information. However, contingent payments in the deal structure are designed to address both adverse selection and moral hazard risks. The empirical literature is mixed, but in general finds that in-licensed products have a higher probability of success than internally developed products, which supports the notion that the stringent due diligence process of alliance formation is more rigorous at weeding out compounds that will ultimately fail, compared to internal R&D review processes within large firms.

In addition to alliances with small firms, several large firms have recently reorganized their drug discovery divisions into small units that attempt to mimic the entrepreneurial spirit and incentives of small firms. The compounds that are produced by these internal units must compete with externally sourced compounds for scarce resources to fund clinical trials. Other attempts to increase R&D productivity within large firms include changes in personnel and organizational structure, and changes in compensation schemes. Despite all these attempts to improve R&D productivity, several large pharmaceutical companies have cut their R&D budgets recently for the first time in decades and instituted share buy-back programs, in response to shareholder concerns about the low return on R&D investment.

Small firms are not immune to the rising costs of R&D and high failure rates. Longer and riskier investment cycles and uncertainty of exit through either acquisition or an initial public offering have also slowed the flow of venture capital into formation of early-stage biotechnology companies. This decline in private equity and venture funding for start-ups has been partially offset by an increase in alliances directly between large pharmaceutical firms and academic institutions and a growth in funding through the corporate venture capital arms of large biopharma firms. These and other creative financing developments suggest that there may be efficiency gains from facilitating mechanisms to finance the development of new products without the formation of new start-up companies around each idea.

Markets for Biopharmaceuticals and Medical Technology

Principles of Optimal Insurance

The market for pharmaceuticals in any country depends on the extent of insurance and on the rules of reimbursement used by payers to control the effects of insurance on prices and utilization. Insurance protects consumers against the financial risk of high drug spending but also makes consumers insensitive to drug prices. Demand-side price sensitivity is further undermined by the fact that physicians who prescribe drugs often lack the information and incentives to make price-sensitive choices. Inelastic demand of insured consumers creates incentives for firms to charge higher prices than they would if consumers were informed decision-makers facing full prices. To address this insurance-induced price insensitivity, insurers in most countries use a range of strategies to control prices and utilization of prescription drugs.

The optimal design of insurance coverage is a critical policy issue that affects patients' access and financial exposure, innovation incentives for firms, and budget impact for taxpayers and consumers. In theory, insurance coverage and eligibility should be designed to encourage optimal utilization of existing drugs (static efficiency) and optimal incentives for R&D investment for new drugs (dynamic efficiency) and provide reasonable financial protection for patients. One proposed approach to achieving these three goals is that copayments should be set at marginal cost while the health insurer pays a top-up payment to the biopharmaceutical firm to reward innovation (Lackdawalla and Sood, 2009). In practice, both marginal cost and appropriate top-up payments are difficult to observe, and this approach ignores appropriate financial protection for patients.

An alternative approach, that could in theory achieve second-best static and dynamic efficiency and appropriate financial protection for patients, is for each payer to make reimbursement of a drug conditional on meeting an incremental cost-effectiveness ratio (ICER) threshold – for example, $50 000 per quality-adjusted life-year (QALY) – that reflects the willingness-to-pay for health gain of that payer's enrollees or citizens (Danzon *et al.* 2012). The firm would be permitted to price up to the ICER threshold, but this implies that the price premium would be constrained by the new drug's incremental benefit relative to the comparator or standard of care. The payer would also define coverage eligibility to assure access for patients for whom the drug is cost-effective at the price charged. Copayments would be modest, to collect some revenue but assure affordability. This approach encourages appropriate innovation, by paying a premium for new drugs

that is based on their incremental value, and assures access for patients. If all countries with comprehensive insurance set ICER thresholds unilaterally, based on their willingness to pay for health, manufacturers would have incentives to set prices that differ across countries, reflecting countries' willingness and ability to pay. This result is broadly consistent with Ramsey pricing principles applied to R&D as a joint cost.

In practice, pharmaceutical pricing and reimbursement regulation differs across countries but follows four broad prototypes: (1) the USA exemplifies free pricing in a pluralistic insurance market with competing health plans; (2) Europe exemplifies several approaches to setting price and reimbursement in universal insurance systems; (3) Japan exemplifies price regulation in a market where physicians traditionally dispensed drugs; and (4) many emerging markets illustrate predominantly self-pay markets for drugs. The following sections describe key economic issues in each of these prototypical markets.

Free Pricing with Competing Payers: The US

In the pluralistic US healthcare system, no single payer has sufficient market power to significantly influence prices. Payers rely primarily on tiered formularies and costsharing to preserve some patient price-sensitivity and to enable payers to negotiate discounts in return for preferred formulary status. Although list prices are unconstrained, tiered formularies have achieved significant discounts in therapeutic classes with close therapeutic substitutes. However, in classes with few and/or differentiated products, which includes most specialty drugs and biologics, payers have not used tiered formularies aggressively to attempt to extract discounts. Rather, they rely increasingly on specialty tiers with 20–30% coinsurance rates. However, most patients are protected by catastrophic limits on costsharing or manufacturer copay coupons, which provides appropriate financial protection but leaves little if any constraint on prices. Launch prices for new drugs therefore continue to rise, with several more than $100 000 per year or per treatment course. Similarly, for physician-dispensed biologics, the reimbursement rules create incentives for high launch prices, with little constraint from patient costsharing.

By contrast, generic markets in the US are highly price competitive. High rates of generic entry and penetration, combined with low generic prices, reflect not only the Hatch–Waxman provisions requiring bioequivalence with low entry costs, but also pharmacy substitution and reimbursement rules that assure price-conscious dispensing choices by pharmacies and patient acceptance of generics. Over the past 15 years, patent expiration on many originator drugs has enabled a massive shift toward generics. In 2012, more than 80% of prescriptions were dispensed generically, up from 47% in 2000, but generics account for only approximately 30% of sales by value, due to their low prices. Generic penetration rates are higher and generic prices are absolutely lower in the US than in many other countries (Danzon and Furukawa, 2011). This has provided significant savings to consumers and created budget headroom for high-priced new drugs. As the flow of new generics declines, attention may shift to better ways to assure value for money while preserving access to new pharmaceuticals in the US.

Effects of cost sharing

Patient cost sharing is an important feature of health-insurance design, particularly in the US. In theory, optimal cost sharing balances financial protection of patients against deterring overuse of services and excessive pricing. If other constraints on pricing or use are also used, then optimal cost sharing can be lower. Conversely, Garber *et al.* (2006) show that at levels of cost sharing that are optimal for patient protection, prices would exceed levels needed to incentivize optimal R&D, assuming current patent design is optimal. Unsurprisingly, cost-sharing levels are highest and studies of cost-sharing effects are most numerous in the US.

Because details of cost-sharing structure, levels, stop-loss, and other controls differ across contexts, generalizations are problematic. With that caveat, the evidence confirms that tiered cost sharing affects choices between drugs. Even modest cost sharing affects utilization and compliance. Recent studies have focused on the interconnection between utilization of drugs and utilization of other services, which may be complements (a physician visit may be necessary to get a prescription) or substitutes (compliance with medications may reduce disease flare-ups and emergency visits). Evidence that even modest cost sharing for some chronic medications can significantly affect utilization of more costly medical services has generated great interest in 'value-based insurance design,' which would take these complementarities into account in designing cost sharing. Further research is needed into how optimal cost-sharing structures differ across disease states and drug types, and how their effects in practice are modified by stop-loss limits, manufacturer coupons, and other offsets.

Price and Reimbursement Regulation: The EU

In most industrialized countries with comprehensive insurance, payers control prices and utilization of biopharmaceuticals, with a view to maintaining access while managing within fixed health budgets. Price regulatory systems use three prototypical approaches to setting prices, and some countries use variants of multiple approaches.

Internal referencing

Internal referencing compares the health outcomes with the new drug relative to one or more existing drugs and grants a price premium only if the new drug demonstrates superior safety, efficacy, or other benefits. In principle, this approach rewards innovation that produces measurable incremental value. It is usually applied only at launch. Postlaunch price increases are generally not allowed, and price decreases may be mandated if total expenditure for a drug exceeds the payer's target based on the expected number of eligible patients. These 'volume-price offsets' reduce the price in proportion to the expenditure overrun. This not only keeps expenditure within target but also deters promotion beyond the target population.

A special case of internal referencing is 'reference price reimbursement,' as implemented in Germany and the

Netherlands, in which the payer groups drugs based on similarity of indication, therapeutic effects, and sometimes mechanism of action. The reference price is the maximum reimbursement price for all drugs in the group, and if the actual price is higher, the patient must pay the excess. The reference price is usually based on a low-priced drug within the group, which could be a generic. If classes are broadly defined and ignore significant differences between drugs, this approach can undermine incentives for incremental innovation within a class. In Germany's post-2010 approach to drug pricing, the first step is a formal review of the new drug, relative to comparators. If the new drug is deemed to offer no significant improvement it is assigned to a reference pricing group and is reimbursed at the prevailing reference price. If it is deemed significantly superior, then a new price is negotiated or determined by arbitration. Thus this approach recognizes the importance of benefit evaluation before assigning a drug to reference pricing.

External referencing

With external referencing, the price of the new drug in country X is set at the mean, median, or minimum price of the same drug in a specified set of other countries. This approach is widely used in the EU, and the external reference may be the EU average price. This approach undermines the firm's ability to maintain price differentials between countries although, as noted earlier, such differentials are consistent with Ramsey pricing principles applied to paying for the joint costs of R&D. Further, external referencing creates incentives for firms to delay or not launch drugs in small, low-priced countries, if these prices might undermine potentially higher prices in other countries. Several studies have found evidence of such delays and nonlaunch due to referencing within the EU. Thus, external referencing by one country can lead to spill-over reductions in access and presumably social welfare in referenced countries.

Parallel trade

Although parallel trade is not a form of direct price regulation, it has effects similar to external referencing, but on a more limited scale. Parallel trade (also called commercial drug importation) permits commercial third parties – usually pharmacies and wholesalers – in one country to import drugs purchased in other, lower-priced countries, effectively arbitraging the price differences. The EU authorizes parallel trade between EU member countries as part of the general policy of free movement of goods within the EU.

Although economic theory generally concludes that free trade increases social welfare by enabling consumers to source products from lower cost producers and benefit from the savings, these conditions are generally not met for parallel trade in drugs. Price differentials for drugs between EU countries reflect differences in income and regulatory systems, not differences in production costs, hence there is no resource efficiency gain from such trade. On the contrary, parallel traded goods often require repackaging or relabeling which adds to resource costs. Further, the savings from arbitraging differences in exmanufacturer prices are largely captured by middlemen and are not transferred to consumers/payers. If the net effect of parallel trade is revenue redistribution from manufacturers to distributors that results in reduced incentives for R&D, then the efficiency effect of parallel trade is likely negative.

Cost-effectiveness review

An indirect approach to price control results when the payer reviews the incremental cost-effectiveness of a new drug, relative to standard of care, as a condition of reimbursement. The UK's National Institute for Clinical Excellence exemplifies this approach, with detailed methodological requirements and an explicit threshold cost per QALY. Other countries, including Australia, Canada, and Sweden use similar approaches. If the manufacturer is permitted to set a price up to the maximum at which the new drug meets the ICER threshold, then this approach acts as an indirect control on price that rewards innovation and enables the manufacturer to capture the benefits produced, as required for dynamic efficiency, but without the payer having to directly regulate the price.

Conceptually, it is a simple step to convert cost-effectiveness analysis (CEA) review into an explicit value-based pricing (VBP) regime. VBP would allow a new drug a price premium over current treatment commensurate with its incremental value, which includes both incremental health benefits plus any cost savings. This VBP might be adjusted postlaunch, if the evidence on incremental benefits changes. Whether the VBP should be adjusted if the price of the comparator changes due, for example, to generic entry, is an important policy question that requires further research.

Measurement of Value

If payers are concerned to get maximum value from their expenditures on medical care, then measurement of value of health gain, using CEA and other approaches, is essential. CEA is used as part of broader health technology assessment (HTA) programs to evaluate the incremental health-related effects and costs of new technologies, including drugs, relative to existing technologies. This approach was adopted in the 1990s in Australia, New Zealand, the UK, and Canada, and variants have since been adopted in an increasing number of countries in Europe and more recently in Asia and Latin America. In the US, there is growing interest in comparative-effectiveness research, but with political reluctance to explicitly use cost per QALY or other outcome measures to make reimbursement decisions. CEA grew out of more general HTA, as payers sought more systematic, evidence-based approaches to resource allocation and adoption of costly new technologies within limited budgets.

Implementing value measurement raises both theoretical and practical issues that are being worked out as payers attempt to apply CEA to regulation of pharmaceutical use and prices. Practical questions include what types of evidence to use and how to deal with the inevitable gaps in evidence, especially at launch; use of risk- or cost-sharing contracts when evidence is uncertain; and use of CEA as one among several criteria considered by decision makers. Considerable progress has been made over the past two decades in both theory and measurement of value, primarily using QALYs. Although many

criticisms remain, similar and other criticisms are likely to apply to any alternative metric that attempts to provide a unidimensional measure of value that can compare outcomes across different health interventions. Until superior alternatives are developed, QALYs are likely to remain widely used.

Physician Dispensing

Pharmaceutical reimbursement raises unique issues in countries with physician dispensing. Japan, Taiwan, and South Korea have traditionally exemplified this approach, but each has recently taken steps to separate prescribing and dispensing, in contrast to China where most drugs are still prescribed and dispensed in hospitals and clinics. Simple economic theory and casual observation suggest that where physicians dispense the drugs that they prescribe and can profit from the margin between a drug's acquisition cost and their reimbursement, manufacturers will offer discounts in order to increase this profit margin. The financial incentives of physicians may, therefore, lead to excessive prescribing and bias toward high-margin drugs. Japan traditionally mitigated this effect by biennial review of acquisition prices and downward revision of reimbursement prices to squeeze the margin.

Since 2000, Japan, South Korea, and Taiwan have all taken steps to encourage switching to pharmacy dispensing. The fundamental challenge is that if dispensing income is a significant fraction of total income for physicians, then payers are under pressure to increase other payments to physicians, in addition to now paying pharmacy dispensing fees, which may increase total expenditures. Japan took a gradual, incentive-based approach, paying increased prescription issuance fees for physicians and dispensing fees for pharmacists. The share of prescriptions dispensed through pharmacies has increased to more than 60% in 2011, but cost savings are uncertain because of the additional fees. Korea abruptly required that physicians cease dispensing drugs, which led to physician protests, increased fees, and apparently a shift to higher priced drugs. In response to physician protests, Taiwan allowed clinics affiliated with physician offices to continue dispensing as long as they hired a pharmacist and paid additional fees. Hence, again there has been no reduction in total medical expenditures. Thus, although the evidence suggests that physician prescribing does distort utilization, changing this is not easy and may lead to higher, not lower expenditures, at least in the short run.

Promotion

Biopharmaceuticals

Because the potential benefits and risks of pharmaceuticals are intrinsically nonobvious, providing information to physicians and consumers about a drug's potential effects is critical to its appropriate use. Such information dissemination is provided and financed largely by pharmaceutical firms, through detailing of physicians, journal advertising, distribution of free samples, and direct-to-consumer advertising (permitted only in the US and New Zealand), subject to regulations that differ across countries. The economic and policy issues raised by such types of pharmaceutical promotion are discussed in another part of this encyclopedia. Estimates of the advertising-to-sales ratio in the US range from 6.7% to 18%. The highest estimates include samples valued at retail prices, which significantly overestimate the cost of samples to firms. High advertising-to-sales ratios reflect both the fact of multiple customers – physicians, patients, and payers – and the incentives created by inelastic demand resulting from extensive insurance coverage and high price-to-marginal cost ratios.

The economic literature on promotion is mainly from the US. It suggests that advertising may be both informative and persuasive, and both characteristics apply to some pharmaceutical advertising. Implications for public health and welfare depend on whether or how far advertising raises brand-specific versus industry-wide demand, impacts drug costs, and impacts competition and prices. Empirical evidence is mixed but suggests that consumer advertising is more effective at enlarging the general market, through more physician contact, expanded treatment, etc., whereas physician advertising is primarily persuasive, although the informative role is likely to be greater early in a drug's lifecycle. There is no strong evidence that either consumer or physician-directed promotion raises prices. An overall welfare assessment would require a balancing of complex benefits and costs, and conclusions may depend on type of drug, stage of lifecycle, and other factors that affect the relative magnitude and value of information versus persuasion.

Medical devices

Promotion of medical devices and equipment varies by sector, depending on the user/decision-maker, usually a hospital. However, for complex, implantable devices such as hips or stents, the surgeons who insert the devices may also be major customers because their ease of use with a device affects their time required and willingness to use a device. Such devices require promotion by technically qualified, skilled salespersons who may also play an important role in training the surgeons on how to use the devices. The empirical evidence suggests significant economies of scale in device marketing. This is plausible, because larger firms that produce a full range of products for a particular medical specialty, for example, orthopedics, can spread the fixed costs of hiring and training a dedicated salesforce that promotes only their products, whereas smaller firms that produce only one product may have to rely on general distributors who handle competitors' products. Such economies of scale in marketing are plausibly one factor accounting for the general pattern that small-device firms with good products are usually acquired by larger firms, rather than attempting to seek external financing to grow as independent competitors. Comprehensive data on promotion, sales, and pricing are not available for devices as it is for drugs, hence this remains an important area for future research.

Emerging Markets: Self-Pay for Pharmaceuticals

Pharmaceutical markets in developing countries differ from those of industrialized countries in that insurance coverage

for drugs is very limited, with most people paying directly out-of-pocket, especially those at lower income levels. Theory suggests that manufacturers might seek to practice price discrimination – charging lower prices in these countries than in higher-income countries – if they were assured that the drugs would not be exported to, or their lower prices would not be referenced by, higher-income countries. Similarly, price discrimination between rich and poor consumers within these countries would also increase sales for companies and access for consumers, if it were feasible. However, government policies, distribution systems, and other factors undermine market segmentation in developing countries, although corporate strategies such as dual branding, direct distribution to providers, and consumer coupons can be effective for some drugs. Inefficient distribution systems also pay a role in raising retail prices to consumers, regardless of prices charged by manufacturers in many developing countries.

The global nature of pharmaceutical R&D raises issues of appropriate cross-national price differentials to share the joint costs. Theoretical models of monopoly pricing using either price discrimination or uniform pricing and models of Ramsey pricing applied to payment for the joint costs of R&D suggest that differential pricing is welfare superior to uniform pricing across countries. Assuming that higher-income countries have more inelastic demand, this implies that richer countries should pay higher prices than poorer countries, and this is consistent with most norms of equity. The principle of differential pricing between the richest and poorest nations is widely accepted in policy debates. However, in practice, consensus breaks down on appropriate price differentials and absolute price levels, particularly for middle-income countries with emerging middle classes but large poor populations.

The evidence suggests that drug prices are higher, relative to average per capita income, in low- and middle-income countries. This applies to generics as well as on-patent drugs. Relatively high prices in low- and middle-income countries partly reflects the highly skewed income distributions, which create incentives for firms to target the more affluent segment (Flynn *et al.*, 2006). Further, because regulatory systems in these countries do not require that generic copies be bioequivalent to the originator, quality uncertainty leads producers to compete on brand, using both brand and high price as a proxy for quality (Danzon *et al.*, 2011). In such markets, only the lowest-quality firms compete on price. However, regulatory requirements for bioequivalence of all generics would likely put many local firms out of business. Thus, the obstacles to reform are primarily political.

Vaccines

Preventive vaccines are biologics but differ from other biopharmaceuticals in important aspects. The external costs of infectious diseases imply external benefits from effective vaccines, and this has motivated public mandates, purchasing, and subsidies for vaccines in most countries and government subsidies to supply for particular products, such as Project Bioshield in the US. Relatively small market size and concentrated purchasing have contributed to the existence of few or sole suppliers of most individual vaccines in the US, which has resulted in shortages when the sole supplier experiences production problems.

A considerable literature has examined the cost-effectiveness of different vaccines in different contexts spanning both developed and developing countries, and appropriate policy responses to both suboptimal private demand and sole supplier markets. Policies to promote investment in vaccine R&D include push and pull incentives for the private sector, public production, and the no-fault Vaccine Injury Compensation Program that was implemented in the US in 1986.

After decades of being considered a neglected R&D sector, the past decade has seen a resurgence of interest in vaccines, with several large pharmaceutical companies and many smaller companies entering the US and EU markets, and several WHO-qualified suppliers of vaccines, from India and South Korea, now selling the majority of vaccines to emerging and middle-income countries. Thus, future research must consider factors that differentiate vaccines from other biologics and are common across all or most vaccines and market contexts versus factors that are specific to a particular vaccine or market context. The conditions for purchasing and supplying vaccines differ significantly across countries. Identifying these differences and their effects is a necessary part of generalizing about vaccine economics and appropriate vaccine policy.

Diagnostic Imaging

Like biopharmaceuticals, diagnostic imaging, including computed tomography, magnetic resonance imaging, positron emission tomography, and other technologies, poses challenges related to achieving appropriate use, pricing, and R&D incentives. However, the context and solutions are very different because these are durable machines with high fixed costs but low marginal cost to hospital or physician purchasers. Although a hospital may own the machine, the decision to order a scan is usually made by a physician who is not the same as the radiologist who interprets the scan and is reimbursed. These basic economic issues related to imaging are discussed in another article, focusing on the USA. Another article reviews the reimbursement approaches used in different countries and then discusses the empirical evidence on differences across countries in number of scanners, rates of scans, and expenditures as a percentage of healthcare spending are described in another part of this encyclopedia. These articles establish a foundation and some interesting facts but point out the need for more research in this important area.

Conclusion

The biopharmaceutical and medical equipment industries pose many interesting economic questions that are different from the textbook economic industries or the health services sectors. Like health services, the role of insurance is fundamental in affecting demand. However, because these are research-intensive

industries, optimal insurance and reimbursement design must consider effects on producers' incentives, short and long run, as well as effects on consumer protection. Much progress has been made in understanding the economics of R&D, effects of regulation, promotion, and pricing and reimbursement, particularly for biopharmaceuticals. But this remains a fertile field for future research.

See also: Cross-National Evidence on Use of Radiology. Markets with Physician Dispensing. Patents and Regulatory Exclusivity in the USA. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Regulation of Safety, Efficacy, and Quality. Research and Development Costs and Productivity in Biopharmaceuticals. Vaccine Economics. Value of Drugs in Practice

References

Danzon, P. M. and Furukawa, M. (2011). Cross-national evidence on generic pharmaceuticals: Pharmacy vs. physician-driven markets. *NBER Working Paper* 17226. Cambridge, MA: NBER.

Danzon, P. M., Mulcahy, A. and Towse, A. (2011b). Pharmaceutical prices in emerging markets: effects of income, competition and procurement. *NBER Working Paper* 17174. Cambridge, MA: NBER.

Danzon, P. M., Towse, A. and Mestre-Ferrandiz, J. M. (2011a). Value-based differential pricing: Efficient prices for drugs in a global context. *NBER Working Paper* w18593. Cambridge, MA: NBER.

Di Masi, J. and Grabowski, H. (2007). The cost of biopharmaceutical R&D: Is biotech different? *Managerial and Decision Economics* **28**(4–5), 285–291.

Ernst and Young. (2012) Pulse of the industry: Medical technology report. New York: Ernst and Young.

Flynn, S., Hollis, A. and Palmedo, M. (2009). An economic justification for open access to essential medicine patents in developing countries. *Journal of Law Medicine and Ethics* **37**(2), 184–208.

Garber, A., Jones, C. I. and Romer, P. M. (2006). Insurance and incentives for medical innovation. Forum for health economics and policy: Vol. 9, Issue 2 Article 4. Cambridge, MA: Biomedical Research and the Economy.

Mestre-Ferrandiz, J. M., Sussex, J. and Towse, A. (2012). *The R&D cost of a new medicine*. London: Office of Health Economics.

Further Reading

Claxton, K., Briggs, A., Buxton, M. J., et al. (2008). Value based pricing for NHS drugs: An opportunity not to be missed? *British Medical Journal* **336**, 251–254.

IMS Market Prognosis (2012). *Report of the IMS Institute of Healthcare Informatics*. Available at: www.ims.com (accessed 20.03.13).

Lakdawalla, D. and Sood, N. (2009). Innovation and the welfare effects of public drug insurance. *Journal of Public Economics* **93**, 541–548.

Malueg, D. and Schwartz, M. (1994). Parallel imports, demand dispersion, and international price discrimination. *Journal of International Economics* **37**, 167–195.

Biosimilars

H Grabowski, Duke University, Durham, NC, USA
G Long and R Mortimer, Analysis Group, Inc., Boston, MA, USA

Introduction

Although the biotech industry is a relatively new source of medical therapies – its first new drug approvals came in the early 1980s – it has recently become a major source of drug industry growth and innovation. New biological entities (NBEs) have a significantly higher likelihood of being a first-in-class or novel introduction compared with other new drug entities (Grabowski and Wang, 2006). For example, the oncology class has experienced the introduction of breakthrough monoclonal antibodies and targeted biological agents resulting from increased knowledge of the molecular mechanisms for cancer (DiMasi and Grabowski, 2007a). Substantial improvements in survival, morbidity, and patients' quality of life have been documented in diseases previously resistant to successful treatment, such as aggressive HER-2 positive breast cancer (Smith *et al.*, 2007) and disability associated with rheumatoid arthritis (Weaver, 2004).

Although NBEs have been an important source of biopharmaceutical innovation, they have also accounted for a rising share of overall drug expenditures in the US and worldwide. They now account for approximately one-quarter of all the US expenditures on pharmaceuticals and represent approximately half of all products in clinical testing (Trusheim *et al.*, 2010). NBEs for oncology patients and other indications also can cost tens of thousands of dollars per course of treatment. They are also frequently targeted to life-threatening and disabling diseases. These facts and trends have made biological entities an increasing focus of attention for policymakers and payers grappling with rising healthcare costs and budgets.

A recent development in Europe and the US is the establishment of an abbreviated pathway for the so-called biosimilars – biological products that are similar to, but not identical with, a reference biological product in terms of quality, safety, and efficacy. Biologics are typically more complex molecules than small-molecule chemical drugs. Biologics are manufactured not through chemical synthesis but through biological processes involving manipulation of genetic material and large-scale cultures of living cells, where even small changes to the manufacturing process may lead to clinically significant and unintended changes in safety and efficacy. As a result, establishing that a biosimilar is 'similar enough' to achieve comparable therapeutic effects in patients is a much more challenging task for companies and regulators than establishing bioequivalence for generic chemical drugs. Biosimilars generally require analytical studies, animal testing data, and some clinical trial evidence on safety and efficacy to gain approval. Biosimilars can provide an important new source of competition to established biological entities. A key issue at the present time is how this competition is likely to develop and how it will influence expenditures for biopharmaceuticals by payers and consumers, investment in innovation, and the research, development, and marketing processes for manufacturers.

The EU has had a framework in place for approving biosimilars since 2005. The European Medicines Agency (EMA) has issued general and class-specific guidelines in six classes and has approved biosimilars in three product classes – somatropins, erythropoietins, and granulocyte colony-stimulating factors (G-CSFs). The experience of biosimilars in various European countries is considered later in this article.

In March 2010, as part of the overall Patient Protection and Affordable Care Act, the US Congress created an abbreviated pathway to approve biosimilars. The Food and Drug Administration (FDA) is in the process of implementing the law, including consulting with potential entrants and developing and releasing for public comment draft guidelines. The US situation is of particular interest as it has been the center of biotech innovation and the country with the largest expenditures on biological products. Although the US has a strong history of generic drug utilization, until the 2010 Act, there was no corresponding pathway for biosimilar entry.

In this article, the authors first discuss regulatory, reimbursement, and economic factors that will affect how competition between branded biologics and biosimilars may evolve. These factors are based on current market dynamics including initial European biosimilar experiences, the provisions of the new US law enacted in 2010, and the US experiences under the Hatch-Waxman Act. Taking into account the scientific, manufacturing, and other differences between biologics and chemically synthesized drugs, and between the regulatory frameworks governing each, expected biosimilar competition is then compared and contrasted with generic competition. Finally, the likely impact of biosimilars on cost savings is briefly assessed and potential impacts on innovation incentives in the biopharmaceutical industry is discussed.

Biosimilar Experience in the European Union

The EU has had in place a well-defined regulatory pathway for biosimilars for several years. In October 2005, the European Commission adopted an EMA framework for the approval of biosimilars. The framework includes an overarching set of principles; general guidelines on quality, safety, and efficacy; and guidelines specific to product classes. To date, the EMA has issued guidelines in six therapeutic classes. Guidance is under development for three other major types of biologics: monoclonal antibodies, recombinant follicle-stimulating hormone, and recombinant interferon beta. Other countries have used a European-like approach, including Canada (where biosimilars are termed 'subsequent entry biologics' (SEBs)) and Japan. Australia adopted the EU guidelines in August 2008.

 doi:10.1016/B978-0-12-375678-7.01208-6

The EMA has required at least one Phase II or III clinical trial for biosimilars to demonstrate similar safety and efficacy to their reference molecules. As opposed to the legislative biosimilar framework in the US, in which the FDA approves applications as biosimilars or interchangeable biosimilars, the EMA framework does not result in any findings of interchangeability, and questions of substitution are left to the member states to regulate. Local substitution laws differ across the EU member states, with some including explicit prohibitions on automatic substitution for biologics (such as Spain and France).

Since 2006, 14 biosimilar products in three therapeutic classes – erythropoietins, somatropin, and granulocyte colony-stimulating factors (G-CSFs) – have been approved, referencing four innovative products, and 13 are currently marketed in Europe. Three applications for biosimilar human insulin (with different formulations) were withdrawn in December 2007, based on failure to demonstrate comparability, and one approved product was later withdrawn (Table 1).

Empirical Evidence from Biosimilars in the European Union

Germany has exhibited the highest level of aggregate demand and market share for any biosimilar product (erythropoietin). To date, Germany's Federal Healthcare Committee, which decides which products and services are reimbursed, has embraced biosimilars wholeheartedly. In addition to a reference pricing system in place for biosimilars, Germany has specific targets or quotas for physician and sickness funds for biosimilars that vary by region. Furthermore, Germany is a main source of biosimilar manufacturing in Europe, and biosimilar companies generally enjoy strong reputations with healthcare providers.

Uptake in other European countries has been slower. In some cases, this reflects later biosimilar entry dates and the timing of reimbursement approval by government payers. Although evidence from experiences in Germany or other European countries with biosimilar substitution are not directly applicable to other markets, given differences in the markets and pricing, access, and reimbursement systems, they nevertheless suggest that over time, payers, physicians, and patients will accept biosimilars.

Table 2 summarizes biosimilar shares in five large European countries: France, Germany, Italy, Spain, and the UK, for the therapies somatropin, erythropoietin alpha, and G-CSF from 2007 to 2009. The extent of biosimilar penetration varied substantially both across therapies within a country and across countries for the same therapy. In Germany, the biosimilar erythropoietin alpha accounted for 62% of total biosimilar and innovator product units sold in 2009, within 2 years of its launch; by contrast, in France, Italy, Spain, and the UK, biosimilar erythropoietin alpha had less than a 5% share in 2009. Biosimilar market shares for G-CSF in 2009 ranged from 21% (UK) to 7% (Spain). However, there is evidence that biosimilar G-CSF shares have grown rapidly in several European countries since 2009 (Grabowski *et al.*, 2012). In particular, a study undertaken by IMS Health found that biosimilars in the G-CSF class had shares more than 50% in Germany, France, and the UK by the third year after launch, and characterized the market for this class in these counties as being commodity-like and mainly controlled by payers (IMS, 2011a). In contrast, the shares for somatropin are lower than the other two classes in most European countries, reflecting conservative physician prescribing and a differentiated market with competition based on price, promotion, and delivery device-based patient convenience.

Biosimilar market development (and share uptake) may differ between European countries and the US, given the differences between their healthcare systems. For example, the US is more litigious than Europe; thus, the FDA may decide to proceed more cautiously and require more clinical data than the EMA has in the past. This broad generalization may not always hold true; however, in the US, the FDA approved Sandoz's enoxaparin sodium abbreviated new drug application (ANDA) as a fully substitutable generic (referencing Lovenox®) requiring no clinical evidence. In contrast, the EMA requires clinical data to approve a biosimilar application for a low molecular weight heparin. Future research comparing biosimilar market attitudes and experience in European countries, countries with a European-like approach (e.g., Australia, Japan, and Canada), the US, and other nations (e.g., the so-called 'BRIC' nations of Brazil, Russia, India, and China) is needed. Given the significant differences in the regulatory, medical delivery, and reimbursement systems between less-developed and more-developed nations, the pattern of biosimilar competition may also be very different.

The United States Biologics Price Competition and Innovation Act

The Biologics Price Competition and Innovation Act of 2009 (BPCIA), enacted as part of the Patient Protection and Affordable Care Act of 2010 (PPACA), created an abbreviated pathway for the FDA to approve biosimilars. This legislation complements the 28-year-old Drug Price Competition and Patent Term Restoration Act of 1984 (generally referred to as the Hatch-Waxman Act), which provides a clear path for generic drug entry in the case of new chemical entities (NCEs) approved under the Food, Drug, and Cosmetic Act (FD&C Act) through the ANDA process. Through that process, generic drugs demonstrated to be bioequivalent to off-patent reference drugs may be approved without the submission of clinical trial data on efficacy and safety. ANDA approval requires a finding that the generic drug is bioequivalent to its reference drug and has the same active ingredient(s), route of administration, dosage form and strength, previously approved conditions of use, and labeling (with some exceptions). Some initially marketed biologic products were approved under the FD&C Act, such as human growth hormones. However, most large molecule biologic medicines were approved under the Public Health Service Act and have not been subject to generic competition under the ANDA process of the Hatch-Waxman Act. Biologic medicines approved under the Public Health Service Act will now be subject to competition from products coming to market through an expedited biosimilar approval process – relying at least in part on the innovator's package of

Table1 European biosimilar regulatory reviews and current marketing status

Trade name	*Active substance*	*Biosimilar sponsor*	*Reference product*	*Therapeutic area*	*Biosimilar decision and date*	
Omnitrope	Somatropin	Sandoz	Genotropin	Turner syndrome, pituitary dwarfism, and Prader–Willi syndrome	Approve	12 April 2006
Valtropin	Somatropin	BioPartners	Humatrope	Turner syndrome and pituitary dwarfism	Approve	24 April 2006
Alpheon	Interferon alpha-2a	BioPartners	Roferon-a		Reject	28 June 2006
Abseamed	Epoetin alpha	Medice	Eprex	Chronic kidney failure, anemia, and cancer	Approve	28 August 2007
Binocrit	Epoetin alpha	Sandoz	Eprex	Chronic kidney failure and anemia	Approve	28 August 2007
Epoetin alfa Hexal	Epoetin alpha	Hexal	Eprex	Chronic kidney failure, anemia, and cancer	Approve	28 August 2007
Retacrit	Epoetin zeta	Hospira	Eprex	Anemia, autologous blood transfusion, cancer, and chronic kidney failure	Approve	18 December 2007
Silapo	Epoetin zeta	STADA	Eprex	Anemia, autologous blood transfusion, cancer, and chronic kidney failure	Approve	18 December 2007
Insulin Rapid Marvel	Insulin	Marvel	Humulin		Withdrawn	16 January 2008
Insulin Long Marvel	Insulin	Marvel	Humulin		Withdrawn	16 January 2008
Insulin 70/30 Mix Marvel	Insulin	Marvel	Humulin		Withdrawn	16 January 2008
Biograstim	GCSF	CT Arzneimittel	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	15 September 2008
Filgrastim ratiopharm	GCSF	Ratiopharm	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve Withdrawn	15 September 2008 20 July 2011
Ratiograstim	GCSF	Ratiopharm	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	15 September 2008
Tevagrastim	GCSF	Teva	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	15 September 2008
Filgrastim Hexal	GCSF	Hexal	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	6 February 2009
Filgrastim Zarzio	GCSF	Sandoz	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	6 February 2009
Nivestim	GCSF	Hospira	Neupogen	Cancer, hematopoietic stem cell transplantation, and neutropenia	Approve	8 June 2010

Abbreviation: GCSF, granulocyte colony-stimulating factor.

Table 2 Initial biosimilar competition in selected EU countries: Market share evidence

	Biosimilar unit share of the molecular entity				
	France	*Germany*	*Italy*	*Spain*	*UK*
Somatropin					
2007	2%	3%	6%	1%	0%
2008	10%	6%	17%	1%	0%
2009	16%	8%	27%	5%	1%
Erythropoietin alpha					
2007	0%	0%	0%	0%	0%
2008	0%	35%	0%	0%	0%
2009	4%	62%	0%	4%	1%
GCSF					
2007	–	–	–	–	–
2008	0%	1%	0%	0%	2%
2009	7%	17%	N/A	9%	21%

Note: Biosimilar share of unit sales are measured based on Defined Daily Dose. Biosimilar G-CSF was not launched until 2008, so biosimilar shares for 2007 are not reported in Table 3. For G-CSF in Italy in 2009, the biosimilar share is recorded as N/A to reflect insufficient data for calculating a biosimilar share – fewer than 5000 DDDs were reported in the data for combined innovator and biosimilar unit sales in Italy that year.

data or a prior FDA approval – for the first time as a result of the BPCIA.

The key provisions of the new legislation establishing an abbreviated pathway for the FDA to approve a biosimilar are:

- Biosimilarity: A biosimilar does not have to be chemically identical to its reference product but must be "highly similar to the reference product notwithstanding minor differences in clinically inactive components" and there must be "no clinically meaningful differences, in terms of safety, purity, and potency." (PPACA, Section 7002 (b)(3))
- Interchangeability: The FDA may deem a biosimilar interchangeable with its reference product if it can be shown that it "can be expected to produce the same clinical result as the reference product in any given patient" and that "the risk in terms of safety or diminished efficacy of alternating or switching between use of the biological product and the reference product is not greater than the risk of using the reference product without such alternation or switch." (PPACA, Section 7002 (k)(4)) The first biosimilar shown to be interchangeable is entitled to a 1-year exclusivity period during which no other product may be deemed interchangeable with the same reference product.
- Regulatory review: The FDA will determine whether a product is biosimilar to a reference product based on stepwise consideration of analytical, animal-based, and clinical studies (including the assessment of immunogenicity and pharmacokinetics or pharmacodynamics). In February 2012, the FDA released the first three documents in a set of guidance documents for the development of biosimilars under BPCIA.
- Regulatory Exclusivity for the innovative biologic: Biosimilar applications may be submitted beginning 4 years after FDA approval of the reference innovative product. Before the FDA can approve a biosimilar using the abbreviated pathway, there is a 12-year period of exclusivity following FDA approval of the innovative biologic. An additional 6 months of exclusivity is available for the reference innovative biologic if pediatric-study requirements are met, which applies to both the 4- and 12-year exclusivity periods. The most important (and contentious) of these exclusivity provisions is the 12 years of exclusivity for an innovative biologic before a biosimilar can enter using an abbreviated application. This 12-year exclusivity term is referred to as regulatory exclusivity in distinction from the exclusivity afforded through patents granted by the US Patent and Trademark Office.
- Limitations on 12-year exclusivity: Several types of licensures or approvals are not eligible for 12-year exclusivity, including: (1) a supplemental biologics license application (sBLA) for the reference biologic; (2) a subsequent BLA filed by the same sponsor, manufacturer, or other related entity as the reference biologic product that does not include structural changes in a biologic's formulation (e.g., a new indication, route of administration, dosing schedule, dosage form, delivery system, delivery device, or strength); or (3) a subsequent BLA filed by the same sponsor, manufacturer, or other related entity as the reference biologic product and that includes structural changes in a biologic's formulation but does not result in improved safety, purity, or potency.
- Reimbursement: A potential disincentive for biosimilar adoption is mitigated by setting the reimbursement for a biosimilar under Medicare Part B at the sum of its Average Selling Price (ASP) and 6% of the ASP of the reference biologic.
- Patent provisions: The BPCIA requires a series of potentially complex private information exchanges between the biosimilar applicant and reference product sponsor, followed by negotiations and litigation, if necessary. In contrast to the patent provisions for NCEs under the Hatch-Waxman Act, there is no public patent listing akin to the Orange Book, no 30-month stay when a patent infringement suit is brought, and no 180-day exclusivity awarded to the first firm to file an abbreviated application and achieve a successful Paragraph IV patent challenge.

Food and Drug Administration Regulations and the Costs of Developing a Biosimilar

The new law authorizing biosimilars gives broad latitude to the FDA to define the process and standards for approval. FDA decisions will affect both the demand for and the supply of biosimilars:

- The level of evidence required will affect the costs of market entry, the number of biosimilar entrants, and the assets and capabilities required to compete successfully.
- The level of clinical trials and other evidence required to establish interchangeability or similarity will also potentially affect the level of market adoption, as greater levels of evidence may increase physicians', payers', and patients' confidence in a biosimilar medicine.

- Naming conventions and pharmacovigilance requirements for biosimilars will affect market entry and perceptions of substitutability by physicians, payers, and patients, as well as safety monitoring after launch.
- Whether data on one indication can be extrapolated to others – absent additional clinical trials in that patient population will have an impact on entry decisions, perceptions of substitutability, and biosimilar market uptake.
- Definitions of what constitutes changes in 'safety, purity, or potency,' as they are applied to determine whether a 12-year exclusivity is to be authorized for next-generation products will affect biotech investor incentives.

Criteria for Establishing Biosimilarity

The initial draft guidance documents released by the FDA in February 2012 state that "FDA intends to consider the totality of the evidence provided by a sponsor to support a demonstration of biosimilarity" (emphasis added). For a given biosimilar application, the FDA draft guidance notes that "(t)he scope and magnitude of clinical studies will depend on the extent of residual uncertainty about the biosimilarity of the two products after conducting structural and functional characterizations and possible animal studies." (Food and Drug Administration (FDA), 2012a, pp. 2, 12). Theoretically, this could encompass, at one extreme, only a bioequivalence study (similar to what is required for generic approval under Hatch-Waxman) or, at the other extreme, when science and experience require more data, a full program of clinical studies equivalent to that included in a biologic's license application.

FDA officials, in a New England Journal of Medicine publication, had previously stated that "[a]lthough additional animal and clinical studies will generally be needed for protein biosimilars for the foreseeable future, the scope and extent of such studies may be reduced further if more extensive fingerprint-like characterization is used." (Kozlowski *et al.*, 2011, p. 386) In the future, the agency hypothesizes, the current state-of-the-art for analytic characterizations may advance to allow highly sensitive evaluations of relevant product attributes and permit a 'fingerprint-like' identification of very similar patterns in two different products (such strategies were cited in the FDA's approval of the Sandoz ANDA for enoxaparin sodium, a complex mixture, mentioned later in this article.)

The costs of an FDA submission for the US approval could be lower for biosimilars already on the market in Europe if the biosimilar can rely on previously undertaken European clinical trials, at least for some products. In its draft guidance documents released in February 2012, the FDA noted it will accept clinical studies undertaken for approval in other jurisdictions under certain circumstances, when justified scientifically and when accompanied by 'bridging' data. However, it also noted, "[a]t this time, as a scientific matter, it is unlikely that clinical comparisons with a non-US-licensed product would be an adequate basis to support the additional criteria required for a determination of interchangeability with the US-licensed reference product," and the specific data requirements for products will be determined by the FDA on a case-by-case basis. (Food and Drug Administration (FDA), 2012b, p. 8.)

If the FDA requires significant clinical trial evidence, approvals for biosimilars, as compared with generics, will require a much bigger investment. The cost for biosimilar approval will depend on the number and size of the necessary clinical trials, the number of indications involved, and other specific FDA requirements. The current requirement for a BLA is typically two large-scale Phase III pivotal trials. If the FDA requires at least one Phase II/III type study comparable to those undertaken by innovators, then the out-of-pocket costs will likely be in the range of US$20 million to US$40 million for the studies alone. In addition, the preclinical costs associated with biosimilars may in some cases be higher for biosimilars than for innovative products, as they entail modifying the production process to achieve a specific profile that very closely approximates the reference product without the benefit of the innovator's experience. Others have estimated that for very complex biologics such as some monoclonal antibodies, biosimilar development costs could total US$100 million to US$200 million and take 8 or more years to bring a product to market (Kambhammettu, 2008). In contrast, the cost of completing bioequivalence studies for generic drugs is estimated to be only US$1 million to US$2 million.

Regulatory Requirements for an Interchangeability Designation

Another key regulatory issue will be the analytical and clinical evidence required to deem a biosimilar interchangeable with its reference product, thus enabling automatic substitution without physician approval, subject to relevant state laws. Under the BPCIA, for products used more than once by patients (the majority of biologics), the biosimilar sponsor will need to demonstrate that switching between the biosimilar and reference product poses no additional risk of reduced safety or efficacy beyond that posed by the reference product alone. Postapproval interchangeability assessments may require a strong postmarketing system and evaluation of postmarketing data.

Achieving an FDA finding of interchangeability may be associated with far greater development costs than achieving a determination of biosimilarity, so it may be limited initially to a select few examples where molecules meet certain tests for establishing 'sameness' through differentiated characterization or other available technology. For instance, the availability of differentiated analytical characterization technology supported the FDA's approval of Sandoz's ANDA for generic enoxaparin sodium (referencing Lovenox®). Although not a biosimilar (Lovenox®, a chemically synthesized product derived from natural sources, has been described as a complex mixture), the factors that the FDA cited in its approval may give some insight into the Agency's current approach and how continued technological change could influence the evidence necessary to establish interchangeability in the future.

For classes of more complex biologics, applications for biosimilarity will likely require some clinical trial data in order to be approved and costly switching trial data in order to be deemed interchangeable. Many firms may elect not to make the investments necessary to pursue interchangeability initially, given the current state of scientific knowledge regarding

biosimilars and high levels of regulatory uncertainty. This is in contrast to small-molecule generic drugs, where an 'A' rating by the FDA recognizes the products as therapeutically equivalent and eligible for substitution by pharmacists without physician approval, subject to state substitution laws, thus driving rapid share loss by the branded reference product.

Manufacturing Costs

The ongoing cost of manufacturing biological entities is also significantly higher than for chemical entities. Biosimilar manufacturers may need to construct expensive plants or obtain long-term lease or purchase agreements with third parties that have an FDA-approved facility if they do not already have excess suitable manufacturing capacity. In any event, the cost of entry for biosimilars in terms of plant capacity is likely to be an order of magnitude higher than for generic drug products (which may total only US$1 to US$2 million) and may be closer to two orders of magnitude higher. The high costs of entry – particularly the substantial capital requirements – are likely to restrict the number and types of biosimilar entrants, at least initially. Furthermore, initial entry is likely to be limited to the biologics with the largest revenues and those where scientific and market feasibility have been demonstrated in Europe.

The Perspectives of Healthcare Payers, Providers, and Patients

Reimbursement and Payer Considerations

Payer reimbursement policies and access control mechanisms also can substantially affect the extent and speed of biosimilar uptake. Consistent with relevant local laws, regulations, and practices, payers will develop coverage and reimbursement policies and make individual pricing, reimbursement, and access decisions for biosimilars and their branded reference products.

Cost sensitivity and willingness to encourage the use of biosimilars in place of their reference therapies may vary across different payers, including private insurers and public payers. Payer controls that restrict patient and physician therapy choice and access may also vary according to the setting in which care occurs (e.g., inpatient hospital or physician office), whether the biosimilar is rated interchangeable, the therapeutic indication and disease severity (e.g., oncology or growth disorders), as well as other factors.

Private insurers

Historically, in the US, managed care plans have been reluctant to restrict access or pursue aggressive cost-control measures because many biologic therapies are: (1) targeted to life-threatening illnesses such as cancer or other diseases that involve serious disability and (2) often lack close substitutes. In addition, biologics that are dispensed by physicians are often managed within plans as medical benefits rather than pharmacy benefits and are typically less subject to centralized controls or formulary restrictions. This has been changing over the past several years, particularly in indications where there is a choice between multiple brand name biologics. The introduction of biosimilars can be expected to accelerate these trends toward more active management of biologic choice, costs, and utilization.

Medicare

Medicare reimburses biologics under either the Part B or the Part D program, depending on the mode of administration. Many biologic drugs are currently dispensed in a physician's office, clinic, or hospital as infused agents. The use of these biologics for Medicare patients is covered under the Medicare Part B program, whereas self-injectable biologics dispensed in pharmacies (including by specialty pharmacy or mail-order programs) are covered by the Part D program.

Medicare Part B

In designing the new abbreviated pathway for biosimilars, Congress acknowledged that the Medicare rules for reimbursement of drugs administered under Part B could provide inadequate financial incentives for providers to utilize lower priced biosimilars. Part B drugs have historically been purchased through a 'buy and bill' approach by providers who also make decisions about which therapies are appropriate for a given patient. The provider is reimbursed by Medicare for administering a Part B drug, and the level of reimbursement is based on the manufacturer's weighted ASP for the category to which the drug belongs (defined by a unique code), plus 6%. When generics are assigned to the same code as their reference new chemical entity, physicians receive the same level of reimbursement, the volume-weighted average ASP for all manufacturers' products, for using either the generic or the reference product. Thus, physicians generally have a strong incentive to utilize the lower cost generic product, (although the physician's choice of generic or reference product also depends on the net acquisition cost of both products to the physician, based on any contracts that may be in place with the brand manufacturer as well as the pricing strategy of the generic entrant).

Because biosimilars are unlikely to be deemed interchangeable by the FDA, at least initially, to the degree they are thus unlikely to be assigned to the same code as the brand product, physicians may have an incentive to utilize the more expensive (higher ASP) reference product for patients, as reimbursement is based on ASP plus 6%. To mitigate potential financial disincentives for physicians to adopt biosimilars, the new legislation sets biosimilar reimbursement under Medicare Part B at the sum of the biosimilar's ASP and 6% of the ASP of the reference biologic product. The reference biologic product will continue to be reimbursed at its own ASP plus 6%. By basing the 6% payment to providers on the reference brand's ASP, the legislation seeks to mitigate provider disincentives to adopt lower cost biosimilars when they are not deemed to be interchangeable and are placed in separate codes. Whether this reimbursement provision will be sufficient to overcome physician experience and loyalty to the reference biologic, as well as other financial incentives, is an open question.

Medicare Part D

Privately offered Medicare Part D drug programs cover drugs available at retail or via mail order, including self-injectable biologics. Biologics accounted for only 6% of total

prescription drug costs in the Medicare Part D program in 2007 (Sokolovsky and Miller, 2009); however, spending for biologics within the Part D program is expected to increase rapidly in the coming years. Between 2007 and 2008, MedPac estimates indicate that prices paid for drugs on specialty tiers (including biologics) in the Part D program grew by 18%, compared with 9% for all Part D drugs. Expenditures for self-injected biologics are expected to continue to grow rapidly as these agents are increasingly used to treat a range of diseases, from rheumatoid arthritis to multiple sclerosis to human growth deficiency, and a large number of new biologics are currently under development. The high price of self-injected biologics relative to traditional NCEs also suggests that biologics will comprise an increasing share of Part D expenditures. This shift may lead payers to pursue pharmacy management techniques aimed at controlling utilization of these biologics.

Many Medicare Part D plan designs include a specialty drug tier, with median coinsurance rates increasing from 25% in 2006 to 30% in 2010 for stand-alone prescription drug plans and to 33% in 2010 for drug plans offered as a part of Medicare Advantage (Hargrave *et al.*, 2010). Coinsurance plan designs could produce strong incentives to utilize biosimilars if substantial discounts emerge for biologic products with expensive courses of treatment for patients. Preferred specialty drugs might be subject to lower rates of coinsurance, to a copayment rather than to coinsurance, or to lower patient out-of-pocket costs at the same coinsurance rate.

One limiting factor to formulary incentives for biologics in Medicare Part D is that enrollees with low-income subsidies make up a disproportionately large share of the market for biologics under the Part D program. Given that these individuals are subject to limited cost sharing, other instruments such as step therapy and prior authorization may be employed to provide incentives for the use of biosimilars.

Medicaid

Medicaid Preferred Drug Lists (PDLs) reflect preferred biologic products in a number of therapeutic categories. Preferred drugs can be dispensed without the access controls (e.g., prior authorization) applied to nonpreferred drugs. For example, online PDLs for Florida, Illinois, New York, Ohio, Pennsylvania, and Texas indicate that rheumatoid arthritis (RA), hepatitis C (HCV), and human growth hormone formularies in these six large states preferred two or three RA agents (of six), one or two HCV agents (of five), and between two and five human growth hormones (of nine agents/forms). Medicaid programs can be expected to encourage biosimilars through PDLs and other medical management instruments. States with managed Medicaid programs apply formulary and access management techniques common in commercial insurance plans, and such managed programs are becoming more common.

Hospitals

Hospitals typically bear the costs of all drugs, including biologics, used during inpatient hospital stays as part of a fixed diagnosis-related group-based reimbursement per admission (DRG) that includes all services and products used during the episode of care. Consequently, these hospitals have incentives to implement formularies of preferred drugs and other mechanisms that encourage the use of lower priced products, possibly including biosimilars. As a result, for biologics that are generally used in hospital settings, hospitals will play a larger role than insurance companies in determining the demand for biosimilars. In the hospital sector, Pharmacy and Therapeutics (P&T) committees review the drugs that are stocked, on standing order forms, and which can be used by physicians. Hospitals also rely on Group Purchasing Organizations (GPOs) to gain leverage in negotiating discounts from suppliers, including biologic manufacturers. Because the hospital GPO market is highly concentrated, favorable contracts with a handful of suppliers can affect product selection. In addition, fixed reimbursement creates strong incentives for input cost reductions. To the degree that biologics used in the inpatient hospital setting are included in the DRG, depending on how significant a portion of spending they represent, hospitals may be more aggressive in implementing access controls to favor the utilization of some biosimilars, if biosimilar prices are not countered by originator manufacturer discounts.

United States healthcare reform initiatives

More widespread adoption of comparative- and cost-effectiveness analyses across the US healthcare system could also influence adoption of biosimilars. Formal cost-effectiveness reviews by payers have been well established in countries outside the US in the form of Health Technology Assessments (HTAs). In the UK, for example, the National Institute for Health and Clinical Excellence's (NICE) coverage recommendations have been based on strict reviews of cost-effectiveness calculations relative to current treatment, with an implied threshold value of an acceptable incremental cost per quality-adjusted life-year (QALY).

Finally, long-term changes in reimbursement policies may also shift financial incentives toward the use of biosimilars. For example, the adoption of global payment strategies, rather than fee-for-service reimbursement, or some form of shared savings, could strengthen the link between physician and/or hospital compensation and the use of lower priced biologics. Global payment strategies provide incentives for the adoption of lower cost treatments (and potentially encourage greater price competition) by setting a fixed payment level for a patient/episode of care, with all or some portion of the cost savings accruing to the care providers. Several states are considering implementing global payment strategies, and it has been suggested that government programs such as Medicaid could be the first to implement these strategies.

Patient and Physician Perspectives

The rate of biosimilar penetration is expected to vary by disease indication, patient type, physician specialty, and other factors. As noted, rates of patient and physician acceptance of biosimilars are expected to be lower when the biosimilar lacks an interchangeability rating. In addition, rates of biosimilar acceptance may vary according to such physician and patient-focused factors as: Whether the physician specialty is historically more price-sensitive or demonstrates greater levels of brand loyalty in therapy choice (for instance, allergists vs. rheumatologists); whether the biosimilars will be used long-term

as maintenance therapy or only once or twice (particularly if long-term clinical data are not available); whether the indication is life threatening or the implications of therapeutic nonresponse or adverse reactions are perceived to be very serious; or whether the difference in ease-of-use or out-of-pocket cost to the patient of the brand instead of the biosimilar is expected to be high.

When patients are stable on a given maintenance therapy, biosimilar substitution may tend to be concentrated among new patient starts. (The same is true of 'switches' between one branded drug and another.) As a result, the penetration of biosimilars for indications with a low rate of turnover in the patient populations may be limited if products are not interchangeable. The degree of biosimilar uptake will also depend on cost differences and the financial incentives to utilize biosimilars employed by managed care and government payers. These incentives, however, are likely to be tempered if existing patients are responding well to an established therapy. Other factors such as specialists' brand loyalty, clinically vulnerable patient populations, and physician conservatism in switching stable patients to new therapies are also likely to keep rates of biosimilar uptake for current patients below those for new patients.

Another important demand-side factor is the perspective of specialist physicians and patient groups concerning biosimilars. Physicians who have years of experience with the reference biologic may be reluctant to substitute a biosimilar even for new patients until sufficient experience has been accumulated in clinical practice settings, as opposed to in clinical trials. To stimulate demand, it may be necessary for biosimilar firms to establish 'reputation bonds' with physicians through strategies similar to those employed by branded firms that communicate information to establish brand value through physician detailing, publications, advertising, and education programs. In addition, patient assistance programs and contracts with health plans, pharmacy benefit managers (PBMs), hospitals, or provider groups, which exercise control over therapy choice, may be used in a targeted way to affect the economic proposition associated with biosimilar adoption. These measures will increase the cost of drug distribution and marketing for biosimilars compared with small-molecule generic drugs, where such marketing and sales costs are minimal and demand is purely driven by lower price and pharmacy contracts for availability.

Biosimilar Competition versus Generic Competition

Since the passage of Hatch-Waxman 28 years ago, generic entry has become a principal instrument of competition in the US pharmaceutical market. Generic products in 2010 accounted for 78% of all the US retail prescriptions, (IMS, 2011b) compared with only 19% in 1984 (Federal Trade Commission, 2002). As discussed, the growth of generic utilization has been accelerated by various formulary and utilization management techniques such as tiered formularies, prior authorization and step therapy requirements, higher reimbursements to pharmacies for dispensing generics, and maximum allowable cost (MAC) programs.

A distinctive pattern of generic competition has been observed in numerous economic studies (Grabowski, 2007). There is a strong positive relationship both between a product's market sales and the likelihood of a patent challenge and between the number of generic entrants and the intensity of generic price competition once the exclusivity period has expired. An increasing number of products are now subject to patent challenges earlier in their product life cycle, as generic firms seek out the 180-day exclusivity period awarded to the first firm to file an ANDA with a successful Paragraph IV challenge. Successful products typically experience multiple entrants within the first several months after patent expiration, and generic price levels drop toward marginal costs rapidly as generic entry increases.

Theoretical Models of Biosimilar Competition

Given the much higher costs of entry for biosimilars compared with generic drugs, as well as the other demand- and supply-side factors discussed in the section Food and Drug Administration Regulations and the Costs of Developing a Biosimilar, the pattern of biosimilar competition is expected to differ from current generic competition. In particular, fewer entrants and less intensive price discounting are expected and competition may resemble branded competition more than generic competition (Grabowski *et al.*, 2006). This is currently the case in the human growth hormone market, where eight products compete both through price, patient support, and product delivery differentiation. In 2006, Sandoz entered the human growth hormone market with Omnitrope® (which referenced Pfizer's Genotropin®, via the section 505(b)(2) pathway of the Hatch-Waxman Act). Omnitrope® has struggled to gain market share. Initially, it was reported to have priced at a 30% discount based on wholesale acquisition cost (WAC) compared with the most widely used biologic in this class, Genotropin®. By 2008, Omnitrope®'s discount had increased to 40% (Heldman, 2008). Despite these discounts, its share of somatropin use remained below 5%. These outcomes may not be reflective of the pattern of substitution for biosimilars generally, given that the human growth hormone market was a mature one with a number of competitors, and also given the differentiation by established brands via sophisticated pen- or needle-free delivery systems in this product class. With the approval of a pen delivery device system, and a strategy that includes physician detailing and patient support services, Omnitrope®'s share of prescriptions dispensed increased to 19% in September 2012.

To date, some theoretical analyses have attempted to model the likely scenarios for biosimilar competition in the US market. One paper implements a simulation approach and projects that the relatively high cost of biosimilar entry will result in relatively small number of entrants even for larger selling biologic products and more modest discounts on biosimilars than in the case of generics (Grabowski *et al.*, 2007). Other research relies on a segmented model of biosimilar competition, where biosimilars would be utilized significantly in price-sensitive segments of the market but less so in the nonprice-sensitive segments (given the reluctance of many providers to utilize biosimilars until considerable

Table 3 Biosimilar competition US market share and price discount economic analyses

Source	*Peak biosimilar penetration*	*Biosimilar discount to preentry brand price*	*Basis*
Grabowski (2007)	10–45%	10–30% (year 1)	Higher estimates correspond to complex small molecules
Congressional Budget Office (CBO) (2008)	10% (year 1) 35% (year 4)	20% (year 1) 40% (year 4)	Similar market situations
Express Scripts (2007)	49%	25% (year 1)	Therapeutic alternatives
Avalere Health (2007)	60%	20% (year 1) 51% (year 3)	Average small-molecule generic drug penetration rates

clinical experience has accumulated) (Chauhan *et al.*, 2008). In this model, average price discounts depend on the relative size of these market segments. The findings indicate that, given a relatively small number of branded biosimilar competitors, the innovator will discount prices from preentry levels but not as much as the biosimilar entrants. This is in contrast to generic competition where branded firms typically do not lower prices postentry but may license an authorized generic when only a small number of generic competitors are expected as a result of a successful paragraph IV entry with a 180-day exclusivity award (Berndt *et al.*, 2007).

Empirical Studies of Generic Drug Analogs

Another line of research attempts to predict how biosimilar competition will emerge by considering analogous situations, including the US generic market for certain products which share some characteristics suggestive of biologics. In one example of this research, small-molecule drugs are divided into two classes, noncomplex and complex, with complex drugs being those that meet two of the following criteria: black box warnings, narrow therapeutic index, prescribed by specialists, oncology products, or manufacturing technology that is available to only a limited number of firms (Grabowski *et al.*, 2011a). Price and quantity data from IMS Health Inc. were analyzed for 35 conventional (nonbiologic) drugs that experienced generic entry between 1997 and 2003, and those drugs classified as complex were found to have significantly lower levels of generic share and price discounts. Furthermore, complex drugs faced only 2.5 generic entrants 1 year following initial generic entry, whereas noncomplex drugs faced an average of 8.5 generic entrants.

Although data from conventional small-molecule generics should not be directly applied to estimate biosimilar shares following market entry, they suggest that uptake rates for biosimilars may be likely to be significantly lower than those for generics, at least initially. Furthermore, these more complex generic drugs are rated therapeutically equivalent (that is, they have an FDA rating of A) and, therefore, benefit from some automatic substitution. To avoid substitution, physicians need to specify in 'do not substitute' orders that prescriptions are to be dispensed as written. At least initially, most biosimilars will not be rated therapeutically equivalent and, therefore, will not be subject to automatic substitution.

Table 3 summarizes other market share and price discount analyses generally based on selective aspects of the US generic market. Most notably, as part of the evaluation of the proposed legislation regarding biosimilars, the Congressional Budget Office (CBO) predicted a penetration rate of 35% with price discounts by biosimilars of 40%. Other estimates of market penetration from a pharmacy benefit management firm, Express Scripts, as well as by Avalere Health, a consulting firm, tend to be somewhat higher than either the Grabowski (2007) or the CBO values, with penetration in the 50–60% range, and somewhat higher discounts in the case of the Avalere study (50% by year 3).

The FDA approval of generic enoxaparin sodium, rated as therapeutically equivalent (having an A-rating) to branded Lovenox®, provides important data about competitive pricing strategy and market acceptance of generics for a complex, 'biologic-like' product. Other notable attributes of Lovenox® include large expenditures by payers (pregeneric entry sales of more than US$2 billion) and a complicated manufacturing process. Currently, the FDA has approved generic enoxaparin applications from two third-party manufacturers, Sandoz (partnered with Momenta) and Amphastar (partnered with Watson), although the latter is the subject of patent litigation. In addition, there had been for a time an 'authorized generic' supplied by Sanofi, the branded manufacturer of Lovenox®. Sales of generic enoxaparin have been robust and there has been rapid erosion of Lovenox®'s revenues and market share.

Projected Savings to United States Consumers

The CBO estimated that the provisions in the current health care law establishing a biosimilar pathway would reduce federal budget deficits by US$7 billion over the 2010–2019 period. This finding is consistent with a 2008 CBO study of a similar Senate bill, which estimated a reduction in federal budget deficits of US$6.6 billion and a reduction in biologic drug spending of US$25 billion for the 2009–18 period. Over the full 10-year period, the US$25 billion in reduced biologic drug spending would represent roughly 0.5% of national spending on prescription drugs, valued at wholesale prices. The bulk of these estimated savings accrue in the last 5 years of the 10-year time ranges analyzed. Savings beyond the 10-year period may increase substantially as more biologics lose patent and 12-year exclusivity protections and as scientific advances reduce the cost of developing and producing biosimilars.

A number of the largest-selling biologic products may face losses of some key patent or 12-year exclusivity protections in the coming years. Determining the effective patent-expiry date

for any given biologic is fraught with uncertainty because of unknowns such as which patents comprise the portfolio protecting an individual biologic, of which there may be many; the strength of those patents in the face of challenges; and the ability of biosimilar manufacturers to work around existing patents. In November 2011, for example, Amgen announced that it had been issued a patent for the fusion protein etanercept (Enbrel®) that could block biosimilar competition until 2028 (the term is 17 years from the date of award, rather than 20 years from the date of application, due to the date of the patent application). Previously, many public sources had anticipated biosimilar entry exposure for Enbrel® as early as 2012. Based on a review of patent-expiry information disclosed in manufacturers' financial reports and supplemented with additional public information from academic literature, research reports, patent filings, and court documents, the earliest publicly reported potential patent-expiry dates for a set of top-selling biologics occur in a timeframe between 2013 and 2018. These biologics include Epogen®/Procrit®, Neulasta®, Remicade®, Rituxan®, and Humira® (all products having multibillion dollar US sales in 2011). The date when these biologics may actually experience biosimilar market entry under BPCIA depends on many technical, market, regulatory, and legal factors, such as whether entry will be at risk, and the outcome of the patent litigation that is likely to ensue.

The extent of biosimilar cost savings will depend on the timing and number of biosimilar entrants, their market share and price discounts relative to the originator's product, and the potential competition from the introduction of 'biobetters' or next generation products in particular product classes. There is likely to be considerable variation in how competition evolves across biological products reflecting molecule complexity, regulatory criteria, the originating firm's patent estates, patient populations and physician specialties, as well as changing reimbursement systems and procedures. In contrast to small-molecule generic competition, there is unlikely to be a 'one-size-fits-all' pattern for biosimilar competition for the foreseeable future.

Innovation Incentives

As it did with Hatch-Waxman, Congress has attempted to balance the objectives of achieving cost savings from an abbreviated pathway for biosimilars with preserving innovation incentives for new biologics. As discussed earlier, NBEs have been an important source of novel and therapeutically significant medicines. Major advances have occurred for several oncology indications, multiple sclerosis, rheumatoid arthritis, and other life-threatening and disabling illnesses. BPCIA differs from Hatch-Waxman in the term of the data exclusivity period for innovators: BPCIA establishes 12-years data exclusivity period for innovative biologics, whereas Hatch-Waxman establishes a 5-year exclusivity period for NCEs. (The FDA cannot approve an abbreviated application relying on the innovator's data until these exclusivity periods expire.) Furthermore, as mentioned earlier, the private information exchange process for resolving patent disputes is very different for biologics under the BPCIA than the 'Orange Book' public disclosure and Paragraph IV challenge framework for NCEs under Hatch-Waxman.

Regulatory Exclusivity and Patent Protection

The process of discovering and developing a new biologic is a long, costly, and risky venture. DiMasi and Grabowski have estimated that the cost to develop a new FDA-approved biopharmaceutical is US$1.2 billion in risk-adjusted costs, capitalized to 2005 dollars using an 11.5% discount rate (DiMasi and Grabowski, 2007b). DiMasi and Grabowski found that NBEs cost more in the discovery phase, take longer to develop, and require greater capital investment in manufacturing plants than NCEs. They found that the probability of success is higher for biologics than for NCEs, but biologics that fail do so later in the Research and Development (R&D) life cycle. After adjustment for inflation and the different time periods studied, the cost of developing an NBE and an NCE are roughly comparable in value.

Intellectual property protection in the form of patents and regulatory exclusivity are the primary policy instruments by which governments encourages risky investment in R&D for new medicines (together with any tax subsidies or direct financial investment programs that may apply). Regulatory exclusivity and patents have separate but complementary roles. The US government awards patents for inventions based on well-known criteria: novelty, utility, and nonobviousness. A regulatory exclusivity period, however, is needed because after invention a long, risky, and costly R&D process remains for the development of new medicines. Effective patent life is often uncertain because significant patent time elapses before FDA approval and because there is uncertainty associated with the resolution of any patent challenges. As a result, regulatory exclusivity provides a more predictable period of protection. It essentially acts as an 'insurance policy' in instances where patents are narrow, uncertain, or near expiry.

The protection afforded by regulatory exclusivity may be particularly important for innovation incentives in biologics to the degree that patents in biologics are narrower in scope than those for small-molecule drugs and more likely to be successfully challenged or circumvented. This may be true to the degree that biologics rely more on process patents, for instance. Given that a biosimilar will be slightly different in its composition and/or manufacturing process, a court may determine that it does not infringe the innovator's patent. This has the potential to lead to a seemingly contradictory outcome where a biosimilar may be 'different enough' not to infringe the innovator's patents but still 'similar enough' to qualify for approval through an abbreviated approval pathway.

As discussed, the BPCIA grants 12 years of exclusivity for innovative biologics during which the FDA may not approve biosimilars referencing them, compared with 5 years of exclusivity for NCEs under the Hatch-Waxman Act, during which an abbrevaited application referencing them cannot be submitted (plus a stay on generic entry for up to 30 months when there is a patent challenge to allow for resolution of litigation). In contrast, the EU has harmonized across member states an '8 + 2 + 1'approach for both NCEs and NBEs (consisting of 8 years of data exclusivity, during which generic competitors may not reference the innovator's data in their applications; 2 years of market exclusivity during which generic marketing authorizations cannot be approved; and a potential additional 1 year of protection for new indications that

demonstrate significant clinical benefits over existing therapies that are approved within the first 8 years after the original molecule's approval).

Economic Analyses of the 12-Year Exclusivity Period

The US 12-year exclusivity period for innovative biologics was the focus of substantial debate by legislators. The 111th Congress considered bills with exclusivity periods ranging from 5 to 14 years. To provide economic analysis to the legislators, Grabowski (2008) developed a breakeven financial analysis using historical data on R&D costs and revenues for new biologics and the risk-adjusted market return on investment in the industry. Under this model, a representative portfolio of biologic candidates would be expected to 'break even' (or recover the average costs of development, manufacturing, promotion, and the industry's cost of capital) between 12.9 and 16.2 years after launch.

A recently published Monte Carlo simulation model examines the interaction between regulatory exclusivity terms and patent protection periods under different scenarios to highlight the circumstances where each is important in maintaining innovation incentives (Grabowski *et al.*, 2011c). The results of this analysis are generally consistent with Congress' determination that a regulatory exclusivity period of 12 years appropriately balances objectives for potential cost savings from biosimilar price competition with long-run incentives for investment in innovative biologics. This study finds that when biologic patents are relatively less certain and expected to have shorter effective lifetimes, an exclusivity period of 12 years greatly enhances investment incentives. However, if biologic patents provide relatively strong protection with significant effective patent life remaining at approval, patents alone will be sufficient to maintain investment incentives in most cases. In those instances, however, the 12-year exclusivity period has only a minimal effect on the timing of potential biosimilar entry and consequently on healthcare costs.

It remains unclear whether the longer exclusivity periods for biologics compared with chemical entities will tilt R&D incentives toward large molecules and whether Congress will consider harmonizing these periods, as is currently the case in the EU.

The Resolution of Patent Challenges

Hatch-Waxman also featured Paragraph IV 180-day exclusivity provisions, under which generic manufacturers could challenge the legitimacy of branded manufacturers' patents or claim that generic entry would not infringe them. Over time, as the law and economic benefits to generics were established, the likelihood of Paragraph IV challenges increased and most drugs became subject to challenges (Berndt *et al.*, 2007; Grabowski *et al.*, 2011a). This has led to uncertainty regarding the effective patent term for new drug introductions, as well as substantial litigation costs early in the product life cycle.

Under the BPCIA, an abbreviated application for a biosimilar can be filed after 4 years. The filing of an application triggers a series of potentially complex private information exchanges between the biosimilar applicant and reference product innovator. These exchanges are followed by negotiations and a process for instituting litigation on the core patents, when necessary. Congress has crafted these patent provisions while eliminating the incentive for litigation associated with a 180-day exclusivity period for the first filer in a successful challenge, as well as the automatic 30-month stay on entry under Hatch-Waxman. By instituting this potentially complex structured process for biologics, legislators hoped that patent disputes would be resolved before the expiration of the 12-year exclusivity period so that biosimilars can enter in a timely fashion. Generic manufacturers have raised concerns about the need to divulge proprietary information, and whether these rules will achieve their intended effects remains unknown.

Firms pursuing a biosimilar strategy could also choose to file a full BLA rather than an abbreviated application. Under the patent resolution provisions of the BPCIA, firms filing an abbreviated biosimilar application are required to disclose information about their manufacturing process and identify potential patent conflicts. By choosing instead to file a full BLA, the biosimilar firm would avoid this disclosure requirement, and, if approved, also be able to enter before the expiration of the 12-year exclusivity period. However, the firm needs to weigh these benefits against the additional investment of expenditures and time associated with filing a full BLA for a biosimilar product. Several firms apparently are considering this strategic option. Teva recently relied on a full BLA filing for its G-CSF filgrastim product, although the original submission to the FDA predated the establishment of a biosimilar pathway in the US. In Europe, the same Teva product is marketed under the name Tevagrastim® and was approved through an abbreviated biosimilar application for the reference product Neupogen® (**Table 1**). The product is scheduled to be launched in the US in late 2013 under a patent settlement with Amgen.

Summary and Conclusion

Biologics have accounted for a significant number of innovative medicines over the past three decades. At the same time, they account for a growing share of drug expenditures in some countries. Policymakers have anticipated the introduction of biosimilars mitigating these cost pressures. Biosimilars have been introduced in various EU countries beginning in 2007. The extent of biosimilar penetration for the biological entities, erythropoietin, G-CSF, and somatropin has varied substantially across therapies within a country and across countries for the same therapy. Germany has experienced the greatest initial uptake of biosimilars reflecting targeted incentives quotas and related factors.

The new US law is designed to balance the objectives of achieving cost savings in the current period and preserving incentives for continued innovation in the future. A number of leading biologic products with significant sales in the US are expected to experience some patent expiration in the next decade, so cost savings could grow significantly over time, depending on how other factors such as regulation, reimbursement, and intellectual property litigation evolve over this period.

In terms of maintaining incentives for future innovation, the US law provides for a 12-year exclusivity period after an innovator's product is approved before a biosimilar referencing can be approved utilizing an abbreviated pathway. This 12-year exclusivity period provides an important 'insurance policy' to the patent system and could be important in the case of biologics where patents may prove to provide less certain protection than those for NCEs. Analysis of a portfolio of representative biological products indicates that 12 years or more of exclusivity from patents or regulatory provisions is generally consistent with achieving breakeven returns that provide a risk-adjusted return on capital and R&D investments.

A number of important issues remain for future research, including how the new law will affect industry structure and incentives for undertaking R&D for biologics versus NCEs. As was the case with Hatch-Waxman, change may be gradual at first, but over time the new law could lead to profound changes in the economics and organization of the biopharmaceutical industry.

See also: Patents and Regulatory Exclusivity in the USA. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA

References

Avalere Health (2007). Modeling federal cost savings from follow-on biologics (study author King, R.). Available at: http://www.avalerehealth.net/research/docs/Follow_on_Biologic_Modeling_Framework.pdf (accessed 18.07.13).

Berndt, E., Mortimer, R., Bhattacharjya, A., Parece, A. and Tuttle, E. (2007). Authorized generic drugs, price competition, and consumers' welfare. *Health Affairs* **790**, 792–797.

Chauhan, D., Towse, A. and Mestre-Ferrandiz, J. (2008). The market for biosimilars: evolution and policy options. *Office of Health and Economics Briefing*, No. 45, 12–14.

Congressional Budget Office (CBO) (2008). S.1695, Biologics Price Competition and Innovation Act of 2007. Available at: http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/94xx/doc9496/s1695.pdf (accessed 18.07.13).

DiMasi, J. and Grabowski, H. (2007a). The economics of new oncology drug development. *Journal of Clinical Oncology* **209**, 214–215.

DiMasi, J. and Grabowski, H. (2007b). The cost of biopharmaceutical R&D: Is biotech different? *Managerial & Decision Economics* 469–475.

Express Scripts (2007). Potential savings of biogenerics in the United States (study authors Miller, S. and Houts, J.). Available at: http://www.express-scripts.com/research/research/archive/docs/potentialSavingsBiogenericsUS.pdf (accessed 18.07.13).

Food and Drug Administration (FDA) (2012a). Scientific considerations in demonstrating biosimilarity to a reference product. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf (accessed 18.07.13).

Food and Drug Administration (FDA) (2012b). Biosimilars: Questions and answers regarding implementation of the Biologics Price Competition and Innovation Act of 2009. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM273001.pdf (accessed 18.07.13).

Federal Trade Commission (2002). Generic drug entry prior to patent expiration: An FTC study. Available at: http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf (accessed 18.07.13).

Grabowski, H. (2007). Competition between generic and branded drugs. In Sloan, F. A. and Hsieh, C-R. (eds.) *Pharmaceutical innovation: Incentives, competition, and cost–benefit analysis in international perspective*, pp. 153–173. New York, NY: Cambridge University Press.

Grabowski, H. (2008). Follow-on biologics: Data exclusivity and the balance between innovation and competition. *Nature Reviews Drug Discovery* **479**, 479–487.

Grabowski, H., Cockburn, I. and Long, G. (2006). The market for follow-on biologics: How will it evolve? *Health Affairs* **1291**, 1291–1301.

Grabowski, H., Kyle, M., Mortimer, R., Long, G. and Kirson, N. (2011a). Evolving brand-name and generic drug competition may warrant a revision of the Hatch-Waxman Act. *Health Affairs* **30**, 2157–2166.

Grabowski, H., Lewis, T., Guha, R., et al. (2012). Does generic entry always increase consumer welfare? *Food and Drug Law Journal*.

Grabowski, H., Long, G. and Mortimer, R. (2011c). Data exclusivity for biologics. *Nature Reviews Drug Discovery* **15**, 15–16.

Grabowski, H., Ridley, D. and Schulman, K. (2007). Entry and competition in generic biologics. *Managerial & Decision Economics* **28**(4–5), 439–447.

Grabowski, H. and Wang, R. (2006). The quantity and quality of worldwide new drug introductions, 1982–2003. *Health Affairs* **25**(2), 452–460.

Hargrave, E., Hoadley, J., Merrell, K. (2010). Medicare Part D Formularies, 2006–2010: A Chartbook, *Report to the Medicare Payment Advisory Commission*. Available at: http://www.medpac.gov/documents/Oct10_PartDFormulariesChartBook_CONTRACTOR_RS.pdf (accessed 18.07.13).

Heldman, P. (2008). Follow-on biologic market: Initial lessons and challenges ahead. *Potomac Research Group, Presentation to the Federal Trade Commission*. Available at: www.ftc.gov/bc/workshops/hcbio/docs/fob/pheldman.pdf (accessed 18.07.13).

IMS (2011a). Shaping the biosimilars opportunity: A global perspective on the evolving biosimilars landscape. Available at: http://www.imshealth.com/ims/Global/Content/Home%20Page%20Content/IMS%20News/Biosimilars_Whitepaper.pdf (accessed 18.07.13).

IMS (2011b). *The Use of Medicines in the United States: Review of 2010*. IMS Institute for Healthcare Informatics. Available at: http://www.imshealth.com

Kambhammettu, S. (2008). The European biosimilars market: Trends and key success factors. *Scicast Special Reports*. Available at: http://scicasts.com/specialreports/20-biopharmaceuticals/2152-the-european-biosimilars-market-trends-and-key-success-factors (accessed 18.07.13).

Kozlowski, S., Woodcock, J., Midthun, K. and Behrman, S. R. (2011). Developing the nation's biosimilar program. *New England Journal of Medicine* **364**(5), 385–388.

Smith, I., Procter, M., Gelber, R. D., et al. (2007). 2-Year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: A randomized controlled trial. *Lancet* **369**(9555), 29–36.

Sokolovsky, J., Miller, H. (2009). Medicare payment systems and follow-on biologics, *Medicare Payment Advisory Commission*. Available at: http://www.medpac.gov/transcripts/followon%20biologics.pdf (accessed 18.07.13).

Trusheim, M. R., Aitken, M. L. and Berndt, E. R. (2010). Characterizing markets for biopharmaceutical innovations: do biologics differ from molecules? *Forum for Health Economics & Policy (Frontiers in Health Policy Research)* **13**(1), 1–45. The Berkeley Electronic Press.

Weaver, A. L. (2004). The impact of new biologicals in the treatment of rheumatoid arthritis. *Rheumatology* **43**(Supplement 3), iii17–iii23.

Further Reading

Grabowski, H., Long, G. and Mortimer, R. (2011b). Implementation of the biosimilar pathway: Economic and policy issues. *Seton Hall Law Review* **41**(2), 511–557.

Rovira, J., Espin, J., Garcia, L., and de Labry, A. O. (2011). The impact of biosimilars entry in EU markets. *Andalusian School of Public Health*. Available at: http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/biosimilars_market_012011_en.pdf (accessed 18.07.13).

Budget-Impact Analysis

J Mauskopf, RTI International, NC, USA

Abbreviations

ACS	Acute coronary syndrome
AIDS	Acquired immune deficiency syndromes
ART	Antiretroviral therapy
CMV	Cytomegalovirus
GBP	Great Britain pound
HER2+	Human epidermal growth factor positive
HES	Hospital episode statistics
HIV	Human immunodeficiency virus
HRG	Healthcare Resource Group
HTA	Health technology assessment
NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NSTEMI	NonST segment elevation myocardial infarction
MI	Myocardial infarction
PCI	Percutaneous coronary intervention
QALY	Quality-adjusted life-year
STEMI	ST segment elevation myocardial infarction
TIA	Transient ischemic attack
UK	United Kingdom
US	United States

Introduction

As healthcare costs increase because of the aging population and technological developments in healthcare, the need by healthcare decision makers for economic evaluations of new healthcare interventions becomes more important. A comprehensive economic evaluation of a new healthcare intervention requires an analysis of both the efficiency of the intervention compared with current treatment patterns and the annual budget impact of the new intervention. An analysis of the annual budget impact might be used to determine affordability of the new intervention, given healthcare budget constraints, or as an implementation tool for newly reimbursed interventions.

A budget-impact analysis typically first identifies, in a national or local health plan, the treated population for the indication for which the new intervention is approved. The analysis then estimates the annual change in healthcare expenditures for the treated population with and without the new intervention in the treatment mix for different rates of uptake of the new intervention. Unlike a cost-effectiveness analysis, which compares the new intervention with a standard of care, the comparison in a budget-impact analysis is between the mix of treatments before the new intervention is reimbursed and the mix of treatments after the new intervention is reimbursed, taking into account the rate of uptake of the new intervention.

There are several published guidelines for budget-impact analyses. These guidelines have been developed either by the health technology assessment (HTA) agencies that require a budget-impact analysis as part of a reimbursement submission (e.g., Pharmaceutical Benefits Advisory Committee (Australia), Canada, Taiwan, and the National Institute for Health and Clinical Excellence (NICE) in the UK) or by independent organizations (e.g., the International Society for Pharmacoeconomics and Outcomes Research). These guidelines describe the estimation framework and data sources that are recommended for performing budget-impact analyses.

Key Elements of a Budget-Impact Analysis

Budget-impact analyses have six primary elements, irrespective of the modeling framework used to derive the estimates: (1) treated population size, (2) time horizon, (3) treatment mix, (4) intervention costs, (5) other healthcare costs, and (6) presentation of results. In addition to these six primary elements, budget-impact analyses generally include sensitivity analyses to test the impact on budget estimates of the uncertainty in the input values used in the analysis or the variability of these inputs among health plans or health systems. Issues that should be considered for each of these elements are described in the following paragraphs.

The first step in a budget-impact analysis is to determine the population currently being treated for the disease indication of interest using epidemiological data. It is critical to estimate not only the size of the treated population but also the mix of disease severity in the population because treatments and disease-related healthcare expenditures may vary with disease severity. For example, individuals with schizophrenia that is refractory to treatment with standard care will have higher annual costs and will use a different mix of treatments than individuals who are responsive to treatment. It also is important to consider a possible 'woodwork' effect with a new intervention, that is, more patients with the indicated condition presenting for treatment when a better treatment becomes available. Finally, for a new intervention that reduces mortality, slows disease progression, and/or changes treatment patterns, changes in the treated population size and the distribution of the population by disease severity must be estimated on the basis of the assumed uptake rates for the new intervention.

The second element, the time horizon for the budget-impact analysis, typically is chosen on the basis of the requirements of the healthcare decision maker, rather than on the duration of the impact of the new treatment (as for a cost-effectiveness analysis). Because healthcare budget holders generally have a short planning horizon, time horizons of

 doi:10.1016/B978-0-12-375678-7.01423-1

3–5 years are usual. With such short time horizons, offsetting cost savings many years in the future from slowed disease progression of chronic diseases or prevention of future cases of the disease or its complications are not captured. But this is an accurate reflection of the costs incurred over the typical planning horizon.

The third element in a budget-impact analysis is the determination of the mix of interventions currently used for the indication and the predicted change in that mix if the new intervention is made available. Unlike cost-effectiveness analyses, which compare the outcomes when taking the new intervention with the outcomes with a standard-of-care intervention, a budget-impact analysis does not assume immediate switch by all patients to the new intervention. Rather, the new intervention is assumed to alter the mix of interventions used for the indication, using estimated or observed uptake data. The budget impact will be higher if the new intervention is used in place of a generic drug than if the new intervention is used in place of another branded drug or a surgical procedure. Also, the budget impact will be higher if the new intervention is combined with current treatments instead of substituted for them.

The costs associated with the current and new interventions should include some or all of the following, depending on the type of intervention: acquisition, administration or labor, other equipment, monitoring, and adverse-event or complication costs. For drugs, generally, wholesale acquisition costs (in the US) or national formulary costs are used as the default values, although the analysis should be designed so that discounts and copays can be subtracted from these costs to provide more accurate estimates of the healthcare decision makers' costs. For devices, wholesale prices should be used; for procedures, standard labor costs should be used. All of these costs are used to reflect the expected costs of current and new interventions to the decision maker for each year of the budget-impact analysis time horizon.

The fifth element, an estimate of the impact of the new intervention on other indication-related costs, excluding intervention costs, is generally but not always included in budget-impact analyses. A simple calculation can be used, based on clinical trial data, for example, to estimate these costs for acute conditions and for those chronic conditions where the full impact on indication-related costs happens within a short period of time or is not likely to change over the model time horizon. Alternatively, changes in indication-related costs for a chronic illness may be estimated by adapting the disease progression model (used to estimate the cost-effectiveness ratios) to calculate annual indication-related costs after reimbursement has been approved for the new treatment. This adaptation involves running the cost-effectiveness model in 'prevalence' mode where the model adds a newly treated cohort each subsequent year, in addition to tracking the starting cohort.

The sixth element in budget-impact analysis is the presentation of the results. Unlike cost-effectiveness analysis, where there may be a societal perspective that can be used as the reference case, there is no reference case in budget-impact analysis. The appropriate perspective for the analysis varies with each decision maker's budget responsibilities, which may range from a pharmacy or department budget to an entire hospital or outpatient clinic budget to countrywide healthcare services. Thus, the model needs to be programed in such a way that it can generate the budget impact from these multiple perspectives. In general, the results are presented undiscounted for year 1, 2, 3, etc. after the new intervention is made available to the decision maker's population. Cumulative, multiyear results also may be presented either discounted or undiscounted.

Clearly, in any budget-impact analysis, there is uncertainty about both model assumptions and input parameter values. In cost-effectiveness analysis, one-way and probabilistic sensitivity analyses are the recommended approaches for presenting the impact of the input parameter uncertainty. For budget-impact analyses, the more common approach to uncertainty analysis is to present a series of scenario analyses, changing input parameter values either one at a time or several at a time to create different scenarios that are meaningful to the decision maker; for example, changing intervention uptake rates and/or expected effectiveness in the decision maker's population. The decision maker may also enter values for input parameters that may vary among health plans or health systems but be known with certainty to each decision maker, such as drug costs, treated population size (based on size of population served and local incidence or prevalence rates), disease severity mix, and patient age distributions. Scenario analyses, which include alternate combinations of uncertain and variable input parameters, provide decision makers with more credible information about the range of possible results, given the specifics of their health plan or health system.

Categorization of Budget-Impact Modeling Approaches

There are three main budget-impact modeling approaches that have been used by HTA agencies and/or in published studies: (1) cost calculator, (2) Markov or state transition model, and (3) Monte-Carlo or discrete-event simulation model. The simplest approach, a cost calculator, is typically used for acute indications and for chronic indications where a static analysis is appropriate; Markov models and discrete-event simulation models are used for chronic indications where a dynamic approach is needed to capture the changes in treated population size, indication severity mix, or treatment patterns.

Budget-Impact Analysis: Cost Calculator Approach

For each drug recommended for reimbursement by the National Health Service (NHS) in England, NICE prepares a costing template for the drug's recommended use where budget impact is assessed to be greater than £1 million or more than 300 patients are affected. The costing template is presented on the NICE web site as a guide to budget planning for decision makers implementing the recommendation in the UK. These costing templates provide excellent examples of static models using a cost calculator approach. The NICE costing templates estimate the expected impact on the NHS budget of the new drug's predicted market uptake over the

next 3–5 years, after considering the current and new drug acquisition costs and associated administration, monitoring, and adverse-event costs. Where credible clinical data are available, the costing templates also estimate changes in disease-related costs associated with the use of the new drug. One-way sensitivity analyses, based on variations in the input parameter values, also are included in the more recent costing templates.

An example of an NICE costing template for prasugrel is presented here to illustrate the cost calculator approach for performing a budget-impact analysis. Prasugrel, when coadministered with acetylsalicylic acid, is indicated in the UK for the prevention of atherothrombotic events in patients with acute coronary syndrome (ACS) (that is, unstable angina, nonST segment elevation myocardial infarction (NSTEMI) or ST segment elevation myocardial infarction (STEMI)) who undergo primary or delayed percutaneous coronary intervention. However, prasugrel was recommended by NICE for reimbursement by the NHS as a treatment option for only a subset of the UK-indicated population: those with STEMI, those with STEMI or NSTEMI and stent thrombosis while taking clopidogrel, and those with NSTEMI and diabetes. The NICE costing template is presented in **Table 1** and includes the six key elements of a budget-impact analysis estimation of the population size, time horizon (1 year), current and projected treatment mix, drug costs, offsetting disease-related cost savings, and presentation of results. The footnotes to the NICE analysis table provided details of the data sources used in the costing template.

In the prasugrel example, because both drugs included in the analysis are oral drugs, there were no costs estimated for administration. Monitoring costs were also not included. Side effect costs, specifically those associated with bleeding events, were included in the rehospitalization rate. The prasugrel costing template included estimates of savings from a reduced rate of rehospitalization in the first year after the ACS

Table 1 Cost calculator model: The NICE costing template for prasugrel

Note	*Description*	*Unit costs*	*Units*	*Total cost*
1	Total population		50 542 505	
1	Population < 35 years		22 263 025	
1	Population 35–74 years		24 365 697	
1	Population 75 +		3 913 783	
2	Estimated annual incidence of ACS, 35–74		0.6%	
2	Estimated annual incidence of ACS, 75 +		2.3%	
3	Number of people diagnosed with ACS each year, 35–74		144 525	
3	Number of people diagnosed with ACS each year, 75 +		89 089	
	Total ACS patients per year		233 614	
4	Proportion needing immediate PCI		16%	
	Number needing immediate PCI		37 430	
5	Proportion without previous TIA or stroke		96%	
	Number without previous TIA or stroke		35 933	
6	Proportion with STEMI		24.6%	
	Number with STEMI		8839	
7	Proportion with STEMI and stent thrombosis on clopidogrel		2.35%	
	Number with STEMI and stent thrombosis who may receive prasugrel		208	
	Number with STEMI without stent thrombosis who may receive prasugrel		8632	
8	Estimated uptake of prasugrel in those with STEMI but without stent thrombosis		70%	
	Estimated number with STEMI who take prasugrel		6250	
9	Proportion who have NSTEMI		75.4%	
	Number who have NSTEMI		27 093	
	Proportion of those with NSTEMI who have stent thrombosis on clopidogrel		2.35%	
	Number with NSTEMI who may received prasugrel		637	
10	Estimated proportion of NSTEMI patients who have diabetes		17.50%	
	Number of NSTEMI patients with diabetes where prasugrel is an option		4630	
	Estimated uptake of prasugrel in NSTEMI patients		70%	
	Estimated total number of NSTEMI patients who may receive prasugrel		3878	
	Estimated total ACS patients who may receive prasugrel		10 128	
11	Current care: People aged less than 75 years			
	Clopidogrel			
	Loading dose: 300 mg	£5.04	1	£5.04
	Maintenance dose: 75 mg day^{-1} (30 day pack) for 1 year	£37.83	12	£453.96
	Cost per patient per year	£42.87		£459.00
	Proportion of patients, 35–74 years		62%	
	Estimated current care costs, 35–74 years		£6265	£2 875 814
12	Current care: People aged more than 75 years			
	Clopidogrel			

(*Continued*)

Table 1 Continued

Note	Description	Unit costs	Units	Total cost
	75 mg day^{-1} (30 day pack) for 1 year: Cost per patient per year	£37.83	12	£453.96
	Proportion of patients 75 + years		38%	
	Estimated current care costs, 75 + years		3862	£1 753 262
	Total costs, current care			£4 629 077
	Proposed care			
	Prasugrel			
	Loading dose 60 mg	£10.20	1	£10.20
	Maintenance dose 10 mg (5 mg for those weighing < 60 kg or 75 + years) for 1 year (28 day pack)	£47.56	13	£618.28
	Cost per patient per year	£57.76		£628.48
	Proportion who may receive prasugrel		100%	
	Total cost of proposed care with prasugrel		10 128	£6 364 958
	Estimated incremental costs of prasugrel			£1 735 881
	Potential disease-related savings			
	Reduction in rate of rehospitalizations		0.87%	
	Number of rehospitalizations avoided		88	
	Weighted average cost of rehospitalization	£5345		–£470 360
	Estimated net budget impact of prasugrel			£1 265 521

Abbreviations: ACS = acute coronary syndrome; HES = Hospital Episode Statistics; HRG = Healthcare Resource Group; NHS = National Health Service; NICE = National Institute for Health and Clinical Excellence; NSTEMI = nonST segment elevation myocardial infarction; PCI = percutaneous coronary intervention; STEMI = ST segment elevation myocardial infarction; TIA = transient ischemic attack; UK = United Kingdom.

Notes:

1. Total population is for England. *Source*: Office for National Statistics population estimates by primary care organization 2006.
2. Calculated incidence from Taylor, M. J., Scuffham, P. A., McCollam, P. L., *et al.* (2007). Acute coronary syndromes in Europe: 1 year costs and outcomes. *Current Medical Research and Opinion* **23**(3), 495–503. For people aged 75 years of age and more than HES 2007–08 data used (codes I20.0–I22.9) to calculate incidence.
3. Age-related incidence from Main, C., Palmer, S., Griffin, S. *et al.* (2004). Clopidogrel used in combination with aspirin compared with aspirin alone in the treatment of nonST segment elevation acute coronary syndromes (ACS). *Health Technology Assessment* 8 (40). ACS incidence significant for age groups from 35 upward. The over 75s category reflects different license indications for the drugs in respect of this age group.
4. Estimate from British cardiovascular intervention society returns (2007) – 53.72% of patients undergoing percutaneous coronary intervention have acute coronary syndrome (nonST segment elevation myocardial infarction (MI)/Unstable Angina and ST segment elevation MI). Total patients 69 677 × 53.72% = 37 430 patients. This is equal to 16% of the total acute coronary syndrome patients per year.
5. Prasugrel-specific product characteristics exclude patients with prior TIA/stroke. This is estimated to be 4% on the basis of the TRITOM TIMI 38 study – Wivott (2007).
6. British cardiovascular intervention society audit returns (2007). Figures taken from manufacturers submission.
7. Results taken from TRITON-TIMI 38 trial included in Evidence review group report (2009). The proportion of patients receiving stents where stent thrombosis has occurred during clopidogrel treatment. Appendix 3 Table 9.6. It has been assumed this proportion applies to nonST segment elevation myocardial infarction patients receiving clopidogrel treatment.
8. Estimate based on expert opinion. Please enter own estimates.
9. British cardiovascular intervention society audit returns (2007). Figures taken from manufacturers' submission.
10. British cardiovascular intervention society audit returns (2007). The figure has been adjusted for people in whom stent thrombosis has occurred during clopidogrel treatment as this group would be recommended prasugrel.
11. Price of clopidogrel: British national formulary 57 edn. (2009). Price of prasugrel as in manufacturers' submission (2009). Proportion is based on annual incidence numbers for people aged 35–74 years of age.
12. Please adjust proportion to reflect local estimates. Where not all people aged 75 years of age and more receive prasugrel for its indicated use, it is assumed that these people would be treated in line with current practice and therefore no incremental cost is likely to be incurred.
13. Daiichi-Sankyo (2009) Eli Lilly and Company Ltd STA submission: Prasugrel for the treatment of acute coronary artery syndromes with coronary intervention. Table 34: TRITON-TIMI rehospitalizations summarized by category with UK NHS reference costs and adjusted to reflect Table 36 – UK rehospitalization rates. The calculated figure for the number of rehospitalizations avoided in the cost per 100 000 columns has been rounded to the nearest whole number which is 1. For smaller populations, savings may not be significant or robust due to the randomness of events. For larger populations, saving results are scaled in the normal way, i.e., rounded to the nearest 1.
14. Reduction in rate of rehospitalizations taken from Table 36 manufacturers submission relating to UK reduction rate. Rehospitalization categories mapped to NHS mandatory tariff 2009/10 and reference costs 2007–08 (where no mandatory tariff). HRG codes used are: AA21Z; AA09Z; AA15Z; EB10Z; EA31Z–34Z; EA14Z–16Z; EA40Z–42Z. Reference cost code used FZ38A.

Source: Adapted from National Institute for Health and Clinical Excellence (2009). *TA 182 Prasugrel for the Treatment of Acute Coronary Syndromes with Percutaneous Coronary Intervention*. London: NICE. Issued October 2009; Current as of January 2013 but could be superceded; available at: www.nice.org.uk (accessed 10.01.13).

episode that was observed in a large head-to-head clinical trial with clopidogrel.

This NICE costing template for prasugrel also included an extensive one-way sensitivity analysis, using maximum and minimum values for the following input parameter values:

- The annual incidence, by age group.
- The proportion of patients with ACS in whom immediate percutaneous coronary intervention is needed.
- The proportion of patients with STEMI.
- The uptake rate in the STEMI population.
- The proportion of the ACS population with NSTEMI.
- The proportion of NSTEMI population with stent thrombosis on clopidogrel.
- The proportion of the NSTEMI population with diabetes.
- The uptake rate of prasugrel in the NSTEMI population.
- The proportion of patients receiving prasugrel who are more than 75 years of age.
- The cost of clopidogrel per patient per year.
- The reduction in rate of rehospitalizations.
- The weighted average cost of rehospitalizations.

The rationale for the selection of the minimum and maximum values for the sensitivity analysis is not provided in the costing template.

Budget-impact analyses using a cost calculator approach have also been published in peer-reviewed journals. For example, in an article by Chang and Sung, the budget impact of using pimecrolimus cream for atopic dermatitis or eczema was estimated using estimates of the number of people seeking care for this condition each year and the average number of physician visits for the condition each year. Chang and Sung used data from a clinical trial of pimecrolimus to estimate likely reductions in follow-up physician visits for those patients who were treated with pimecrolimus. Although the condition is chronic, it is not progressive or life threatening. Therefore, the use of a static cost calculator approach is appropriate. Chang and Sung estimated the budget impact for a single year, on the basis of observed market share for the first year the drug was introduced, but tested the impact of changes in market uptake in a sensitivity analysis.

Using a static cost calculator approach to budget impact analysis for chronic progressive and/or life-threatening diseases may underestimate the budget impact. For example, Smith and colleagues used a static approach to estimate the budget impact of valsartan for the treatment of patients with heart failure. The authors' estimates were based on the number of enrollees with heart failure in a US health plan and on the average number of hospitalizations each year for these patients. The authors used data from a clinical trial of valsartan that showed a reduction in the number of hospitalizations and in the length of hospital stay for patients treated with valsartan. However, annual mortality rates with heart failure are significant, and the valsartan clinical trial also estimated a reduction in mortality for patients on valsartan. Such a reduction in mortality would result in an increased number of patients being treated for heart failure at any one time and an associated increase in treatment and monitoring costs for the health plan. This increase in the population size being treated was not included in the Smith and colleagues' budget-impact analysis. A dynamic disease progression model could have been used to estimate the change in the size of the prevalent population over time, given the reduction in mortality rates. Alternately, estimates of the change in life expectancy with valsartan could have been derived from the clinical trial data and used to estimate the change in the treated population size at steady state and used in the cost calculator approach.

A budget-impact analysis by Dee and colleagues estimated the budget impact of natalizumab over a 3 year time horizon for multiple sclerosis, a slowly progressing chronic disease. In this analysis, the authors explicitly captured the budget impact of the increasing uptake of natalizumab over time. The budget-impact estimates in the Dee and colleagues' study were based on the reduced costs for treating relapses of multiple sclerosis and the increased drug costs for natalizumab, including administration costs and monitoring for serious side effects such as progressive multifocal leukoencephalopathy. The authors also considered different payer perspectives and adjusted the budget impact depending on which payer perspective was considered. However, this static cost calculator approach ignored the impact on multiple sclerosis treatment costs of slowing the rate of disease progression that is associated with natalizumab treatment.

In the Smith and colleagues' study, the estimated budget impact of the new treatment did not include the additional drug-related and disease monitoring and symptomatic treatment costs in the extra months of life for the patient. But for patients with heart failure in these studies, the additional life expectancy may be short and the impact on the size of the treated population relatively small. Similarly, the budget impact of slowing disease progression in multiple sclerosis, omitted from the Dee and colleagues' study, is likely to be small within the time horizon of the budget-impact analysis. But in other chronic conditions, the impact on life expectancy could be significant, for example, for human immunodeficiency virus (HIV) infection. In this case, a dynamic budget-impact model, using either a Markov model or simulation approach, might be more appropriate.

Budget-Impact Analysis: Markov Model Approach

A study by Mauskopf demonstrated how a Markov model can be used to develop both cost-effectiveness and budget-impact estimates for a hypothetical new treatment for HIV infection. To develop the budget-impact estimates, it was first necessary to understand the current distribution of HIV patients among different HIV health states, measured in terms of ranges of CD4 cell counts. This distribution was obtained for a cohort of patients who were not treated, using natural history data that provided estimates of the time spent in each health state. Using these estimates and the number of new patients diagnosed and their CD4 cell-count distribution, the Markov model was run, adding a newly diagnosed cohort each year, until a steady state was reached for the number of patients in each health state without treatment. The introduction of the hypothetical antiretroviral therapy drug regimen was assumed to shift the CD4 cell-count up by one CD4 cell-count range for all patients in the treated cohort and to hold the cohort there for 4 years before disease progression resumed. The Markov model was rerun with the hypothetical antiretroviral drug

regimen. For each cycle of the model, the number of individuals alive in each health state was generated. A new steady state was reached in 10–20 years. For each health state, treatment costs, rates of opportunistic infections, and days in the hospital were estimated. Population estimates for all of these outcomes were generated for each year after introduction of the antiretroviral drug regimen.

A Markov budget-impact model can be programed to capture only the budget impact for newly entering cohorts cumulatively in each year after a new drug becomes available; alternatively, the model can be programed to assume that all prevalent patients also immediately switch to the new treatment or that a certain proportion of the prevalent patients switch each year. In the Mauskopf model, there were 10 680 persons alive in the UK with HIV in 1994 and an incident cohort of 1258 persons per year. The treatment regimens compared were no antiretroviral treatment and a hypothetical antiretroviral drug regimen that was assumed to stop disease progression for 4 years but to be taken for 6 years. All persons alive with HIV were assumed to switch immediately to the antiretroviral drug regimen, as were those individuals newly diagnosed during the model time horizon. Selected model inputs and outcomes are shown in Tables 2 and 3, respectively.

In this model, the impact of antiretroviral therapy (ART) on life expectancy for people with HIV infection was large, resulting in a significant increase in the number of individuals living with acquired immune deficiency syndromes (AIDS) and HIV infection and a shift to less severe disease stages. In this analysis, other outcomes that are of importance to patients and health planners were estimated, including the number of cases of opportunistic infections, illustrated in Table 3 by the number of cases of CMV infection, as well as the number of hospital days used by individuals with HIV infection. This latter value can be very useful for planning for hospital care for those with HIV infection.

Mar and colleagues presented a similar approach to budget-impact analysis using a Markov model for a Basque population to estimate the impact of the use of thrombolysis for patients with stroke on the prevalence of different degrees of residual disability in patients with stroke and the associated budget impact. In their study, the current prevalent population in different poststroke health states (death, disability, autonomous, and recurrent stroke) without thrombolysis was estimated using data on stroke incidence stratified by age and sex, all-cause mortality rates, stroke excess mortality risk, and disability outcomes from stroke. The budget impact associated with the use of thrombolysis was estimated using trial data indicating that the percentage of patients with residual disability was lower when thrombolysis was used than when it was not used. Thus, the Markov model was run over a 15 year time horizon with the two different rates of disability, as well as changing numbers of strokes due to the aging population. The results for the Basque population are shown in Table 4. In the Mar's study, the current population health state prevalence rates, as estimated by the Markov model for patients

Table 2 Markov model: Selected input data for HIV model

Input data	*CD4 cell-count range*				
	> 500	*350–500*	*200–349*	*100–199*	*< 100*
Average time in disease state: No ART (years)	2 (after diagnosis)	1.8	1.8	1.5	1.3
Transition probability to next worse state: No ART[a]	0.5	0.5556	0.5556	0.6667	0.7692
Annual healthcare costs: Excluding ART[b]	£1834	£1834	£1834	£1912	£7490
Annual community service costs[b]	£1137	£1137	£1137	£1378	£2230
Annual CMV incidence	0.0024	0.0024	0.0024	0.0750	0.2550
Annual hospital days	1.13	1.13	1.13	2.87	29.9

[a]Transition probability is equal to (1/time in state).

[b]1995 Great Britain pounds inflated to 1999 Great Britain pounds, using the hospital and community health services price index.

Abbreviations: ART = antiretroviral therapy; CMV = cytomegalovirus; HIV = human immunodeficiency virus.

Source: Adapted from Table 1, reprinted from Mauskopf, J. (2000). Meeting the NICE requirements: A Markov model approach. *Value in Health* **3**(4), 287–293.

Table 3 Markov model: Annual outcomes with and without ART for HIV infection

Annual outcomes	*Year one*		*Year three*		*Year six*	
	No ART	*ART*	*No ART*	*ART*	*No ART*	*ART*
Cost GBP ($\times 10^6$)	29.1	66.9	29.1	123.4	29.1	151.6
Number of persons treated	10 680	11 938	10 680	14 454	10 680	17 804
Cost per person	2 725	6 260	2 725	9 353	2 725	8 829
CMV cases	581	149	581	155	581	502
Hospital days	62 775	16 200	62 775	19 000	62 775	60 665

Abbreviations: ART = antiretroviral therapy; CMV = cytomegalovirus; GBP = Great Britain pound; HIV = human immunodeficiency virus; QALY = quality-adjusted life-year.

Source: Adapted from **Table 3** in Mauskopf, J. (2000). Meeting the NICE requirements: A Markov model approach. *Value in Health* **3**(4), 287–293.

without thrombolysis, were validated on the basis of population registry data and an alternative modeling approach for estimating poststroke life expectancy.

Two other published studies illustrate the use of Markov models to capture the dynamic aspects of budget-impact analysis. In the budget-impact analysis for trastuzumab in early breast cancer by Purmonen and colleagues, a 4 year time horizon was modeled using a state transition model with two health states: free of distant recurrence and with distant recurrence. The budget impact was estimated as the difference in cumulative undiscounted 1, 2, 3, and 4 year costs for all cohorts starting treatment during the model time period with or without the use of adjuvant trastuzumab. The model was based on the number of early breast cancer patients, human epidermal growth factor positive (HER2 +) prevalence, length and cost of adjuvant treatment, and the effectiveness of the treatment. All HER2 + patients were assumed to be treated with trastuzumab. Sensitivity analyses included a scenario analysis that looked at different treatment patterns, prevalence of HER2 +, and treatment costs. In addition, a probabilistic sensitivity analysis was included that estimated the impact of the following uncertain or variable parameter inputs: number of early breast cancer patients, HER2 + prevalence in those with early breast cancer, disease-related transition probabilities, and treatment costs. The results of the probabilistic sensitivity analysis were presented as an affordability curve in which the probability of the budget impact being below different budget constraints was presented (see **Figure 1**).

In a combination of cost-utility and budget-impact analysis of third-generation aromatase inhibitors for advanced breast cancer, Marchetti and colleagues used a state transition model to estimate the life expectancy and lifetime costs for a single annual cohort of patients newly diagnosed with advanced breast cancer and starting treatment with or without the use of anastrozole or letrozole. The authors estimated the budget impact for a single cohort under the assumption that all

Table 4 Markov model: Stroke outcomes with and without thrombolysis

Annual Outcomes	*2000*	*2005*	*2010*	*2015*
Stroke number	4 541	5 176	5 812	6 447
Dependent patients: no thrombolysis	6 505	8 478	10 450	12 423
Dependent patients: 10% with thrombolysis	6 505	8 368	10 232	12 095
Difference in dependent patients	0	109	219	328
Number with thrombolysis	454	518	581	645
Reduced costs for dependency (€)	0	1 132 000	2 264 000	3 396 000
Increased costs for thrombolysis (€)	1 223 000	1 395 000	1 566 000	1 737 000
Gain in QALYs	0	36.59	73.19	109.78

Abbreviation: QALY = quality-adjusted life-year.

Source: Adapted from **Table 5** in Mar, J., Sainz-Ezkerra, M. and Miranda-Serrano, E. (2008). Calculation of prevalence with Markov models: Budget impact analysis of thrombolysis for stroke. *Medical Decision Making* **28**(4), 481–490. Copyright © 2008 by Sage Publications. Reprinted by Permission of SAGE Publications.

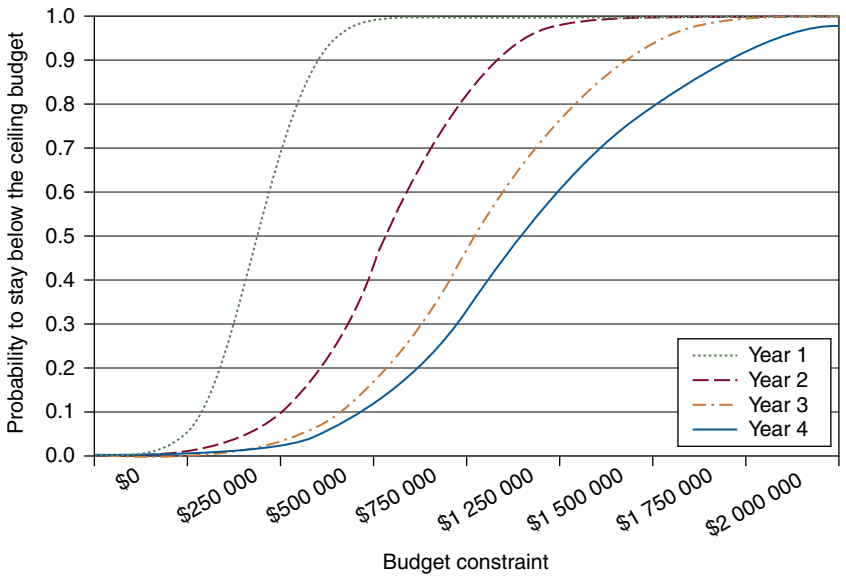


Figure 1 Markov model: Results of the probabilistic sensitivity analysis for trastuzumab in early breast cancer. Reprinted from Figure 3 in Purmonen, T. T., Auvinen, P. K. and Martikainen, J. A. (2010). Budget impact analysis of trastuzumab in early breast cancer: A hospital district perspective. *International Journal of Technology Assessment in HealthCare* **26**(2), 163–169. Reproduced with permission from Cambridge University Press.

patients in the cohort are treated with either anastrozole or letrozole. This focus on a single cohort and assumption of 100% uptake is typical for cost-effectiveness analyses but is less often used for budget-impact analyses.

Budget-Impact Analysis: Simulation Model Approach

Another type of disease model frequently used in cost-effectiveness analyses of new treatments is Monte Carlo or discrete-event simulation. In simulation models, the disease pathway is simulated for a group of individual patients with different characteristics for the duration of the disease episode or for lifetime (for chronic diseases). This approach to disease modeling has several advantages over a deterministic Markov approach: variability among patients in disease outcomes and in the impact of the treatment is captured explicitly; all relevant patient, system, and treatment characteristics can be captured without requiring an expansion of health states; disease and treatment history over time can be accounted for in the analysis; and multiple events can occur at the same time. Discrete-event simulation models track patients on the basis of the time to the next event, whereas Monte Carlo simulation models typically track the patients at specific time points. The disadvantage of the simulation approach is that it generally requires additional data inputs and additional computation time compared with the Markov modeling approach.

As with Markov modeling, the discrete-event simulation approach can be used to generate budget-impact as well as cost-effectiveness estimates by simulating a prevalent population rather than a single population cohort. Martin and colleagues presented the results of a budget-impact analysis for expanded screening for HIV in the US, using a Monte Carlo simulation model that included screening and treatment for HIV infection. This model has been used extensively for cost-effectiveness analyses of alternative management strategies for HIV infection. In their publication of the simulation model's results, the authors estimated the number of prevalent cases of HIV infection that were currently undetected and the annual number of new cases of HIV infection, using national prevalence and incidence data. Using a series of published studies and reports, the authors also estimated the proportion of these individuals that would be eligible for government-sponsored HIV screening, as well as the CD4 cell-count and viral load distributions for those persons unaware of their HIV status. The authors then entered this patient population into the screening module of their Monte Carlo simulation model and tracked costs over a 5 year time frame, with and without the introduction of a new screening program. Martin and colleagues presented the additional number of cases identified from expanded screening each year for 5 years and the undiscounted budget impact of expanded screening and the associated earlier treatment by discretionary and entitlement programs (see **Tables 5** and **6**; **Figure 2**).

Discrete-event simulation models also have been used to estimate budget impact of drug treatments, tracking both the prevalent and incident populations to determine the annual budget impact. Caro and colleagues used a discrete-event simulation model to estimate the budget impact over 100 days for alternative treatments of bipolar-associated mania, using estimates of changes in response to therapy in the Young Mania Rating Scale over time. Mar and colleagues used a discrete-event simulation model to estimate the budget impact of thrombolysis in patients with stroke, using estimates of a reduced number of patients with residual disability after stroke in those patients given thrombolysis treatment. In both

Table 5 Simulation model: Clinical characteristics of newly detected HIV-Infected individuals eligible for care through discretionary and entitlement programs

	Current practice	*Expanded screening*
Number identified over 5 year period		
Prevalent cases in year 1	54 343	63 747
Prevalent cases in year 2	18 362	24 062
Prevalent cases in year 3	17 276	19 755
Prevalent cases in year 4	14 759	15 106
Prevalent cases in year 5	11 366	10 651
Total prevalent cases in period	116 107	133 321
Incident cases in year 1	4 099	6 701
Incident cases in year 2	8 379	13 258
Incident cases in year 3	12 340	18 764
Incident cases in year 4	16 086	23 417
Incident cases in year 5	19 618	27 361
Total incident cases in period	60 523	89 501
Mechanism of detection, prevalent cases		
Screening (%)	19.7	33.1
Opportunistic infection (%)	68.3	57.8
Never detected (%)	12.0	9.1
Mechanism of detection, incident cases		
Screening (%)	39.3	60.2
Opportunistic infection (%)	49.0	32.3
Never detected (%)	11.7	7.5
CD4 count at detection		
Prevalent (mean cells mm^{-3})	122	140
Incident (mean cells mm^{-3})	251	312

Abbreviations: HIV = human immunodeficiency virus; QALY = quality-adjusted life-year.

Source: Adapted from **Table 2** in Martin, E. G., Paltiel, A. D., Walensky, R. P. and Schackman, B. R. (2010). Expanded HIV screening in the US: What will it cost government discretionary and entitlement programs? A budget impact analysis. *Value in Health* **13**(8), 893–902.

Table 6 Simulation model: Incremental quality-adjusted survival per person

Cases	*Current practice*	*Expanded screening*
Prevalent cases (ΔQALYs)	–	2.0
Incident cases (ΔQALYs)	–	3.2

Note: These numbers refer to the quality-adjusted survival over the newly detected cases' lifetime and not just the 5 year time horizon of the budget-impact analysis.
Abbreviation: QALY = quality-adjusted life-year.

Source: Adapted from **Table 2** in Martin, E. G., Paltiel, A. D., Walensky, R. P. and Schackman, B. R. (2010). Expanded HIV screening in the US: What will it cost government discretionary and entitlement programs? A budget impact analysis. *Value in Health* **13**(8), 893–902.

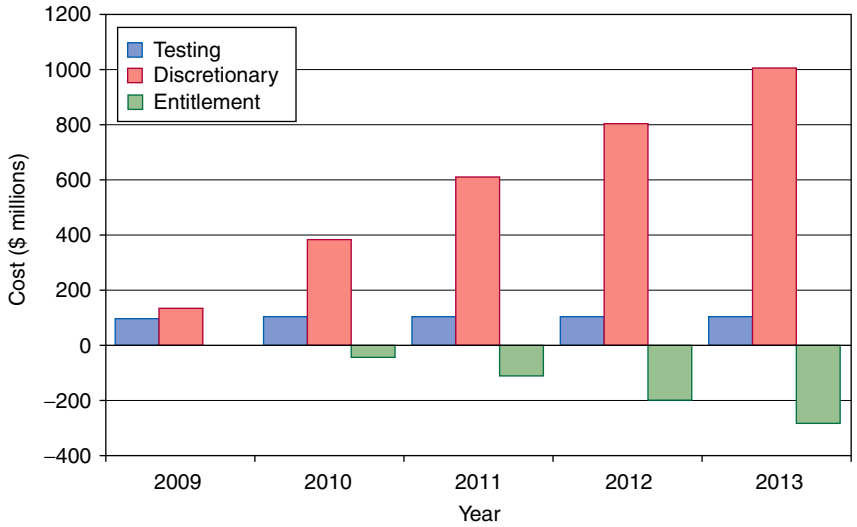


Figure 2 Simulation model: Results. Reprinted from Figure 1 in Martin, E. G., Paltiel, A. D., Walensky, R. P. and Schackman, B. R. (2010). Expanded HIV screening in the US: What will it cost government discretionary and entitlement programs? A budget impact analysis. *Value in Health* **13**(8), 893–902.

models, prediction equations were estimated by using patient-level data to estimate time to the primary events included in the model.

The advantages of using Monte Carlo or discrete-event simulation models to estimate budget impact of alternative disease management strategies are that, generally such models have been previously validated for the cost-effectiveness analyses and the inputs are consistent for both types of estimates. In addition, changes in disease severity and life expectancy over time can be included in the model. This is very important for HIV infection or stroke, where alternative screening or treatment strategies can have a major impact on the treated population size and/or severity mix, and thus on healthcare decision makers' budgets.

Conclusions and Where Next

As illustrated in this article, budget-impact models can be developed using a variety of approaches: a cost calculator approach or disease progression modeling approaches using either Markov or simulation models. Generally, the simpler approach is preferred by healthcare decision makers because such an approach is more transparent and can more readily be run using individual health plan characteristics. The cost-calculator approach can be used for acute illnesses, as well as for chronic illness where changes in disease severity, life expectancy, or treatment patterns (1) do not occur, (2) occur very rapidly and can readily be captured in a cost-calculator model, or (3) occur beyond the time horizon of the budget-impact analysis. In instances where the changes in disease severity, life expectancy, and/or treatment patterns cannot be credibly captured in a cost-calculator model, a disease progression modeling approach might be needed.

A disease progression modeling approach may be more desirable when an integrated cost-effectiveness and budget-impact model is desired. However, care needs to be taken to ensure that the budget-impact estimates are generated for the prevalent population rather than for the single-disease cohort that is typically used for cost-effectiveness analysis. The budget-impact model should also compare a mix of current and future treatments rather than a simple comparison of all patients treated with either a current treatment or a new treatment, as is typically seen in a cost-effectiveness analysis. In addition, the appropriate costs for the budget holder are their actual prices paid net of discounts and copays while opportunity costs are more appropriately used for cost-effectiveness analyses.

The question of how to reflect the uncertainty or variability in the inputs to a budget-impact analysis is also important. There are several different types of uncertainty or variability that can be present in the input parameter values, uncertainty about the estimates of the efficacy of the new and current interventions, variability in patient characteristics and current treatment patterns in different healthcare settings, and both uncertainty and variability in the changes in expected treatment patterns with the availability of the new intervention. Because these analyses are aimed to help healthcare decision makers understand the budget impact on the population for which the decision makers have responsibility, budget-impact analyses most commonly include either one-way sensitivity analyses, using ranges of both uncertain efficacy inputs and differences in patient characteristics and current and future treatment patterns (e.g., NICE cost calculators), or scenario analyses where several of these input parameter values may be changed to produce a scenario that most closely matches the healthcare decision maker's population. Probabilistic sensitivity analyses are sometimes included in published budget-impact analyses, but these are probably not very useful for healthcare decision makers because the sensitivity of the budget-impact analyses results to parameter uncertainty may be less than the sensitivity of the budget-impact analysis to variabilities in the healthcare decision maker's population characteristics and treatment patterns. The concept of the

affordability curve for different budget constraints as used in Purmonen and colleagues' article may be a useful way to present the results of a probabilistic sensitivity analysis. However, the probabilistic sensitivity analysis presented in Purmonen and colleagues' study included both uncertain parameters (HER2+ prevalence and transition probabilities reflecting efficacy) and variable parameters that would probably be known with certainty by the decision maker (price of trastuzumab and number of patients), thus reducing the value of their probabilistic sensitivity analysis.

Although the primary purpose of a budget-impact model is to estimate the annual impact on a health plan budget after a new intervention is reimbursed for the health plan's covered population, a budget-impact models may also generate estimates of the associated changes in population health outcomes during the same time period. These estimates may be used in the budget-impact analysis to estimate the changes in disease-related costs, but the population health estimates also can provide useful information for healthcare decision makers. For example, estimates of changes in disease cases or hospital days may be useful for health services planners. These population-based estimates of these outcome changes should be presented for each year after introduction of the new intervention along with the budget-impact estimates.

See also: Adoption of New Technologies, Using Economic Evaluation. Analysing Heterogeneity to Support Decision Making. Biopharmaceutical and Medical Equipment Industries, Economics of. Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties. Economic Evaluation of Public Health Interventions: Methodological Challenges. Economic Evaluation, Uncertainty in. HIV/AIDS, Macroeconomic Effect of. HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Infectious Disease Modeling. Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity. Observational Studies in Economic Evaluation. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Problem Structuring for Health Economic Model Development. Public Health: Overview. Searching and Reviewing Nonclinical Evidence for Economic Evaluation. Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies. Statistical Issues in Economic Evaluations. Synthesizing Clinical Evidence for Economic Evaluation. Valuing Informal Care for Economic Evaluation

Further Reading

Caro, J. J., Huybrechts, K. F., Xenakis, J. G., et al. (2006). Budgetary impact of treating acute bipolar mania in hospitalized patients with quetiapine: An economic analysis of clinical trials. *Current Medical Research and Opinion* **22**, 2233–2242.

Chang, J. and Sung, J. (2005). Health plan budget impact analysis for pimecrolimus. *Journal of Managed Care Pharmacy* **11**, 66–73.

Danese, M. D., Reyes, C., Northridge, K., et al. (2008). Budget impact model of adding erlotinib to a regimen of gemcitabine for the treatment of locally advanced, nonresectable or metastatic pancreatic cancer. *Clinical Therapeutics* **30**, 775–784.

Dasbach, E. J., Largeron, N. and Elbasha, E. H. (2008). Assessment of the cost-effectiveness of a quadrivalent HPV vaccine in Norway using a dynamic transmission model. *Expert Review of Pharmacoeconomics and Outcomes Research* **8**, 491–500.

Dee, A., Hutchinson, M. and De La Harpe, D. (2012). A budget impact analysis of natalizumab use in Ireland. *Irish Journal of Medical Sciences* **181**, 199–204.

Mar, J., Arrospide, A. and Comas, M. (2010). Budget impact analysis of thrombolysis for stroke in Spain: A discrete event simulation model. *Value in Health* **13**, 69–76.

Mar, J., Sainz-Ezkerra, M. and Miranda-Serrano, E. (2008). Calculation of prevalence with Markov models: Budget impact analysis of thrombolysis for stroke. *Medical Decision Making* **28**, 481–490.

Marchetti, M., Caruggi, M. and Colombo, G. (2004). Cost utility and budget impact of third-generation aromatase inhibitors for advanced breast cancer: A literature-based model analysis of costs in the Italian National Health Service. *Clinical Therapeutics* **26**, 1546–1561.

Martin, E. G., Paltiel, A. D., Walensky, R. P. and Schackman, B. R. (2010). Expanded HIV screening in the U.S.: what will it cost government discretionary and entitlement programs? A budget impact analysis. *Value in Health* **13**, 893–902.

Mauskopf, J. (2000). Meeting the NICE requirements: A Markov model approach. *Value in Health* **3**, 287–293.

Mauskopf, J., Murroff, M., Gibson, P. J. and Grainger, D. L. (2002). Estimating the costs and benefits of new drug therapies: Atypical antipsychotic drugs for schizophrenia. *Schizophrenia Bulletin* **28**, 619–635.

Purmonen, T. T., Auvinen, P. K. and Martikainen, J. A. (2010). Budget impact analysis of trastuzumab in early breast cancer: A hospital district perspective. *International Journal of Technology Assessment in Health Care* **26**, 163–169.

Smith, D. G., Cerulli, A. and Frech, F. H. (2005). Use of valsartan for the treatment of heart-failure patients not receiving ACE inhibitors: A budget impact analysis. *Clinical Therapeutics* **27**, 951–959.

Sullivan, S. D., Mauskopf, J. A., Augustovski, F., et al. (forthcoming). Budget Impact Analysis – Principles of Good Practice: Report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. *Value Health*.

Wiviott, S. D., Braunwald, E., McCabe, C. H., et al. TRITON-TIMI 38 Investigators (2007). Prasugrel versus clopidogrel in patients with acute coronary syndromes. *New England Journal of Medicine* **357**, 2001–2015.

Collective Purchasing of Health Care

M Chalkley and I Sanchez, University of York, Heslington, York, UK

The term collective purchasing is often used interchangeably with cooperative purchasing, group purchasing and collaborative purchasing and sundry other expressions. A fuller list of terms is set out by Schotanus and Telgen (2007) and Tella and Virolainen (2005) who provide a useful starting point for investigating the wider use of these arrangements.

There are a number of notions of collective purchasing in health care and here three are considered: collective purchasing of health-care inputs, collective purchasing of health insurance, and collective purchasing of health-care treatments or interventions. The details are set out below.

Collective Purchasing of Health-Care Inputs

In the first notion of collective purchasing, health-care providers cooperate in respect of their purchasing of medical supplies. Nollet and Beaulieu (2005) provide a useful overview of these arrangements between health-care providers, which often mirror those that arise in other settings. The central idea is that a number of independent organizations, or more colloquially firms, agree amongst themselves to negotiate collectively with the suppliers of their inputs.

Advantages and Disadvantages

The motivation for such arrangements is primarily seen as being to reduce costs, by some combination of negotiating a lower price, reducing administration, or economizing on utilization. Studies of such collective purchasers typically report that they achieve price reductions in the order of 10–15%. Economists would argue that this is probably a consequence of the purchasing collective representing countervailing monopoly power and thus reducing the economic rents of their suppliers. Reductions in administrative costs will result from conventional sources such as economies of scale and scope and consolidation of the purchasing function. Some studies, for example, Schneller, 2000, report savings of as much as 40% in this respect but it is not clear that all costs are being recorded. Exactly how a purchasing collective might reduce utilization of inputs is less clear. One idea is that the collective standardizes its purchases and thus avoids unnecessary duplication of inputs. It is difficult to obtain hard evidence of this in practice and it should be noted that standardization requires coordination but not necessarily cooperative purchasing. In terms of problems of collective purchasing, the usually cited limitations of these arrangements are the problems of reflecting the potential diverse objectives of the members of the collective and the possible antitrust implications of collective action. As suggested above there are a number of sources of further reading on this use of collective purchasing in health care and it corresponds to a broad literature on supply chain management.

Collective Purchasing of Health Insurance

The second notion of collective purchasing arises specifically in the US health-care sector and originates from a system in which health insurance is often provided as a part of employment. Small employers who have to purchase health insurance on behalf of their employees may be at a disadvantage relative to larger employers in terms of dealing with the providers of health insurance. By forming health insurance purchasing cooperatives they might be able to redress this disadvantage. Wicks (2002) provides a good starting point for further reading in respect of these arrangements; their purported advantages and their potential problems. More recently the term health insurance purchasing cooperative has also been applied to any collective of individuals, as distinct from companies, seeking to purchase health insurance as a group. Moreover, there is contemporary policy debate concerning whether such arrangements can increase the coverage of health insurance.

Advantages and Caveats

If employee benefits are considered to be simply another input into production, then this second notion of the term collective purchasing is very closely related to the first use described in Section Collective Purchasing of Health-Care Inputs. By cooperating, small employers may achieve a lower price or achieve scale economies in their purchase of health insurance coverage. The literature on supply chain management referred to above again provides the details. But it can be argued that health insurance is a sufficiently idiosyncratic 'input' that additional issues arise in terms of benefits of forming a collective. The most often discussed issue – and again Wicks (2002) is the best starting point for further investigation – is that of risk pooling. A purchasing cooperative may help to balance high- and low-risk individuals and thus achieve coverage for some employees who might otherwise be precluded by their high-risk premiums. This idea is, however, contentious. If health insurers can discriminate between high and low risks they have an incentive to offer better rates to the lower risk types. So if two employers, one with a high-risk group of employees and the other with a low-risk group of employee form a cooperative to purchase insurance, there is a good chance that the low-risk employer would be offered better terms outside of the purchasing cooperative; the purchasing cooperative will fail. In reviewing the evidence regarding the effect of health insurance purchasing cooperatives, Wicks (2002) draws attention to the greater choice that individuals are faced with when a purchasing cooperative is in place. This is an interesting contrast with the more usual outcome of collective purchasing – greater standardization.

 doi:10.1016/B978-0-12-375678-7.00811-7

Collective Purchasing of Health-Care Treatments: The Role of Insurers

Although the first two notions of collective purchasing described in Sections Collective Purchasing of Health-Care Inputs and Collective Purchasing of Health Insurance arise in particular jurisdictions or in particular institutional settings, the third notion is close to ubiquitous in health-care markets. Although physicians or health-care organizations are the supplier of care and individuals in need of that care are the recipients, for most individuals in most circumstances the terms under which their care is provided – how much will be paid for it under various scenarios – is determined as a part of an agreement entered into by their insurer with health-care providers. This concept of collective purchasing seems to have been first exposited in relation to public-health insurance by Evans (1987) but, as one will see, can equally be argued to apply increasingly to private insurance. To understand this notion of collective purchasing, and just how substantially the consumer of health care differs from the consumer of apples or pears, it is useful to start by reconsidering the usual concept of purchasing (demand) in economics. This supposes that there is a defined good or service, a price that is specified by the seller, and a consumer whose role it is to specify the quantity they wish to purchase. Almost none of this applies in health-care markets. The services that constitute health care are many and varied and patients are more interested in getting better than in receiving those services *per se*. Service is not well-defined up until delivery (treatment) and suppliers do not compete in the conventional sense of offering a known product at a given price. The quantity that a person wants is 'enough to make me better.' And pertinent to a discussion of collective purchasing, consumers seldom act unilaterally because health-care insurance often involves insurers reimbursing health-care suppliers directly. The topic of health insurance is a vast one and its emergence and growth in health-care provision a substantial area of study, but the interested reader can consult McGuire (2011) or Pauly (2011) for recent overviews. A crucial element of insurance is that it makes the insurer a third-party purchaser of health care and this element of health-care provision gives rise to a number of concerns, especially in terms of the lack of incentives that the recipients of services have to regulate or monitor suppliers. This is another substantial topic for which Stinchcombe (1984) and Enthoven (1994) provide an entry point for further reading.

Alternatives to Collective Purchasing under an Insurance Scheme

Following Section Collective Purchasing of Health-Care Treatments: The Role of Insurers, the intermediation of insurance seems to make collective purchasing of health-care treatments commonplace. That does not need to be the case; traditional arrangements termed indemnity insurance allow insured individuals free reign to choose their health-care supplier, with the insurer reimbursing, subject to rules regarding copayment, stop-losses, etc., the provider of treatment. However, increasingly fewer private insurance arrangements allow consumers to unrestrictedly choose their supplier, or permit suppliers to dictate the price of a service, preferring instead to manage the treatment pathway by selectively contracting with specific providers or even integrating providers into the organization through employment contracts. Under managed care arrangements, as described by Dranove (2000), Newhouse (2002), and Baker (2011), insurers enter into various arrangements with providers on behalf of their enrollees. This kind of management of treatment provision, where the insurer collectively purchases on behalf of their enrollees is, if anything, more prolific in the realm of public-health insurance which conditions treatment on contracts with health-care providers with terms and conditions set on behalf of all covered patients/consumers (Blomqvist, 2011). Thus collective purchasing and health insurance would seem to go hand in hand; insurers collectively purchase health care on behalf individuals and the extent of such arrangements can vary according to the number of consumers covered (from employees in a single company, to all members of the population of a region or even a nation) or the services covered (from a single health-care intervention to an integrated treatment system) and may encompass many different payment mechanisms (from fixed price per treatment item, to price per illness of a fee per patient).

Advantages and Caveats of Health-Care Insurance

A first approach to explaining the above phenomenon might be to consider the same motives for collective purchasing as described in the first two notions of that term described above; by seeking to purchase on behalf of a large population the insurer might be able to negotiate lower prices and save resources relative to what each individual would have to expend in dealing with their own provider. One key problem is that providers of health-care have informational advantages and third-party arrangements such as insurance mean that even the limited information that patients have may not available to the payer. This results in a lack of information, incentives, and buying power on the demand side of health-care markets. The result is effective monopoly power on the part of service suppliers and one interpretation of collective purchasing arrangements by insurers is that they provide some countervailing buyer power. In simple terms, a single patient, consumer, or even small insurer may be at the mercy of a health-care provider who dictates a high price; a purchasing collective may achieve a lower price. This mirrors the traditional role of collective purchasing in other contexts except that in health care a need for countervailing market power may by more pervasive; it is not only lack of competing suppliers that creates seller power, it is lack of buyer information.

The previous approach does not, however, recognize the very distinctive features of health-care provision regarding which a large literature has developed in health economics and which can begin to rationalize the third notion of collective purchasing much more convincingly. Elsewhere in this volume there are extensive discussions of agency, imperfect information, and transactions costs and the implications of these for health-care delivery and understanding these concepts is central to appreciating a long tradition in health economics focusing on the consequence of insurance in terms of the extent that

consumers who are insulated from cost will not have incentives to control the cost of their treatment. It thus becomes important for insurers to contain costs but, given the general lack of information that patients and consumers have it is also important to maintain incentives for a good quality of service.

In this setting, collective purchasing of health treatments becomes a method of dealing with multiple agency issues. One approach emphasizes selective contracting, the purchaser's decision about which providers to contract as suppliers. By limiting the set of suppliers, the purchaser generates bargaining power that may counteract market power of sellers or allow a buyer to influence the cost and/or quality of care. A second possibly complementary approach focuses on contract design. Rather than just negotiating on price, in their role as collective purchasers of health-care insurers may dictate the form of contract that the health care is provided under and thereby seek to influence, through the design of appropriate incentives the cost and quality of health care that patients receive. Viewed in this context a collective purchasing contract is a means of trying to align the incentives of health-care suppliers with those of the purchaser of health care. A great deal of attention has, for example, been directed at the question of whether a simple fixed-price arrangement as embodied in the Medicare Prospective Payment System in 1983, and much emulated since, can achieve both cost control and appropriate quality of care. A recent summary of the extensive adoption of such systems in Europe and the claims that are made in terms of cost control are documented in Brusse *et al.* (2011). This transition from purchasing through reimbursement of costs to determining an *ex ante* fixed price gives perhaps the best illustration of the potential of collective purchasing to effect change in a health-care system.

Agency theory also highlights how difficult it might be in practice to design good collective purchasing contracts for health care. As one problem is resolved so others may materialize. One concern that has developed is that while moving toward predetermined prices based on a particular characterization of a patients' medical condition (so called prospective price contracts) may drive down costs, it may also give rise to attempts to select easier to treat patients – cream skimming – or avoid expensive ones. Thus for a collective purchaser the design of appropriate contracts can be very complex matter.

Summary

In many areas of economic activity, purchasers find it in their interests to act collectively to get a 'good deal' from their supplier. Traditional explanations of collective purchasing rely on the concept of buyers achieving some monopsony power to offset the monopoly power of sellers, or on the achievement of scale economies in purchasing. These explanations apply equally well in health care in regard to supply chain management in health-care organizations such as hospitals, who cooperate to purchase medical supplies, and can be extended to understand health insurance purchasing cooperatives. The possible disadvantages of these arrangements are that they fail to correctly reflect the diversity of preferences of their constituent members, or that they may run foul of the law in terms of antitrust or anticompetitive practices. But there is another notion of collective purchasing in health care that is more prolific and requires a rather more involved explanation. Individual consumers of health care do not for the most part act unilaterally in dealing with a health-care supplier – insurers, both public and private, act as intermediaries and very often as the collective purchaser. This manifestation of collective purchasing is intricately linked with the prevalence of health insurance, which is an arrangement concerned with insulating individuals from the costs of their health care and where individuals are so insulated agency problems arises. Insurers may try and contain costs and ensure adequate quality by setting terms and conditions for the supply of health treatments and thus act as collective purchasers. These sorts of arrangements go under different names such as managed care or health-care contracts depending in part on whether they are instigated by private or public insurers, but they are in essence collective purchasing.

See also: Health Insurance in the United States, History of. Managed Care. Markets in Health Care. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Physician-Induced Demand

References

Baker, L. (2011). Managed care. In Glied, S. and Smith, P. C. (eds.) *The Oxford handbook of health economics*. Oxford: OUP.
Blomqvist, A. (2011). Public sector health care financing. In Glied, S. and Smith, P. C. (eds.) *The Oxford handbook of health economics*. Oxford: OUP.
Brusse, R., Geissler, A., Quentin, W. and Wiley, M. (2011). *Diagnosis-related groups in Europe*. Burr Ridge, IL: McGraw-Hill.
Dranove, D. (2000). *The economic evolution of American health care*. Princeton, NJ and Oxford: Princeton University Press.
Enthoven, A. C. (1994). On the ideal market structure for third-party purchasing of health care. *Social Science Medicine* **39**(10), 1413–1424.
Evans, R. G. (1987). Public health insurance: The collective purchase of individual care. *Health Policy* **7**, 115–134.
McGuire, T. G. (2011). Demand for health insurance. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *The handbook of health economics*, vol. 2, pp. 317–396. Boston, MA, USA: Harvard Medical School.
Nollet, J. and Beaulieu, M. (2005). Should an organization join a purchasing group? *Supply Chain Management an International Journal* **10**(1), 11–17.
Pauly, M. (2011). Insurance and the demand for medical care. In Glied, S. and Smith, P. C. (eds.) *The Oxford handbook of health economics*. Oxford: OUP.
Schotanus, F. and Telgen, J. (2007). Developing a typology of organizational forms of cooperative purchasing. *Journal of Purchasing & Supply Management* **13**, 53–68.
Stinchcombe, A. L. (1984). Third party buying: The trend and the consequences. *Social Forces* **62**(4), 861–884.
Tella, E. and Virolainen, V. M. (2005). Motives behind purchasing consortia International. *Journal Production Economics* **93–94**, 161–168.
Wicks, E. K. (2002). *Health insurance purchasing cooperatives*. The Commonwealth Fund. Dublin, Ireland: Economic and Social Research Institute.

Further Reading

Fraser Johnson, P. (1999). The pattern of evolution in public sector purchasing consortia. *International Journal of Logistics Research and Applications: A Leading Journal of Supply Chain Management* **2**(1), 57–73.
Mello, M. M., Studdert, D. M. and Brennan, T. A. (2003). The leapfrog standards: Ready to jump from market place to courtroom? *Health Affairs* **22**(2), 46–59.
Sullivan, S. (1993). Collective purchasing and competition in health care. *Health Policy Reform* **10**, 259–268.

Comparative Performance Evaluation: Quality

E Fichera, S Nikolova, and M Sutton, University of Manchester, Manchester, UK

Glossary

Agency relationship The relationship between an agent and a principal. Classically in health care, the role of a physician or other health professional in determining the patient's (or other client's) best interest and acting in a fashion consistent with it. The patient or client is the principal and the professional is the agent. More generally, the agent is anyone acting on behalf of a principal, usually because of asymmetry of information. In health care, other examples include health managers acting as agents for their principals such as owners of firms or ministers, regulators as agents for politically accountable ministers, ministers as agents for the electorate. In health care, the situation can become even more complicated by virtue of the facts, first, that the professional thereby has an important role in determining the demand for a service as well as its supply and, second, that doctors are expected (in many systems) to act not only for the 'patient' but also for 'society' in the form, say, of other patients or of an organization with wider societal responsibilities (like a managed health care organization), or taxpayers, or all potential patients. There can be much ambiguity, as in seeking to understand the agency relationships in overseas aid giving and management, and as in establishing the extent to which formal contracts can enhance efficiency.

Incentive contracts The contracts between insurers or other third party payers and the providers of health care that embody incentives and penalties (both usually financial) for failing to meet particular conditions.

Yardstick competition An industrial regulatory procedure under which the regulated price is set at the average of the estimated marginal costs of the firms in the industry.

Zeckhauser's dilemma A problem with incentive contracts when those who are incentivized to behave in particular ways cannot fully control the consequences of their actions. They then require compensation in some form to offset this increase in the risk they face of failure, which raises the cost of the contract relative to the benefits anticipated by the principal.

Introduction

Health care purchasers and regulators often make comparisons between providers on indicators of quality. In this article the rationale for such comparisons is described, the options for this form of monitoring are considered and how this type of evaluation has evolved over time is outlined. Then, using a recent example of a quality program that links financial rewards to comparative performance in the UK, the key issues with this kind of performance evaluation are highlighted.

Principal–Agent Problems

The health care sector is characterized by a series of agency relationships. Patients delegate decision-making to doctors and payers give responsibility for supplying health care to providers. This delegation of decision-making or provision would be unproblematic if there was symmetric information and identical objectives were shared between the parties. In reality, two general problems are suggested by the principal–agent analysis. First, the task itself (i.e., delivering health care) is only partially observable or verifiable. This is called the moral hazard or hidden action problem. Second, the agent's capabilities are unknown to the principal but are known to the agent before the parties enter into the contract. This may adversely affect the principal's payoff and is called the adverse selection or hidden information problem.

The solution adopted in practice is to use a set of performance indicators to measure the output of the agent. However, this is only a partial solution because the information problems persist when the correlation between such indicators and the agent's effort is noisy and determined by a random component that often varies across agents. The extent to which the agent is in control of such variation is also unknown to the principal. The principal must therefore design a contract or system of incentives that elicits a second-best outcome from the agent.

Problems with Incentive Contracts in Healthcare

It is often claimed that the design of incentive contracts is more difficult in the health care sector than in other sectors. This is particularly the case when the principal's problem of ensuring that the agent delivers a high quality service is considered. There are five problems that are germane:

1. One of the best known concerns about incentive contracts is the trade-off between incentives and risk (so-called Zeckhauser's dilemma). Theoretically, incentive contracts impose a risk on agents and risk-averse agents will require a higher mean level of compensation. This premium will increase with the riskiness of the environment. Although empirical research has not found convincing evidence that higher incentives are given in riskier environments, health care providers provide an uncertain output (the well-being of the patient) which is only partially dependent on their actions.
2. When multiple actions are substitutes, incentive schemes may cause diversion of effort. For instance, an incentive

scheme focused on observable indicators will induce health care providers to game the system and reduce their effort on unobservable dimensions.
3. Because patients differ in their expected health outcomes and the agent has more information on the expected health outcomes than the principal, the agent may engage in 'cherry-picking' of patients, providing care only to patients at low risk of adverse outcomes.
4. As health care provision often requires input from more than one agent, the aggregation of agents into groups (for example, hospital teams) for incentive contracts creates externalities. These externalities may be positive (through monitoring of effort by close peers) or negative (caused by free-riding on others' efforts).
5. Health care providers have a social role and are trained to adopt professional ethics. Their utility functions are typically assumed to contain an altruism component that values the benefits to their patients. Incentives have the danger of crowding-out this intrinsic motivation.

Information on Quality

Quality assessment in health care is bedeviled with measurement problems. The measurement of output, or more strictly the agent's effort in producing output, is particularly difficult. Quality can be measured in terms of the quality of inputs, processes, or outcomes. Input quality measurement, for example, would involve assessing the capabilities and training of the labor force, the standard of the capital facilities and equipment, and the input mix. Such an approach is often taken by health care regulators seeking to maintain a register of qualified providers. Process quality measurement, however, would involve assessing whether agents are performing actions that are most likely to generate good quality outputs. In health care, this might involve assessing whether providers are adhering to best-practice guidelines and offering patients effective treatment regimes. Finally, quality output measurement would focus on the benefits that have been achieved for patients, regardless of how they have been achieved. Such benefits should include gains in survival and quality of life and increasingly capture patients' experience of using health care services.

The difficulty for the principal is to know which type of quality measurement offers the most accurate information on the agents' efforts. Quality inputs are a necessary but not sufficient condition for quality outputs. When assessing the quality of processes, principals are frequently forced to rely on agents' reports of their processes. These may be deliberately misreported, or may be applied to the least-costly patients who may be less likely to gain substantial benefits. The main problem with direct measurement of the quality of the agent's outputs is that these are noisy signals of their effort because patient outcomes reflect historical events, the patient's own actions, and the actions of other agencies. These are largely unobservable and contain a substantial random element.

For these reasons, principals often adopt a portfolio of quality indicators across each of these levels. This reduces, but does not eliminate, the problems with each of the individual indicators. However, it generates new problems of how the agent's performance on each indicator should be aggregated to form an overall signal of their effort.

Comparative Performance Evaluation

Broadly speaking, incentive contracts can be classified into two types of performance measurements: (1) absolute and (2) comparative performance. Under absolute performance, the agent is set standards on performance measures that must be achieved, for example, 80% compliance with a care guideline. Under a comparative performance scheme, the agent's performance is benchmarked against a relative standard.

The relative standard in comparative performance evaluation can be set on two dimensions: time and reference group. The time dimension of comparative performance can be implemented in a static or in a dynamic setting (i.e., current or historical performance). The reference group dimension of comparative performance can be implemented across groups of agents within or between health care organizations. Although dynamic comparative performance may or may not be implemented across reference groups, static comparative performance is always relative to a reference group.

To set an absolute performance standard, the principal needs to have good information on the effort that the agent will need to make to reach that standard. Setting a relative standard based on the agent's own historical performance ensures that the agent improves quality (and thereby increases effort) period-on-period but can fall foul of secular trends and does not seek to induce equal effort across agents. Use of a static reference group benchmark isolates performance measurement from (common) secular trends, but relies on choice of an appropriate reference group and places the agent at higher risk.

If the reference group approach is selected, comparative evaluation can involve two broad types of comparisons against the other agents. It can involve comparison to the average (which is called benchmarking) or it can involve the construction of league tables (known as a rank-order tournament in the sport sector).

Benchmarking versus Rank-Order Tournaments

The primary purpose of relative performance evaluation is to mitigate the principal's imperfect information. However, comparative performance evaluation has a 'yardstick competition' effect as well as an information effect. Because rank-order tournaments will increase competition more than benchmarking, the latter is a lower-powered incentive whereas the former provides sharper incentives. Previous research has shown that wider variation in levels of performance will be induced by rank-order tournaments. The risk of such tournament-based incentives is that contestants who think they have little chance to earn a prize are not motivated by the scheme and wider variations in performance are created.

Comparative performance evaluation is optimal only when all agents face common challenges. When this is the case, the performance of one agent allows the principal to infer information about another agent's performance. However, if worse health conditions adversely affect performance and these are concentrated in specific areas, then these factors should be filtered out by comparing providers within the same area. However, an agent's rank-order within an area contains less information on the performance of an individual agent and will not generally represent an efficient use of information. Instead, aggregate measures like averages of similar organizations are more efficient because they provide sufficient information about common challenges.

Benchmarking is able to reduce the 'feedback' and the 'ratchet' effects of the reward mechanism. Feedback occurs whenever one agent's action affects the incentive scheme and thus changes the agent's own reward as well as the reward for other agents. As the number of agents affecting the overall standard is higher under benchmarking, the feedback effect will be lower than in the case of rank-order tournaments. The ratchet effect is in essence the dynamic counterpart of the feedback effect. Good agents may be better off by hiding or misreporting their 'true' performance for fear that the principal may raise the current target on the basis of past performance. Unless collusion between agents occurs, this gaming is mitigated by benchmarking.

More fundamentally, any judgment on which type of relative performance evaluation is most effective depends on the goals the principal is trying to achieve. The principal may be primarily concerned with maximizing efficiency or with minimizing inequity. If the principal is mainly concerned with increasing the efficiency of health care provision then they will seek to use comparative performance evaluation to increase the average level of performance and will likely adopt a rank-order tournament. Alternatively, the principal may be motivated by the distribution of agents' performance levels as they care most about equity of health care provision. In this case, they will seek to use comparative performance evaluation to close the gap between outstanding and poorly performing health care providers. In this case, the principal may be reluctant to use rank-order tournaments as this may increase the gap in performance between agents at the top and the bottom of the league.

The Development of Comparative Performance Evaluation

Comparative performance evaluation began as an informal exercise in the private sector and became more structured in the late 1970s in response to Japanese competition in the copier market. It typically took the form of rank-order tournaments as the extent of market competition was high.

More recently, benchmarking has been used in the public sector. For example, from April 1996 the Cabinet Office and HM Revenue and Customs in the UK have run a project, the Public Sector Benchmarking Service, to promote benchmarking and the exchange of good practice in the public sector.

Box 1 shows some key definitions of benchmarking. It highlights the competitive definition of benchmarking by the private company Xerox and the less competitive definition of benchmarking, focused on learning from comparisons, by the public sector.

These developments have been mirrored in the health care sector. Initially, governments in their roles as payers and regulators, made use of the availability of electronic information to give feedback to providers on their relative performance. These initiatives were frequently undertaken under the auspices of professional bodies and the focus was deliberately on information-sharing and supporting intrinsic motivation. Providers were often given data on their own performance and the performance of the average provider or their rank in the distribution of performance over anonymized providers.

Later, these data were deanonymized and sometimes publicly reported. This was viewed as a natural progression. Once providers were content that the information on their performance was accurately recorded and consistently collected across providers, the public could be reassured that quality in the public health care sector was consistently high.

However, when quality first became linked to penalties and rewards, it was typical to use absolute performance standards. The introduction of waiting time targets in the UK National Health Service (NHS), associated with stringent monitoring and strong personal penalties, for example, was enforced using absolute maximum standards. These were frequently criticized for distorting priorities and inducing gaming, though the empirical evidence on patient reprioritization is scant and previous research finds no support for gaming. Similarly, the introduction of highly powered financial incentives for UK general practices in the form of the Quality and Outcomes Framework were based on absolute standards. The lack of data on baseline performance meant that these standards were set too low and that only modest gains in quality were delivered, some of which have been shown to be due to gaming of the self-reported performance information.

The second generation of financial incentives for improving quality in the UK NHS make greater use of comparative performance evaluation. There are a number of national schemes that emphasize local flexibility and payment for quality improvement rather than achievement of absolute standards. The forerunner to these was introduced in one region in England

Box 1 Definitions of benchmarking

The Public Sector Benchmarking Service defines benchmarking as: 'Improving ourselves by learning from others.'

The Cabinet Office calls benchmarking: 'The process of comparing practices and performance levels between organizations (or divisions) to gain new insights and to identify opportunities for making continuous improvements.'

The European Benchmarking Code of Conduct states that: 'Benchmarking is simply about making comparisons with other organizations and then learning the lessons that those comparisons throw up'.

Xerox, a pioneer of private sector benchmarking in the copier market says that it is: 'The continuous process of measuring products, services and practices against the toughest competitors or those companies recognized as industry leaders.'

and provides a good example of the limitations of using financial incentives linked to comparative performance evaluation. This scheme is described in the next section.

The Advancing Quality Program

The Advancing Quality (AQ) program was launched in October 2008 for 24 acute hospital trusts in the North West of England. Trust performance is summarized by an aggregate measure of quality – the composite quality score – within each of five clinical domains. The five incentivized clinical conditions are acute myocardial infarction, coronary artery bypass graft surgery, hip and knee replacements, heart failure, and pneumonia. The composite quality scores are derived by equally weighing achievement on a range of quality metrics which include process and outcome measures. Table 1 lists the quality metrics used in AQ.

Table 1 Quality measures used in the advancing quality program

Patients with acute myocardial infarction
Aspirin at arrival
Aspirin prescribed at discharge
ACEI[a] or ARB[b] for LVSD[c]
Adult smoking cessation advice/counseling
Beta blocker prescribed at discharge
Beta blocker at arrival
Fibrinolytic therapy received within 30 min of hospital arrival
Primary PCI[d] received within 90 min of hospital arrival
Standardized survival index
Patients with heart failure
Evaluation of left ventricular function
ACEI or ARB for LVSD
Discharge instructions
Adult smoking cessation advice/counselling
Patients receiving coronary artery bypass grafting
Aspirin prescribed discharge
Prophylactic antibiotic received within 1 h before surgical incision
Prophylactic antibiotic selection for surgical patients
Prophylactic antibiotics discontinued within 48 h after surgery end time
Patients receiving hip and knee replacements
Prophylactic antibiotic received within 1 h before surgical incision
Prophylactic antibiotic selection for surgical patients
Prophylactic antibiotics discontinued within 48 h after surgery end time
Recommended venous thromboembolism prophylaxis ordered
Received appropriate venous thromboembolism prophylaxis within 24 h of surgery
Readmission avoidance rate – 28 days post discharge
Patients with pneumonia
Oxygenation assessment
Initial antibiotic selection for immunocompetent patients
Blood culture performed in A&E before initial antibiotics received in hospital
Initial antibiotic received within 6 h of hospital arrival
Adult smoking cessation advice/counseling

[a]Angiotension converting enzyme inhibitor.
[b]Anguitensin receptor blocker.
[c]Left ventricular systolic dysfunction.
[d]Percutaneous coronary intervention.

The AQ scheme is similar to the Hospital Quality Incentive Demonstration (HQID) in the US. Both schemes started as pure rank-order tournament systems. At the end of the first year, hospitals in the top quartile received a bonus payment equal to 4% of the revenue they received under the national tariff for the associated activity. For trusts in the second quartile, the bonus was 2% of the revenue. For the next two quarters, the reward system changed to the same structure that was adopted by HQID after 4 years; bonuses were earned by all hospitals performing above the median score from the previous year and hospitals could earn additional bonuses for improving their performance or achieving top or second quartile performance. There was no threat of penalties for the poorest performers at any stage.

Evidence from HQID and AQ initiatives suggests that providers quickly converge to similar values on the process metrics and differences in performance must be measured at a very high level of precision to discriminate among providers. In addition, on some of the process measures most providers scored (close to) maximum scores. Because of the small variability in the measures and these ceiling effects, the schemes end up rewarding trusts based on small differences in performance.

Under the HQID and AQ scoring mechanisms, all of the targeted indicators are given equal weight regardless of their underlying difficulty. Thus, the quality score methodology involves a risk that providers will divert effort away from more difficult tasks toward easier tasks. However, despite the clear incentive to do so, research from the US suggests no consistent evidence that providers engaged in such behavior.

From the perspective of public health and policy making, the more important question, however, is whether health outcomes have changed as a result of the introduction of HQID and AQ initiatives. Here, the US and UK experiences are contradictory. A comprehensive US study found no evidence that HQID had affected patient mortality or costs. The first evidence from the UK shows that the introduction of AQ initiative was associated with a clinically significant reduction in mortality.

In both countries, studies have found weak links between process measures and patient mortality and ruled out causal effects on the health outcome. This appears to show that improved performance on the process measures alone could not explain the association with reduced mortality in the North West.

The critical questions now are how and why AQ scheme was associated with robustly estimated mortality reductions when similar studies have found little evidence of an effect of process metrics on patient outcome.

The qualitative evaluation of the AQ scheme found that participating hospitals adopted a range of quality improvement strategies in response to the program. These included employing specialist nurses and developing new and/or improved data collection systems linked to regular feedback of performance to participating clinical teams.

Compared to HQID, the larger size and greater probability of earning bonuses in AQ may explain why hospitals made such substantial investments. The largest bonuses were 4% in AQ compared to 2% in HQID and the proportion of hospitals

earning the highest bonuses was 25% in AQ compared to 10% in HQID.

In addition, the participation process may be important. To participate in HQID, hospitals had to (1) be subscribers to Premier's quality-benchmarking database, (2) agree to participate, and (3) not withdraw from the scheme within 30 days of the results being announced. The 255 hospitals that participated represented just 5% of the total 4691 acute care hospitals across the US. In contrast, the English scheme was a regional initiative with participation of all NHS hospitals in the region. This eliminated the possibility of participation by a self-selected group that might already consist of high performers or be more motivated to improve. Further experiments would be required to identify whether pay for performance schemes are more effective when participation is mandatory or targeted at poor performers.

Despite the 'tournament' style of the program, staff from all AQ participating hospitals met face-to-face at regular intervals to share problems and learning, particularly in relation to pneumonia and heart failure, where compliance with clinical pathways presented particular challenges and where the largest mortality rate reduction can be found. Similar shared learning events were run as 'webinars' for HQID. The face to face communication, regional focus, and smaller size of the scheme in England may have made interaction at these events more productive.

The fact that a scheme that appeared similar to a US initiative was associated with different results in England reinforces the message from the rest of the literature that details of the implementation of incentive schemes and the context in which they are introduced have an important bearing on their effects.

Concluding Remarks

To summarize, the asymmetry of information between the principal and the agent is particularly acute in the case of information on quality. Principals design incentive contracts under these circumstances to induce agents to increase their effort. One way in which principals can retrieve information on the efforts being made by agents is through comparisons of performance across time and/or across agents. Such comparative performance evaluation can involve comparison to own historical achievements or a reference group's achievements. The principal can benchmark agents to the average or create a rank-order tournament.

Although both types of comparative performance evaluation can improve efficiency by reducing the principal's information problem, rank-order tournaments are more likely to increase the gap between performance at the top and the bottom of the league. Benchmarking minimizes feedback and ratchet effects, but it can also weaken competition between agents. Ultimately, the choice between benchmarking and rank-order tournaments depends on the objectives of the principal.

In practice, comparative performance evaluation for improving quality was used quite widely and with little controversy when it appealed only to intrinsic motivation. Linkage of comparative performance evaluation to financial rewards, however, has led to a sharper focus on its limitations. In this regard, the experiences with the HQID and AQ initiatives display many of the conundrums of using comparative performance evaluation. There is a great deal of uncertainty over, and little empirical evidence to support, the choice of comparator. The frequently adopted strategy of using a portfolio of indicators leads to problems of appropriately weighing the calculation of overall performance to avoid re-prioritization of effort. Finally, incentivization of improvements in the quality of processes reported by agents does not in itself lead to outcome improvements.

Overall, the evidence base on the effects of comparative performance evaluation is weak. Although there has been a great deal of (well-intentioned) experimentation, these initiatives have been adapted too frequently and have not been rigorously evaluated. Ultimately, the main challenges for principals considering the use of comparative performance evaluation are how to measure hospital quality, how to identify similar agents to make accurate comparisons, whether to appeal to extrinsic or intrinsic motivation, and how to devise and implement the pay-for-performance initiative given the context in which it is introduced.

See also: Competition on the Hospital Sector. Heterogeneity of Hospitals. Markets in Health Care. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs

Further Reading

Benabou, R. and Tirole, J. (2006). Incentives and prosocial behaviour. *American Economic Review* **96**(5), 1652–1678.

Burgess, S. and Metcalfe, P. (1999). Incentives in organisations: A selective overview of the literature with application to the public sector. *CMPO Working Paper Series No.00/16*.

Burgess, S. and Ratto, M. (2003). The role of incentives in the public sector: Issues and evidence. *Oxford Review of Economic Policy* **19**(2), 285–300.

Chalkley, M. (2006). Contracts, information and incentives in health care. In Jones A. M. (ed.) *The Elgar companion to health economics*, pp. 242–249. Cheltenham: Edward Elgar Publishing Ltd.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources* **37**(4), 696–727.

Frey, B. S. (1997). A constitution for knaves crowds out civic virtues. *Economic Journal* **107**, 1043–1053.

Glaeser, E. L. and Shleifer, A. (2001). Not-for-profit entrepreneurs. *Journal of Public Economics* **81**(1), 99–115.

Grout, P. A., Jenkins, A. and Propper, C. (2000). *Benchmarking and incentives in the NHS*. London: Office of Health Economics, BSC Print Ltd.

Holmström, B. and Milgrom, P. (1991). Multitask principal–agent analysis: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* **7**, 24–52. (Special Issue).

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* **89**(5), 841–864.

Lindenauer, P. K., Remus, D., Roman, S., et al. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine* **356**(5), 486–496.

Nicholas, L., Dimmick, J. and Iwashyna, T. (2010). Do hospitals alter patient care effort allocations under pay-for-performance. *Health Services Research* **45**(5 Pt 2), 1559–1569.

Prendergast, C. (2002). The tenuous trade-off between risk and incentives. *Journal of Political Economy* **110**(5), 1071–1102.

Propper, C. (1995). Agency and incentives in the NHS internal market. *Social Science and Medicine* **40**, 1683–1690.
Ryan, A. M. (2009). Effects of the premier hospital quality incentive demonstration on medicare patient mortality and cost. *Health Services Research* **44**(3), 821–842.
Ryan, A. M., Tomkins, C., Burgess, J. and Wallack, S. (2009). The relationship between performance on Medicare's process quality measures and mortality: Evidence of correlation, not causation. *Inquiry* **46**(3), 274–290.
Shleifer, A. (1985). A theory of yardstick competition. *RAND Journal of Economics* **16**(3), 319–327.
Siciliani, L. (2009). Paying for performance and motivation crowding out. *Economics Letters* **103**(2), 68–71.

Competition on the Hospital Sector

Z Cooper, Yale University, New Haven, CT, USA
A McGuire, LSE Health, London, UK

Glossary

Concentration The degree to which a given number of producers (in this case hospitals) share the total level of output (treatments) in a given geographical area.
30-Day AMI mortality A death from acute myocardial infarction (AMI) within 30 days of admission to hospital.
Hospital competition Hospital behavior that arises when hospitals are contesting for patients due to incentive mechanisms imposed by funding bodies. Hospitals might compete on the basis of lowering prices or increasing quality of care, or a combination of the two, to attract patients and funding. Quality competition under fixed prices currently is predominant.
Hospital prices The charges either set by the hospital or by the funder or the regulator for the treatments and other services provided. The level of hospital costs are one determinant of prices; other factors include the degree of competition, the level of unsecured costs (e.g., to compensate for teaching provision, charitable provision, or new innovation), and the type of financial return sought (e.g., whether the hospital is for-profit or not-for-profit).
Hospital quality The quality of service provision attained by a hospital. Quality may be judged across many different dimensions and measured in different ways (ranging, e.g., from in-hospital mortality rates to level of overall patient satisfaction).
Market power The ability of any individual producer to control a dimension of the market it operates in. Within the hospital sector market power is normally related to the degree to which a hospital is able to capture potential patients. The higher the concentration of patients treated within a given geographic region by a given hospital tends to form the basis of the measurement of hospital power.

Introduction

A range of specific policies designed to increase both patient choice and hospital competition has been introduced in, amongst other countries, England, Denmark, Sweden, Norway, and the Netherlands. A primary concern arising from such reforms is the effectiveness of hospital competition to provide improvements in quality, responsiveness, and efficiency. Theory would suggest that if hospital prices are not fixed but endogenously determined by the hospitals themselves, and quality is not easily observed or verifiable, then hospitals may react to increased competition for funds by offering lower quality at a given price, thus chiseling on quality, attracting higher volume and funding but producing lower quality output. Competition may be introduced, but it may not produce the desired effect.

Theory also suggests, however, that if prices are set exogenously increased competition will lead to higher quality, although, it has also been noted that, if provider preferences are sufficiently altruistic, high quality provision can also occur within a restricted competitive environment. Indeed, theoretically, if altruism is sufficiently high there may be a negative relationship between competition and quality provision. Thus examination of the incentive structures and the environment into which these are introduced is critical. This has been the subject of debate, at the core of which is the notion that, given a regime of fixed prices, hospitals will compete for patients and therefore revenue, through improving the quality of care offered. Fixed hospital prices are essentially associated with Diagnostic Related Group (DRG) prices for predefined case groupings. Those in favor of hospital competition argue that with fixed price competition for patients, efficiency and quality improve as hospitals increase their performance or risk losing their market share. Those against competition argue that such market-based reforms can destabilize hospitals, increase transaction costs, and possibly even harm patients.

This article examines the empirical evidence on patient choice and hospital competition to consider whether competition is associated with an improvement in hospital quality and patient outcomes. To do so, the general literature that considers hospital competition and quality is assessed. Before this examination of the literature however, the conceptual difficulties of measuring competition in this sector are discussed.

Issues in Measuring Competition

To assess the impact that hospital competition has on clinical quality there has to be an agreed definition of market power. The major challenge is the estimation of the size of the competitive market and the power exercised by individual hospitals. It is obvious that incorrect definition of the potential market would result in biased assessment of the impact of competition.

In product markets price relationships, in particular own-price and cross-price elasticities, may be examined to aid definition of the relevant market. In the hospital sector this is not relevant as prices, even if known, are highly regulated. Typically, investigators calculate hospital market size through concentrating on the definition of geographic area instead and do so in one of three ways. First, geographic market area may be defined as based on a fixed radius, defined by a largely arbitrary distance that creates a circular market of radius *r*. Investigators then calculate the degree of competition inside that market. Fixed radius measures have the possibility of both

overestimating and underestimating the actual size of the market. The shortcomings of such fixed radius measures is that they do not take account of potential demand when they estimate market size. As a result, the fixed radius measures may suffer from urban density bias and overestimate competition in urban areas. However, an advantage of this type of fixed radius market definition is that the market size tends not to be endogenous to any other factors, such as hospital quality.

A second option is to create a variable radius market where the radius *r* that dictates the size of the market varies according to preexisting referral patterns, actual patient flows, or hospital catchment areas. For instance, a variable radius *r* could be set at a length that captures the home addresses of 75% of patients at a particular hospital. Variable radius measures tend not to be as affected by urban density bias but some argue that, when the radius *r* that defines the size of the market is based on existing referral patters or hospital catchment areas, the market size they estimate may again be biased. For example, a high performing hospital may have a larger catchment area than a lower quality competitor.

A third option is to create a radius that varies according to travel distance. An example of a travel-based radius would be to define radius *r* as the distance that captures the hospitals within a 30-min travel time from a particular patient's home address. Market definitions based on existing referral patterns may be related to the real or perceived quality of local hospitals, but can suffer from referral patterns reflecting quality. Some argue that any estimates of competition that rely on actual patient flows may still be biased. Rather than using actual patient flows, predictions of patient flows to specific hospitals may be used to reduce this bias. Some studies have used predicted demand to estimate market size, based on travel distance for patients, arguing that their method mitigates the problems of traditional fixed and variable market measures of competition. However, in practice, sizes of markets defined using radii derived from travel distances tend to be highly correlated with the sizes of fixed radius markets. Because the two market definitions produce results, which are so closely correlated, they both tend to be affected by urban density bias. The key issue with both market definitions is that they require a largely arbitrary definition of the size of the market, such as 30 km for fixed measure and a 30-min travel time for time variable measure. Both market definitions may therefore either overestimate or underestimate the true size of the market depending on how the upper boundary of the market is set by researchers.

All three approaches have been applied to the hospital market; none is perfect. Each measure has its own strengths, weaknesses, and inherent bias. A practical approach in considering which method to employ is to assess the compatibility of the data with the various measures, to trade-off the inherent bias contained in each method by comparisons across a number of measures and to explore the use of instrumental variables to overcome any endogeniety.

General Evidence on the Relationship between Hospital Competition and Clinical Quality

The largest volume of literature assessing the relationship between hospital competition and quality comes from the USA (see Gaynor, 2006 for an overall review). The bulk of the existing US literature has investigated the relationship between competition, prices, and capacity and is rather out of date. There is a related small, but growing literature in the US that looks directly at the impact of hospital competition on clinical performance. A number of studies consider endogenous price environments and, unsurprisingly, the general finding with respect to the influence of increased competition on outcome quality is ambiguous.

A smaller number of recent studies on competition and quality tends to the conclusion that, under exogenously determined fixed-price competition, higher levels of competition generally lead to improvements in clinical performance. The bulk of this US literature on hospital competition and clinical quality examines the outcomes of Medicare beneficiaries and within the timeframe of these studies Medicare operated an exogenously determined DRG pricing scheme. Findings generally support a positive relationship between in-hospital mortality and increased hospital concentration (Kessler and McClellan (2000) is a prime example). One study found that competition was associated not only with improved outcomes in the Medicare population but also with more intensive treatment for sicker patients and less intensive treatment for healthier patients who needed less care.

The literature outside the US is smaller but supports the general findings. There is a growing, recent literature on hospital competition within the National Health Service (NHS) in England, for example. It is based on the introduction of a purchaser–provider split, where GP practices purchased secondary hospital care on behalf of their patients. As initially introduced, these reforms were said to have created an internal market in health care. They were based on various contractual arrangements. Hospital prices were generally not fixed and can therefore be assumed endogenous. There is a wide consensus that the internal market never created high-powered incentives for hospitals or developed a significant degree of competition. Notwithstanding this criticism, there is some evidence that prices fell during the internal market. One study also found that, during the initial phase of the internal market, higher competition was not associated with lower quality.

Examination of the impact of the NHS internal market on patient waiting times and length of stay for hip replacement from 1991 to 1994/5, using survival analysis to look at hospital level data during the internal market reform period, found that waiting times for hip replacements fell and so did patients' average length of stay. This study found that, after the internal market was introduced, patients were more likely to be transferred to another facility rather than remaining in the hospital where they had the surgery until they were ready to be discharged home.

The strongest evidence on the impact of hospital competition on patient quality in the NHS comes from a number of English studies. This article considers various aspects of increased competition on hospital quality. The dominant quality measure, 30-day AMI mortality, was chosen because, being tied to an emergency treatment and largely associated with in-hospital mortality, it is not easily manipulated by hospital admission policies. The mechanism through which AMI-mortality may be used as a proxy for general hospital quality is not always made explicit, but hinges on the

presumed correlation between the management of AMI treatment and wider hospital practices. One study of the impact of the internal market (presumed competitive) on hospital quality as it had been before 1999, i.e., a period before the fixing of hospital prices, used a 30-min drive time from ward centers as the competitiveness measure. Using hospital-level data and controlling for hospital and local area characteristics, it was found that the internal market led to a small but statistically significant increase in 30-day AMI mortality, the adopted measure of quality (Propper, 1996).

A further study (Propper *et al.*, 2008) used a longer time period to assess whether more competitive areas had higher or lower AMI mortality over the period 1991–1999. Once again this is a period of endogenously determined prices. Similar to the findings from their previous work, the report that higher competition during periods of competition was associated with higher AMI mortality, i.e., higher competition is associated with lower hospital quality in this dimension. They argue that it is not credible that hospitals deliberately sought to curtail quality in this manner – hospitals did not deliberately worsen 30-day AMI mortality. Rather it is suggested that as the internal market increased competitive pressures hospital resources were shifted from quality domains that were not fully observable and verifiable such as the impact of hospital care on health outcomes, to those, such as waiting times for elective procedures that were easily measured and were being targeted.

The introduction of DRG-type prices into the English NHS in 2005/06 fixed hospital tariffs at the same time as competition within the NHS was strengthened. Two recent studies have used difference-in-difference estimators to examine the impact of this increase in competition on hospital quality using 30-day AMI as the measure of hospital quality. Cooper *et al.* (2011) found that AMI mortality decreased more quickly for patients living in more competitive areas than that in less competitive areas. Specifically in the three-year period after the reforms were introduced, a one standard deviation increase in hospital competition was associated with approximately a 1% decrease in AMI mortality. Gaynor *et al.* (2010) found a similar impact of the increase in competition on hospital quality, again measured through 30-day AMI death rates, over the period 2003 to 2007. Both studies, therefore, find that increased competition under a fixed price regime within the English NHS over the period 2002–8 improved hospital quality even though a different aggregation of data and different methods are used.

There is also a small empirical literature that considers the impact of increased hospital competition on equity and patient access. The hypothesis is that competition may have a detrimental effect on equality of access for NHS patients. Waiting times for patients having an elective hip replacement, knee replacement and cataract repair over the period 1997 and 2007 in England seem to have generally decreased as competition increased, with the variation in waiting times for those procedures across socioeconomic groups also greatly reduced. Cookson *et al.* (2010) examined the impact of the internal market on equity, measured as the association between patient deprivation and hospital utilization. They compared competitive and noncompetitive areas, where competition was measured using a Herfindahl–Hirschman (HHI) index in a fixed radius market and also found that there was no evidence that competition had a worsening effect on socioeconomic health care inequality.

Conclusions

This short review has confirmed what was to be expected from theory: Under exogenous fixed-price regimes health care reforms, which increase competition among hospital providers, can lead to improved outcome of quality. There is not a large volume of empirical evidence that can be used to test this theoretical conclusion but what does exist is rather robust. The methods used tend to be similar and reliant on robust estimation procedures, including difference-in-difference estimation and large data sets. One criticism of these findings is that a large number of studies use a similar proxy measure of hospital quality: 30-day AMI mortality. There are justifiable reasons for the choice of this measure: It is associated with an emergency admissions and treatment, which is difficult to manipulate by the hospital providers. It is nonetheless a one-dimensional measure of quality and the generalizability of the empirical findings rest on a belief that there is a strong correlation between this dimension and other less verifiable dimensions of hospital quality. It is perhaps not too difficult to buy into the belief that if hospitals have good management structures all dimensions of quality will trend in a similar manner. Other empirical research has indeed found that hospitals with better overall management skills had lower mortality from AMI. Moreover, recent studies show that this measure of hospital quality (30-day AMI mortality) is indeed correlated with other hospital outcome measures. The policy implications appear clear that with a fixed price regime competition can be improving. That this is not found when prices are set endogenously is perhaps an unsurprising lesson.

See also: Comparative Performance Evaluation: Quality. Empirical Market Models. Evaluating Efficiency of a Health Care System in the Developed World. Heterogeneity of Hospitals. Markets in Health Care. Switching Costs in Competitive Health Insurance Markets. Theory of System Level Efficiency in Health Care

References

Cooper, Z. N., Gibbons, S., Jones, S. and Mcguire, A. (2011). Does hospital competition save lives? Evidence from the NHS patient choice reforms. *Economic Journal* **121**, *F228–F260.*

Gaynor, M. (2006). Competition and quality in health care markets. *Foundations and Trends in Microeconomics* **2**, 441–508.

Gaynor, M., Moreno-Serra, R. and Propper, C. (2010) Death by market power: Reform, competition and patient outcomes in the National Health Services. *CMPO Working Papers.* UK: University of Bristol.

Kessler, D. P. and Mcclellan, M. B. (2000). Is hospital competition socially wasteful? *Quarterly Journal of Economics* **115**, 577–615.

Propper, C. (1996). Market structure and prices: The responses of hospitals in the UK National Health Service to competition. *Journal of Public Economics* **61**, 307–335.

Propper, C., Burgess, S. and Gossage, D. (2008). Competition and quality: Evidence from the NHS internal market 1991–1996. *Economic Journal* **118**, 138–170.

Further Reading

Baker, L. C. (2001). Measuring competition in health care markets. *Health Services Research* **36**, 223–251.

Bloom, N., Propper, C., Seiler, S. and Van Reenan, J. (2010). The impact of competition on management quality: Evidence from public hospitals. *CEP Working Paper – 14 February 2010 Draft*. London, UK: Centre for Economic Policy, London School of Economics.

Breeke, K., Siciliani, L. and Straume, O. (2009). Hospital competition and quality with regulated prices. *CESinfo Working Paper – 2010*. Munich, Germany: Centre for Economic Studies and Ifo Institute for Economic Research (CESifo).

Cookson, R., Dusheiko, M., Hardman, G. and Martin, S. (2010). Competition and inequality: Evidence from the English National Health Service 1991–2001. *Journal of Public Administration Research and Theory* **20**, 181–205.

Cooper, Z. N., McGuire, A., Jones, S. and Le Grand, J. (2009). Equity, waiting times, and NHS reforms: Retrospective study. *British Medical Journal* **339**, b3264.

Gaynor, M. (2004). Competition and quality in hospital markets. What do we know? What don't we know? *Economie Publique* **15**, 3–40.

Kessler, D. P. and Geppert, J. J. (2005). The effects of competition on variation in the quality and cost of medical care. *Journal of Economics and Management Strategy* **14**, 575–589.

Le Grand, J. (2009). Choice and competition in publicly funded health care. *Health Economics, Policy and Law* **4**, 479–488.

Tay, A. (2003). Assessing competition in hospital care markets: The importance of accounting for quality differentiation. *Rand Journal of Economics* **34**, 786–814.

Cost Function Estimates

K Carey, Boston University School of Public Health, Boston, MA, USA

Glossary

Average cost Total cost divided by the rate of output.
Behavioral cost function A cost function that includes amongst its determinants not only the cost of inputs but also things such as length of stay, case-mix, and quality of care.
Cost function A mathematical relationship between the costs of inputs in the production process and the rate of output.
Economies of scale Also known as increasing returns to scale: The amount of resources used per unit of output falls at higher output rates.
Economies of scope Also known as 'scope effects.' Economies of scope enable a firm to produce several goods or services jointly more cheaply than producing them separately. The simultaneous production of hospital care and medical teaching is an example.
Fixed cost A cost that does not vary with output either because input prices are constant or because decision makers have decided not to vary the input in question. Few, if any, inputs are technically fixed in the sense of being unalterable.
Long run A period of time in which all inputs are treated as variable.
Marginal cost The additional cost incurred if the output rate is increased by a small amount.
Production function A technical relationship between 'inputs' and the maximum 'outputs' or 'outcomes' of any procedure or process. Also sometimes referred to as the 'technology matrix'. Thus a production function may relate the maximum number of patients that can be treated in a hospital over a period of time to a variety of input flows like doctor- and nurse-hours, and beds.
Short run A period of time in which one or more inputs are treated as fixed.
Stochastic frontier cost function An empirical method of estimating the maximum outputs obtainable from given resources and, hence, the degree to which actual operations fall short of the most efficient way or operating.

The Economic Cost Function: Foundations

Microeconomics contains a theoretically based framework that describes how an individual business enterprise chooses to optimize production and cost efficiency, given existing technologies and prices of inputs. Within this supply side structure, the production function models the relationship between outputs produced and inputs used in the process, and the cost function models the relationship between the production cost of different levels of output accounting for input prices. The two functions are related in the sense that the production function shows the various ways of combining inputs to produce outputs, given the state of technology, and the cost function shows how to do it at minimum cost. Given certain basic mathematical properties, a duality or one-to-one correspondence exists between a set of production possibilities and the respective minimum cost function. In modeling the provision of health care services, economists often prefer the cost function to the production function because input prices are plausibly assumed to be determined outside of the model of firm behavior, whereas the selection of inputs in the production process are not.

The cost function is a powerful tool in the econometric application of the theory of production. In health economics, the preponderance of cost function estimation studies have focused on the hospital, which lies at the nexus of health care services and is the foremost component of health care spending. A number of issues involved in cost function estimation in health care have been addressed in empirical studies of US hospital costs. The remainder of this article will highlight the key issues involved in cost function estimation largely in that context.

Approaches to Cost Function Estimation

Short-Run Versus Long-Run Cost Functions

In any cost function estimation, a fundamental determination facing the researcher is whether to adopt the short-run or the long-run perspective. The distinction lies in assumptions regarding the state of equilibrium, or whether the firm has set all its inputs at their cost-minimizing levels. A variable cost function assumes the short-run scenario in which a firm's capital costs are fixed, whereas a total cost function takes the long-run perspective, in which all costs are variable and inputs have been chosen such that total costs are minimized. In the short-run variable cost function specification, the dependent variable measuring costs does not include capital costs; however, the fixed measure of capital is included as an explanatory variable. In the long-run total cost function, the dependent variable includes capital costs.

The appropriate choice of the short-run versus the long-run approach draws on both theoretical and practical considerations. If the firms are believed to be employing all inputs at the cost minimizing levels, then the long-run total cost function is indicated by theory. However, if firms cannot adjust their capital stock quickly in response to changing output

levels or input prices, a short-run variable cost function is the preferable specification. From a practical perspective, estimation of a long-run cost function requires a measure of capital costs, which are often difficult to observe. In addition, the long-run cost function should include measures of all input prices including those of capital, which in most applications can be achieved only as rough approximations.

In hospital studies, it is generally agreed that capital stocks are adjusted over time horizons exceeding the periods of study included in most datasets. Moreover, the industry has experienced considerable organizational, regulatory, and demand side changes over recent decades. These factors, together with the challenges of measuring capital costs and capital input prices, generally have led economists to estimate short-run variable cost functions for hospital studies. This specification does require reliable measures of fixed inputs. It also assumes that those inputs are exogenous, or that hospitals do not have the opportunity to significantly adjust their physical plant size.

Structural Versus Behavioral Cost Functions

In pure theoretical form, costs are modeled solely as a function of output levels and prices of inputs, controlling for fixed inputs or capital in the case of the short-run variable cost function. However, in empirical applications, cost functions generally incorporate other observable factors that have been found both conceptually and empirically to account for significant variation in the costs of producing specific products or services. This is particularly important in the health services literature where such cost estimations are alternatively referred to as behavioral cost functions or hybrid cost functions as opposed to structural or pure theoretic cost functions.

In the hospital literature, variables included in behavioral cost functions may not have a particular role in the microeconomic theory of the firm, but they incorporate real world differences in hospitals and reflect patterns of variation found in actual hospital cost data. Typically, hospital cost functions contain a primary measure of output such as number of admissions, one or more measures of input prices, and a measure of fixed capital such as the number of beds or the amount of total fixed assets. Admissions alone do not capture variation in hospital output. Other product descriptor variables commonly included are average length of inpatient stay, a case-mix index that is usually based on the relative costliness of the diagnosis-related groups assigned to admitted Medicare patients, and the number of hospital outpatient visits.

Other key variables that have been demonstrated to account for variation in hospital costs and are often included as controls in the cost function are measures of local market competition, ownership status (for-profit, not-for-profit, or government), and the presence of a teaching mission. Market competition is often measured using a Herfindahl–Hirschman index of market concentration. The index, calculated as the sum of the squared market shares of individual firms competing in the same market, is a function of the number of competitors and the distribution of their relative market shares. Its values fall in the range of 0–1 where lower measures signify many hospitals competing within the market and higher measures indicate fewer hospitals. Teaching hospitals are more costly because of the extra resources involved in performing an educational, in addition to a therapeutic, mission. These costs are sometimes captured by a binary variable such as membership in the Council of Teaching Hospitals or alternatively by a continuous variable measuring the number of medical residents affiliated with the hospital.

Challenges in Cost Function Estimation

Measuring Output: The Multiproduct Cost Function

Health care provision is highly complex, and measuring the output of a firm that supplies health care services is often complicated. For example, a typical general hospital treats patients with a large number of diverse conditions using thousands of different medical procedures. Resource utilization for surgical inpatients is greater than for medical inpatients, and inpatients are more resource intensive than outpatients. In physician practices, office visits for established patients have cost implications that are unlike those driven by visits with new patients, emergency room visits, or hospital visits. Nursing homes provide distinct levels of care for their residents, and skilled nursing patient days have different cost implications than intermediate care or other patient days.

Most health care cost estimations rely on the multiproduct cost function (also referred to as the multiple output cost function), which defines the cost of producing more than one type of output assuming that all inputs are used efficiently. Incorporating more than a single output into the cost function adds realism to the model. The multiple output specification also allows for a richer set of theoretical constructs useful in applications of cost function results. However, greater output complexity also introduces additional challenges in capturing unit costs of production. These issues are discussed in further detail in the section on Average Costs.

Controlling for Quality

Microeconomic theory assumes that the firm minimizes cost in choosing inputs to the production process to produce outputs at a given level of quality. Although measurement of firm cost is generally straightforward and measures of output are usually feasible, the quality of health care service provision is multidimensional and difficult to quantify. Yet, it has long been established that if quality of service is not controlled in a cost function, biases result.

Variation in quality levels also complicates the theoretical modeling of health care cost. In the case of hospitals, high nurse staffing ratios, the extra resources required by teaching hospitals, sophisticated information systems, and/or innovative high technology services are cost increasing features that have been found to be associated with higher observed hospital quality. Yet, low quality also can be cost increasing if it is related to lapses leading to preventable adverse events or postoperative complications that require additional services. These dynamics are interrelated. For instance, higher nurse staffing levels and/or sophisticated information systems not only have a direct and positive impact on costs but also reduce the probability of expensive adverse events, thereby simultaneously having an

indirect effect that is cost reducing. Overall, the theoretical relationship between costs and quality is complex, consisting of the joint effects of many different factors operating simultaneously.

Quality of health care also has presented repeated problems of measurement and data availability. Consequently, many cost function studies have not included explicit quality measures, confounding the impact of cost containment policies. In the absence of observed measures, some hospital cost functions have incorporated unobserved quality by building on the economic theory or by exploiting the structure of the error term in regression models. Studies that have included observed quality controls have relied heavily on structural measures of hospital quality such as teaching activity. There is widespread agreement that quality of care tends to be higher in teaching hospitals, which have access to the newest technologies. Yet, patient satisfaction and continuity of care are often worse in teaching hospitals and reports of resident exhaustion not uncommon. Teaching *per se* also represents the specific hospital output of medical education so that teaching is at best a proxy variable for hospital quality. Other structural measures include the presence of high-technology services, board certification of staff, hospital accreditation, and registered nurses as a percentage of full-time nursing staff. Finally, process measures such as outpatient follow-up to inpatient care, or outcome measures such as readmission, mortality, or adverse event rates have been used as quality controls.

The Profit Maximization Assumption in Health Care

The empirically estimated cost function derives from a theoretical framework, which assumes that the firm's fundamental goal is profit maximization. However, it generally is agreed that producers of health care services are often motivated by other objectives. For-profit enterprises constitute a minority of general hospitals in developed countries, and a large percentage of nursing homes are nonprofit organizations. Although a number of theoretical models have been developed in order to explain the objectives of nonprofit firms in the health sector, the empirical cost function literature on hospitals does not find that ownership drives cost differences. Growing competition in the hospital industry may force nonprofit hospitals to behave much like for-profit hospitals to remain viable.

Useful Constructs

The magnitudes of coefficients on independent variables generated by the cost function are not in themselves meaningful. However, a number of constructs fundamental to the theory of the firm can be determined using the cost function estimates. Key measures include marginal cost, average cost, economies of scale, and economies of scope. These represent a highly constructive set of tools that frequently are used in cost function applications to research and policy.

Marginal Costs

Marginal cost is the increment in cost that occurs when the output produced is increased by one unit. More formally, it is the derivative of the total cost function with respect to output. Marginal costs are important because economic decisions are made at the margin. For example, the economic decision of a physician practice to expand or reduce a particular service in response to a change in fixed payment rates will depend on the marginal cost of producing that service.

Average Costs

Average cost is defined as the total cost of production divided by the number of output units. Although a conceptually simple construct, calculation of average costs is complicated in health care cost functions. Because of the multiproduct nature of production, it is difficult to describe output in a single utilization measure. The American Hospital Association Annual Survey Database contains measures of 'adjusted' discharges and patient days where these outputs are inflated by the ratio of total (inpatient plus outpatient) revenues to inpatient revenues. These measures are widely accepted and used in hospital cost function estimations; however, it is recognized that they are biased to the extent that hospitals cross-subsidize across inpatient and outpatient services. Although the ratio of costs rather than revenues would be a more accurate economic adjustment, separation of costs in this way is not generally available in hospital accounting systems.

Economies of Scale

Economies of scale refer to the notion that average cost falls as the firm expands. Conversely, diseconomies of scale occur when expansion incurs increasing average costs. From a technical standpoint, a measure of economies of scale is equivalent to the ratio of marginal to average costs. This is because if cost at the margin is lower than average cost, then average cost will fall with increased output.

In the multiproduct context, there are two distinct economies of scale concepts. Product specific economies of scale characterize the cost effects of expanding each output separately while holding production levels of other outputs constant. The alternative adaptation is ray scale economies, which assumes a proportional increase in cost resulting from a simultaneous proportional increase in all outputs. Either construct may be appropriate; the choice depends on the context involved in the specific analysis.

Economies of Scope

The nature of multiproduct cost functions also gives rise to the related concept of economies of scope. Typically, a health care enterprise will produce more than one product because sharing of resources generally means that it is cheaper to produce products together than to produce them separately. Economies of scope refer to the savings incurred as a result of joint production.

Functional Form of the Cost Function

The cost function is not derived from a specific production technology; hence, no particular functional form is called for

in estimation. Yet, because the functional form of the minimum cost function is unknown to the researcher, there is a risk of misspecification, in which the model may yield poor or even erroneous predictions. Some judgment is called for in selecting a functional form for the cost function, and the econometrician practices a degree of art as well as science in formulating the econometric model.

A variety of specifications are employed in practice. The most commonly used in the health industries is the translog, a 'flexible functional form,' which represents a local second-order Taylor approximation to any true differential function. The translog involves logarithmic transformation of the dependent and independent variables and includes squared terms as well as interactions among outputs as independent variables. An important drawback to the translog that estimates a large number of parameters is the problem of multicollinearity among its many terms so that some precision of the estimates is sacrificed for functional flexibility, a trade-off that may or may not be warranted depending on the size of the dataset being used and the objectives of the particular research question. The problem is exacerbated in multiproduct cost functions and increases with finer disaggregation of outputs.

An alternative to the translog that often has been adopted in hospital and nursing home studies is a model that is logarithmic in costs with cubic polynomials on output. Although less flexible than the translog, the cubic specification is consistent with the classic U-shaped average cost function. It is particularly useful when the focus of the research is on marginal effects. There are other functional forms that have been used to estimate hospital cost functions. Of particular mention are the generalized translog, which often is used for multiproduct cost functions in cases where an output takes a value of 0 for some firms, and the generalized Leontief, which is useful in studies where the determination of input substitutability is of particular interest.

Some Applications

Health economists have used the cost function approach to address an extensive array of research questions. A description of the full range is beyond the scope of this narrative. However, this section highlights several notable issues that have been explored using cost function estimates. The purpose is to provide insight into the usefulness of the cost function approach in addressing important health policy concerns.

An economic question that lies at the core of the theory of the firm is optimization of firm size and the related issue of scale economies. The importance of economies of scale as a determinant of industry structure underlies economic arguments that have been put forth as justification for various forms of hospital regulation. A wave of hospital mergers in the 1980s and 1990s, for example, led the US federal antitrust authorities to develop guidelines for hospital mergers that allowed for demonstration of economic efficiency stemming from economies of scale. Economists have used the cost function to estimate the optimal hospital size, measured in patient days, or alternatively in number of beds. More recent policy concern has been over rapid growth of small physician-owned specialty hospitals. The economic cost function approach has been used to address the question of whether these hospitals are large enough to capture scale and scope efficiencies.

The cost function approach also has been applied to changes occurring in the internal organization of hospitals over the past two decades. Steep declines in the length of hospital inpatient stays began in the 1980s in response to insurer and government payer pressures on hospitals to absorb greater financial risk in their treatment decisions. The cost function has been used to examine the marginal cost of patient days over the course of a hospital stay. If the marginal cost of a patient day is relatively small, because the patient is in the recuperation stage and resource utilization is relatively low, then shortening the stay may or may not be an effective cost containment strategy.

An interesting policy question relating to the production of physician services is whether physician payments reflect marginal costs. For example, the Resource-Based Relative Value System through which US physicians are paid under the Medicare system was designed to reimburse at cost; however, the formulae used by Medicare is based on accounting cost systems that may not accurately reflect true production costs. A multiple output physician cost function is a tool that can more accurately reveal how marginal costs of production vary across different physician services that may be reimbursed at the same rate under administered pricing or privately negotiated rates.

The multiproduct cost function is well suited to empirical analysis of the US nursing home industry, which serves residents under explicitly distinct payment mechanisms: Rates received for Medicaid patients covered under various state programs for the poor are known to be considerably lower than those charged to self-paying patients. The cost function is a useful tool for exploring a number of policy questions. Are Medicaid rates paid by states to nursing homes for providing care for their poor elderly populations equal to the cost of treatment? Conversely, do higher rates charged to self-paying patients cross-subsidize Medicaid patients?

Stochastic Frontier Cost Function Estimation: Measuring Inefficiency

As expenditures on health care in developed countries have mounted in recent years, the goal of improving efficiency in health care provision has become a central objective for policy makers. At the same time, the demand for improved capability in measuring provider performance has stimulated the development of frontier analysis, which generates empirically based inefficiency measures at the provider level. Frontier studies define inefficiency as the extent to which an organization's performance exceeds the optimum (or frontier) as predicted by either production function or cost function estimates.

Within this empirical framework, the stochastic frontier cost function is the principal econometric technique for identifying the cost inefficiency of an individual provider. In contrast to a typical cost function that fits the average level that best fits the data, the stochastic frontier cost function traces out the least cost locus econometrically for varying output

levels and in that sense is more consistent with the theoretical concept of cost minimization. Inefficiency is inherently unobservable and assumed to be absorbed in the residual term. Allowing for unobserved firm-specific random shocks, the technique identifies an inefficiency term according to the deviation of the firm's actual cost to the least possible cost as determined by the cost function. Focus on the inefficiency term in stochastic frontier cost function analysis differs from traditional cost function analysis, in which interest is centered on estimated coefficients. In examining the performance of hospitals over the past decade, stochastic frontier analysis has been more prevalent in the literature than traditional cost function estimation.

A particular challenge for stochastic frontier cost function estimation is the ongoing difficulty in adequately controlling for quality. In hospital studies, for example, if quality is cost increasing overall, failure to account for it will result in confounding the inefficiency measures because it is not possible to differentiate between higher residual costs resulting from unobserved superior quality and higher costs resulting from managerial inefficiency or slack.

Further Reading

Aletras, V. H. (1999). A comparison of hospital scale effects in short-run and long-run cost functions. *Health Economics* **8**, 521–530.

Carey, K. (2000). Hospital cost containment and length of stay: An econometric analysis. *Southern Economic Journal* **67**, 363–380.

Carey, K. and Burgess, J. F. (1999). On measuring the hospital cost/quality trade-off. *Health Economics* **8**, 509–520.

Carey, K., Burgess, J. F. and Young, G. J. (2008). Specialty and full service hospitals: A comparative cost analysis. *Health Services Research* **43**, 1869–1887.

Carey, K. and Stefos, T. (2011). Controlling for quality in the hospital cost function. *Health Care Management Science* **14**, 125–134.

Escarce, J. and Pauly, M. V. (1998). Physician opportunity costs in physician practice cost functions. *Journal of Health Economics* **17**, 128–151.

Harrison, T. D. (2011). Do mergers really reduce costs? Evidence from hospitals. *Economic Inquiry* **49**, 1054–1069.

Rosko, M. D. and Mutter, R. L. (2008). Stochastic frontier analysis of hospital inefficiency: A review of empirical issues and an assessment of robustness. *Medical Care Research and Review* **65**, 131–166.

Troyer, J. L. (2000). Cross-subsidization in nursing homes: Explaining rate differentials among payer types. *Southern Economic Journal* **68**, 750–773.

Vita, M. G. (1990). Exploring hospital production relationships with flexible functional forms. *Journal of Health Economics* **9**, 1–21.

Cost Shifting

MA Morrisey, University of Alabama at Birmingham, Birmingham, AL, USA

Cost Shifting

Cost shifting exists when a hospital, physician group, or other provider raises prices for one set of buyers because it has lowered prices for some other buyer. The term has also been applied to managed care firms that are similarly said to have raised premiums for one set of purchasers because it had to lower premiums for some other set. Cost shifting is often confused with price discrimination. Health service providers commonly price discriminate; that is, they charge different prices from different payers. However, such differential pricing strategies are not evidence of cost shifting.

Cost shifting frequently enters into debates over government payment polices for Medicare and Medicaid and is prominent in health-care reform debates. Some have argued, for example, that efforts to reduce Medicare expenditures by lowering payments to hospitals under the Medicare Prospective Payment System or through the encouragement of Medicare managed care plans may save money for Medicare, but it will increase expenditures by private payers. This is said to occur because hospitals simply raise their prices to private insurers to make up the difference. Insurers, facing higher hospital prices, will then tell employers that they have to raise health insurance premiums because they are 'being cost-shifted against' by hospitals.

Analogously, proponents of health-care reform will often argue that systemwide reforms are needed because efforts to control government expenditures will simply increase private expenditures. It is argued that private payers should support coverage for the uninsured because the costs of the subsidy will be less than they appear because the hidden cost shift will be eliminated. Any piecemeal effort to control costs will ultimately be eroded by increases in costs for some other payer with the result that costs are not controlled. Subsidizing care for the uninsured and reforming the health-care system are important goals, but cost shifting is unlikely to be a serious component of the underlying economics.

The Economics of Cost Shifting

Morrisey (1994) used the Frank Capra movie *It's a Wonderful Life* as a vehicle to describe the economics of cost shifting. In the movie, Mr. Potter owned most of the town of Bedford Falls and he was the meanest man in town. He charged high rents on his apartments and high interest rates on loans from his bank. Suppose he also owned and operated Potter Hospital, the only hospital in town.

As a profit-maximizing old man, Potter would charge the most people would be willing to pay for each hospital day. He would determine the extra revenue and extra costs associated with each day of hospital care and produce the number of hospital days for which the extra revenue just equaled the extra costs. If he produced less, he was giving up profit he could have had; if he produced more, he would lose money because the extra cost was greater than the extra revenue.

Suppose Potter had two sets of hospital service buyers. The first set includes private purchasers who are willing to pay according to their downward-sloping demand curves. At lower prices they will buy more hospital days. The second set comprises government-sponsored patients who pay only the amount set by the government. They cannot pay more and the government will not pay less. To keep the story simple, suppose that each group of patients costs the same to treat and that marginal costs increase over the relevant range of output.

Potter faces two questions: first, should he provide any care to government-sponsored patients, and second, if so, what price should he charge private patients. The answers are straightforward business economics. The objective is to extract as much profit out of each market segment as possible. On the government side, he will admit patients until the extra revenue, the government fee, is just equal to the extra cost of care. On the private side, things are a bit more complicated. He can charge only a single price in this market. A lower price implies more units sold, but he can collect the lower price only from people who would have paid more. So Potter must find the price at which the extra revenue is just equal to the extra costs of treating these patients. And that extra revenue can be no lower than what he could get from a government-sponsored patient. The result of these calculations is that Potter will admit patients until the marginal revenue from private patients is equal to the marginal revenue from government-sponsored patients and is equal to the marginal cost of care.

This is shown graphically in **Figure 1**. The analysis is a simple case of price discrimination on the part of a monopolist with two buyers. The government price ($P_{Gov't}$) is fixed by government fiat and the hospital can get all the government patients it wants; thus, the government demand curve is also the government market marginal revenue. The private market yields a downward-sloping demand curve and its associated marginal revenue curve. The profit-maximizing hospital would (conceptually) trace out the envelope of the highest marginal revenue available from each market for every unit of service. This yields the kinked dark line that incorporates parts of each of the private and government marginal revenue lines in the figure.

Potter Hospital would produce hospital services to the point where marginal cost equals the envelope marginal revenue. That is, it would supply the quantity Q_T. Potter would sell the amount between Q^* and Q_T to the government because the marginal revenue from the government is greater than that from the private market. He would sell the private market the quantity from the origin to Q^* because over this range the marginal revenue from the private payers is greater than that offered by the government. Notice that, like a good monopolist, Potter charges the private market the most it will pay for the quantity up to Q^*. That is shown by the private demand curve with the price $P_{Private}$.

 doi:10.1016/B978-0-12-375678-7.00927-5

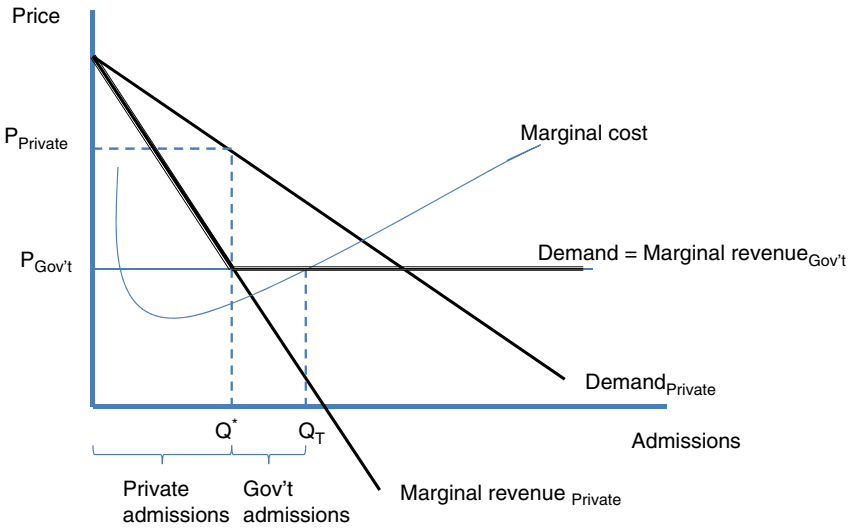


Figure 1 Monopoly price discrimination with two buyers.

Thus, because he has market power, Potter can charge different prices to different purchasers. This is classic price discrimination. A firm with market power will charge different prices to different purchasers as long as the purchasers have different degrees of price sensitivity and as long as one group cannot resell to the other. Thus, Potter charges a higher price to private purchasers (who have less price sensitivity) and a lower price to government-sponsored buyers (who are not allowed to pay even a dollar more than the government rate). Similarly, airlines charge higher prices to those who have to travel on specific dates and lower prices to those who have flexible schedules.

Now suppose the government lowers the price it will pay for hospital care. The cost-shifting argument says that Potter would accept the lower government price and 'make it up' by charging more in the private market. The economics imply the contrary. A lower government price signals that government patents are less profitable. Potter immediately sees that some private patients are willing to pay more than the new lower government rate. He shifts some hospital capacity to the private market. But to sell these services he has to lower the private price to everyone. Thus, government action lowering its price does not lead to higher private prices; rather lower private prices result as a profit-maximizing provider tries to shift capacity to the private segment of the market. Thus, standard theory indicates that cost shifting will not occur.

Graphically, the result is easily shown. See **Figure 2**, which adds a new lower government payment level to the earlier discussion. Note that the envelope of marginal revenue shifts down in its second segment. Potter Hospital reduces its total output from the old Q_T^1 to Q_T^2 to reflect the lower price available. The smaller quantity is now reallocated with more going to private patients and less to government-sponsored patients. However, the only way that Potter can sell the extra private services is to lower the price, as the figure indicates, from P^1_{Private} to P^2_{Private}.

Suppose the hospital were nonprofit and therefore did not 'maximize profits.' To see this, consider George Bailey from the *Wonderful Life* movie. He has a good heart and wants to help people. Suppose he ran the hospital in Bedford Falls. In particular, suppose George wanted to have the newest technology and to provide care to the indigent who are not eligible for the government care and cannot pay for private care. Note that if these things paid, Mr. Potter would have provided them as well.

If the hospital is to be all it can be, George has to generate as much 'surplus' as he can. Surplus, of course, is just another word for profits. The business problem is exactly the same for George Bailey as it was for Mr. Potter. If the hospital wants to provide as much charity care and new technology as it can, it must charge what the traffic will bear in each of its markets. The only difference between the two is how they spend the 'surplus.' Thus, when the government cut its price, George would shift capacity to the private market segment and lower its price as well. Potter ended up with fewer profits, and George Bailey ended up being able to provide less charity care and less new technology. Again, no cost shifting is predicted.

Cost shifting requires that a hospital or provider, more generally, raises its price for the private patient when the government price is reduced. This result can be consistent with standard economics, but it requires some special circumstances. First, the provider has to have market power. Without it, it cannot charge different prices. Second, it has to 'favor' paying patients. This means it has to charge them prices that are below the profit-maximizing price. Another way to say this is that the provider has to have 'unexploited market power.' Some commentators have described nonprofit hospital boards as not permitting charges to be set at levels above that needed to provide quality. This could be construed as favoring paying patients with prices below 'surplus maximizing' levels.

Under this scenario the hospital could be thought of as spending surpluses it could have had on lower prices to paying patients. Then, when the government lowers its price, the hospital has less surplus to subsidize its paying patients and raises its private price. This is cost shifting as envisioned by its proponents.

Several hypotheses emerge from this analysis. First, market power is a necessary condition for cost shifting. If health-care

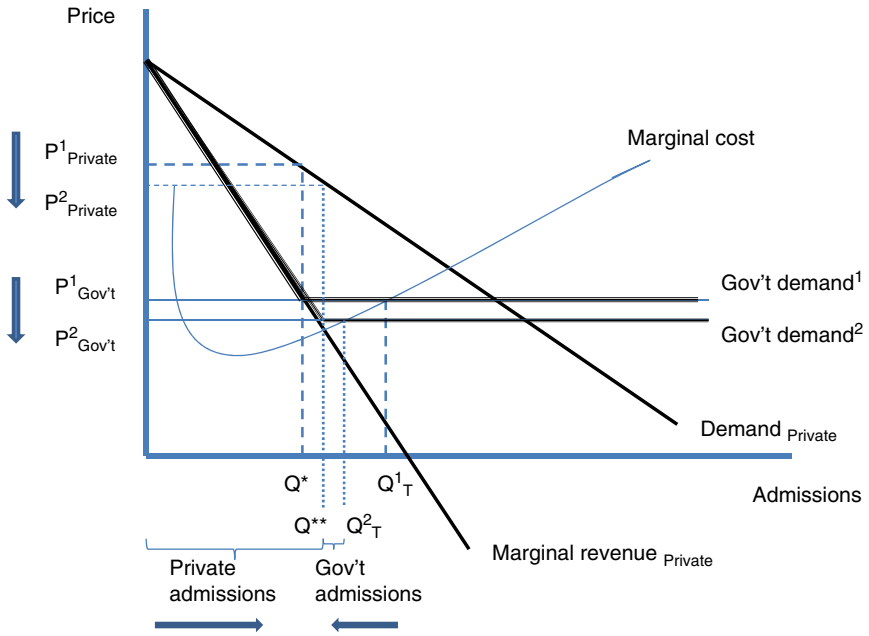


Figure 2 Effect of lowering of the government price.

markets are competitive, then cost shifting cannot exist because efforts to raise prices to one market segment would be thwarted by a willingness of others in the market to provide services at the old price.

Second, profit maximization implies no cost shifting. If a provider is indeed maximizing profits, by definition it has no unexploited market power. As a consequence, if investor-owned hospitals are profit maximizers, one would not expect to see them engaged in cost shifting.

Third, nonprofit status with market power by itself does not imply the ability to cost shift. The issue is the objectives of the organization. Cost shifting requires that the organization value setting prices to private patients at levels below those that would maximize profits.

Fourth, the model implies that cost-shifting behavior is limited. Once a provider exploits its unexploited market power, it has no further ability to cost shift.

Empirical Evidence on Cost Shifting

Ultimately the existence and magnitude of cost shifting is an empirical issue. The empirical evidence with respect to cost shifting has been mixed, but the rigorous research largely concludes that if it exists, its magnitude is modest at best. Unfortunately, much of the work simply misses the point because it seeks to show that different payers pay different prices for essentially the same services. This is true, but price discrimination is not cost shifting. Other work tries to use cross-sectional comparisons to test for the presence of cost shifting. This is difficult to achieve because cost shifting is a dynamic phenomenon. However, there have been five relatively recent papers that test for cost shifting using hospital behavior over time. See Morrisey (1994 and 1996) and Frakt (2011) for detailed reviews of the literature.

Hadley *et al.* (1996) used a national sample of hospitals over the 1987–89 period to examine the effects of financial pressure and competition on the change in hospital revenues, costs, and profitability, among other things. They found that hospitals with lower base-year profits increased costs less and increased their efficiency. With respect to cost shifting, "[w]e found no evidence to suggest that cost shifting strategies that might protect hospital revenues in the face of financial pressure were undertaken successfully" (Hadley *et al.*, 1996, p. 217).

It is also noteworthy that this study, and all of those reviewed here, control for hospital ownership status, but do not formally test for differences in behavior by ownership type. This is a lost opportunity. The exception is the work by Zwanziger *et al.* (2000).

Dranove and White (1998) used 1983 and 1992 California hospital data to examine the effects of reductions in Medicaid and Medicare volume on changes in price–cost margins (i.e., net price minus average costs all divided by net price) of privately insured patients in Medicaid-dependent hospitals. They found "no evidence that Medicaid-dependent hospitals raised prices to private patients in response to Medicaid (or Medicare) cutbacks; if anything, they lowered them" (p. 163). They also found that service levels fell for Medicaid (and Medicare) patients relative to privately insured patients and fell by more in Medicaid-dependent hospitals.

Zwanziger *et al.* (2000) used California hospital data from the same source over the full time period 1983 through 1991 and reached decidedly different conclusions. They computed the average price per discharge for Medicare, Medicaid, and privately insured patients. Controlling for average costs in a two-stage model, they found that lower Medicare and Medicaid prices were associated with higher private prices. A one percentage point decrease in the Medicare average price was estimated to increase private prices at nonprofit hospitals by 0.23–0.59 percentage points. The larger price increases were

found in markets with less hospital competition. They also found evidence that investor-owned facilities also engaged in cost shifting. Similar analysis by Zwanziger and Bamezai (2006) for the 1993–2001 period concluded that "cost shifting from Medicare and Medicaid to private payers accounted for 12.3 percent of the total increase in private payers' prices from 1997 to 2001" (p. 197).

Cutler (1998) examined whether lower Medicare payments led hospitals to greater cost cutting or cost shifting. Using data from Medicare cost reports over the 1885–1990 and 1990–95 periods, he found that in the early period, hospitals shifted costs dollar for dollar to private payers – an effect larger even than the Zwanziger *et al*. study. However, over the later period he found no evidence of cost shifting. Cutler attributes the difference in the results to the advent of selective contracting in the early 1990s that increased the extent of price competition among hospitals.

The most extensive analysis of cost shifting undertaken to date is that of Wu (2010). She uses Medicare data to examine the long period from 1996 to 2000 focusing on the effects of the effect of the Balanced Budget Act on Medicare hospital prices. Unlike earlier work, she treats the Medicare variable as endogenous. Wu finds that hospitals shifted approximately 21 cents of each Medicare dollar lost to private payers. Cost shifting varied by the bargaining power of the hospital. When a hospital had more power vis-à-vis insurers; it was able to shift more costs.

Conclusions

The most rigorous of the studies conducted in the past decade provide mixed evidence of the existence and magnitude of cost shifting in hospitals. Taken as a whole, the evidence does not support the claims of its proponents that cost shifting is a large and pervasive feature of US health-care markets. Only an early analysis by Cutler (1998) finds dollar-for-dollar increases in private prices as a result of lower Medicare payments. Even this finding is contained to a single short-run period. At best, one can argue that cost shifting, over the 15–20 years covered by the recent analyses, resulted in perhaps one-fifth of Medicare payment reductions being passed on to private payers. At worst, the majority of the rigorous studies found no evidence of cost shifting.

The theoretical literature strongly suggests that cost shifting can take place only if providers have unexploited market power. Once exploited, this avenue of response to changes in government payment policies disappears. This, together with the empirical findings, has three implications. First, policy advocates should worry much less about cost shifting. Although it can exist, other factors appear to be much more important in determining provider pricing. Second, the bulk of burden of reductions in government programs are borne by public patients. The consequences of such decisions cannot be shuffled off to private payers. Finally, health-care competition matters. One should look for evidence of cost shifting in markets and times that are characterized by provider concentration. If one is worried about cost shifting, encourage greater competition among hospitals, physicians, and insurers.

See also: Competition on the Hospital Sector. Managed Care. Markets in Health Care. Medicare. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment

References

Cutler, D. (1998). Cost shifting or cost cutting? The incidence of reduction in Medicare payments. *Tax Policy and the Economy* **12**, 1–27.

Dranove, D. and White, W. (1998). Medicaid-dependent hospitals and their patients: how have they fared? *Health Services Research* **33**(2), 163–185.

Frakt, A. B. (2011). How much do hospitals cost shift? A review of the evidence. *Milbank Quarterly* **89**(1), 90–130.

Hadley, J., Zuckerman, S. and Iezzoni, L. I. (1996). Financial pressure and competition: changes in hospital efficiency and cost-shifting behavior. *Medical Care* **34**(3), 205–219.

Morrisey, M. A. (1994). *Cost shifting: Separating evidence from rhetoric.* Washington, DC: AEI Press.

Morrisey, M. A. (1996). *Hospital cost shifting, a continuing debate. EBRI Issue Brief*, no. 180. Washington, DC: Employee Benefit Research Institute.

Wu, V. (2010). Hospital cost shifting revisited: New evidence from the Balanced Budget Act of 1997. *International Journal of Health Care Finance and Economics* **10**(1), 61–83.

Zwanziger, J. and Bamezai, A. (2006). Evidence of cost shifting in California hospitals. *Health Affairs* **25**(1), 197–203.

Zwanziger, J., Melnick, G. A. and Bamezai, A. (2000). Can cost shifting continue in a price competitive environment? *Health Economics* **9**(3), 211–225.

Cost-Effectiveness Modeling Using Health State Utility Values

R Ara and J Brazier, University of Sheffield, Sheffield, UK

Glossary

Cost-effectiveness analysis A method of comparing the opportunity costs of various alternative health or social care interventions having the same benefit in terms of a common unit of output, outcome, or other measure of accomplishment.

Cost-effectiveness threshold The maximum incremental cost-effectiveness ratio that is acceptable to a decision-maker. A rational community health-maximizing decision maker judges this threshold in terms of the health forgone elsewhere in the system if resources were to be devoted to one particular purpose rather than being available elsewhere in the system – the opportunity cost in terms of health.

Genomics The science of the function and structure of genomes, i.e., the DNA within a single cell of an organism.

Incremental cost-effectiveness ratio The ratio of the difference between the costs of two alternatives and its effectiveness or outcomes.

Meta-analysis A statistical technique for combining data from multiple studies used to identify the overall estimate of treatment effect.

Opportunity cost The value of a resource in its most highly valued alternative use. In a world of competitive markets, in which all goods are traded and where there are no market imperfections, opportunity cost is revealed by the prices of resources: the alternative costs forgone in order to pursue a certain action.

Introduction

There has been a growing use of quality-adjusted life-years (QALYs) in the assessment of the cost-effectiveness of healthcare interventions. There are now many agencies around the world using evidence on the incremental cost per QALY to inform reimbursement decisions or clinical guidelines. The QALY provides a metric for valuing the impact of healthcare interventions on survival and health-related quality of life (HRQL) on a common scale. It achieves this by assigning a utility value for each health state on a scale where 1 is for full health and 0 for dead, with the possibility of negative values for states regarded as worse than dead. There are many different ways for deriving such health state utility values (HSUVs). At the same time, there has been an increasing use of decision-analytic models to provide the main vehicle for conducting the assessment of cost-effectiveness. These overcome the limitations of relying on single clinical trials, which often do not use measures for generating HSUVs, have a limited sample size (particularly for some rare events), insufficient follow-up periods, an unrealistic protocol and setting, and may be difficult to generalize from. Models provide a means of combining evidence from a variety of sources on the clinical efficacy of the interventions, resource use, costs of resources, and HSUVs in a way that addresses the decision problem in a more relevant way than a clinical trial. HSUVs are a key parameter in such models. There is a separate article on the derivation of HSUVs and the different instruments used. This article is concerned with the methodological issues associated with using HSUVs in cost-effectiveness models.

There are many different types of models used to assess cost-effectiveness including decision trees, Markov models, and discrete event simulation. All seek to represent reality in terms of health states likely to be experienced by patients in the decision problem, transition probabilities between the states, and costs and utility value associated with each state. The states may be defined in different ways including whether a patient has a condition, severity of condition, key events (e.g., fractures in the case of osteoporosis), adverse events, and various comorbidities. These events may occur multiple times and there may be cases of multiple conditions. This article addresses four sets of methodological issues around the use of HSUVs to populate such cost-effectiveness models. (1) The selection of the measure for generating the HSUVs that best meets the requirements of policy makers and measurement criteria like validity. (2) The source of HSUV data such as the main clinical efficacy trials or whether to seek more relevant values for the model population from observational datasets, or to search, review, and synthesize an ever-growing literature. (3) Suitable utility data using the required measure may not be available from relevant studies, and in these cases regression techniques may be used to map from various health or clinical measures onto the selected utility measure. (4) Technical problems in using HSUVs in cost-effectiveness models, including how to adjust values over time, estimate values for those not in the condition of interest, and estimate the impact of conditions (comorbidities) or adverse events. This article considers the technical issues alongside the common requirements of policy makers around the world. Many of the decisions are not technical ones alone but involve normative judgments that in many cases will be made by policy makers requiring cost-effectiveness evidence. This is intended to be a practical guide aimed at analysts who are building cost-effectiveness models.

What Measures Should be Used?

There are four broad approaches for generating HSUVs: Generic preference-based measures (also known as multiattribute utility instrument), condition-specific preference-based measures, bespoke vignettes, and patient's own valuation. The most

 doi:10.1016/B978-0-12-375678-7.01411-5

widely used of these in recent years has been the generic preference-based measure of health. These measures have two components. One is a descriptive system that is composed of several multilevel dimensions. For example, the EQ-5D has five dimensions (mobility, self-care, usual activities, pain and discomfort, and anxiety and depression) each with three levels (a five-level version has recently been developed) and defines 243 health states. Each one of these states has a value on the QALY scale that was obtained by interviewing a sample of the general population. This descriptive system is usually completed by patients or their proxies in clinical studies and so provides a direct link between QALY estimates and the reported experiences of patients. By collecting EQ-5D or some other measures over time, it is possible to calculate the QALY gain in a trial setting (as the area under the curve) or to value states used in the model from observing patients in different clinical states.

These generic measures are designed for use in all conditions and patients. However, there are concerns that no one measure is sensitive or relevant to all conditions or patient groups. For this reason condition-specific preference-based measures have been developed by Brazier and colleagues. The problem with condition-specific measures is a concern with the lack of comparability between different instruments. This will be a problem where the model contains states from different conditions, as is often the case, and where the policymaker is making resource allocation decisions between conditions. Another approach has been to develop specific vignettes where there is no patient-reported information on the impact of a condition or its treatment. These vignettes can be specifically designed to describe the states in the model. However, in addition to the concern about comparability, vignettes do not have a direct link to evidence on patient experience that is achieved by the other two approaches, because they are not based on patient completion of a descriptive system but usually involve the views of experts (all be it informed by patient experience). The final approach avoids having to describe health states altogether and instead ask patients to value their own state using one of the preference elicitation techniques, such as time trade-off. Most agencies prefer health states to be described by patients and then valued by members of the general public, but one or two have specifically requested valuations directly from patients and this approach continues to be used.

A key problem is that these different approaches to valuing health produce different values. Indeed, different generic instruments have been shown to generate HSUVs that differ to a significant degree. The selection of instrument will have important implications for the incremental cost-effectiveness ratio. There is a literature on how to select the right measure in a given case, and this considers issues around the validity of the descriptive system for the condition, valuation methods, and source of the values. The decision about the right measure should not only consider these issues but will also be constrained in some cases by the policy makers to whom the model is going to be submitted. Some agencies have adopted a reference case that includes a preferred measure or approach. The most prescriptive has been the National Institute for Health and Clinical Excellence (NICE) in England who state a preference for the EQ-5D, and those submitting evidence need to demonstrate the EQ-5D is not appropriate in order to submit cost-effectiveness models using HSUVs from other measures. In some other countries, there is merely a preference for a generic measure. In others still, there is no preference expressed as to which type of measure should be used.

The final choice of measure used to derive the HSUVs will depend on some combination of the requirements of the policymaker, psychometric and other criteria, and also availability. In many cases, there is very limited evidence on HSUVs from a preferred measure or approach, and the analyst must make best use of available evidence. This may include the use of nonreference-based measures of health or clinical measures through the use of mapping (see Section Predicting Health State Utility Values When Preference-Based Data are Not Available). It will increasingly involve reviewing a range of possible sources including trials, observational and routine datasets, and the literature.

Source of Health State Utility Values

Clinical Trials

An appropriate source for the data on HSUVs may be the main clinical trial(s) used to inform the evidence on effectiveness. This enables the trial data to be used directly within the analysis of HRQL, eliminates concerns about the applicability of the health data to populations from which the effectiveness estimates are obtained and enables all the effects of treatment to be included directly in the estimate, including any side effects of treatment, without the need for adjustment. However, there may be concerns about the generalizability of effectiveness and/or HRQL data to the population in the model. There may be other circumstances where health state utility data are not best collected within the clinical trials, for example, if adverse events related to the condition or treatment are rare and not likely to be captured in the trials, or where the outcomes of interest are too long-term to be captured in a typical trial duration, or when the trial does not reflect common practice. In these circumstances observational studies may be more appropriate for capturing the impact of the event on HRQL.

Observational

HSUVs are often sourced from observational sources conducted for the purpose. Such tailored studies have the advantage of being designed for the purpose of populating a specific model and so can be designed to value the specific states defined in the model. However, this will often not be possible. Another data source is routine datasets such as general population health surveys (e.g., Medical Expenditure Panel Survey in USA and Health Survey for England in England) or routine surveys of patient-reported outcomes (e.g., the UK Patient Reported Outcome Measures program). For any observational source a key concern will be the extent to which HSUVs are caused by the condition. Patients who had a recent fracture, for example, have a lower score than those who do not. However, the differences found from cross-sectional observational studies tend to exaggerate the impact of hip

fracture because they often do not take into account their prefracture health status. As for evidence on efficacy, longitudinal evidence is better evidence than cross-sectional, as the impact of specific events or disease onset can be controlled for covariates.

Reviewing the Literature

There are published lists of HSUVs for a wide range of conditions and this literature is growing all the time. There is a risk that model builders will be tempted to use the first suitable value or even use those values that support the cost-effectiveness argument that is being made in a submission to a reimbursement authority. The larger the literature, the more prone the selection of values is to bias. For this reason, it is beholden on analysts to justify their selection of values. This implies a need for HSUVs, like other important model parameter values, be obtained from a systematic review of the literature in order to minimize bias and through appropriate synthesis of available values, capture the uncertainty, and improve the precision in the values used.

There are rarely the resources available to do a full systematic review in searching, reviewing, and synthesizing the evidence. Furthermore, reviewing HSUV studies is different from the conventional hierarchy of evidence used for clinical effectiveness. Simply looking for HSUVs from a search for efficacy evidence will fail to retrieve many, if not most, published HSUVs for the health states in the model because randomized controlled trial are often not the main important source for HSUVs and the models may include other conditions and adverse events. A model examining the cost-effectiveness of strategies for managing osteoporosis had states for various factures (e.g., hip, vertebra, and shoulder), breast cancer, coronary heart disease, and no event. A systematic literature review by Peasgood and others on the impact of osteoporosis fractures identified 27 articles from an initial set of 1000 papers reporting potentially relevant HSUVs for the model. As can be seen in Figure 1, there is a substantial difference in the HSUVs reported for the same time periods and although there is a trend for recovery following hip fracture, none achieve the prefracture values, and one study reported a decline in HSUVs over a period of 4–17 months.

The key considerations in searching and reviewing HSUVs are: (1) Do the HSUVs meet the methodological requirements of the policymaker – in the case of NICE, the focus may be on obtaining EQ-5D values (using the UK tariff of values), (2) have the HSUVs been obtained from a population relevant to the population in the model (e.g., in terms of severity of condition, age, and gender), and (3) what is the quality of the study including recruitment and response rates? These considerations do not operate in a dichotomous way because the analyst is looking for the best estimates and not necessarily the perfect ones, and these requirements may be relaxed depending on the available evidence base. Concerns about the relevance or quality of data should be fully explored in the cost-effectiveness model through the use of sensitivity analyses.

There are a number of search strategies for identifying HSUVs. However, a full search of the literature may yield many hundreds of values, and so the reviewer may wish to use more focused search strategies limited to identifying existing reviews or key papers and following up references in those articles, as described by Papaioannou.

For many conditions, there are a large number of HSUVs available in the literature and considerable variation in the values for what seem to be similar states. A review of values for use in a cost-effectiveness model of osteoporosis, for example, found values for hip fracture to vary from 0.28 to 0.72 and

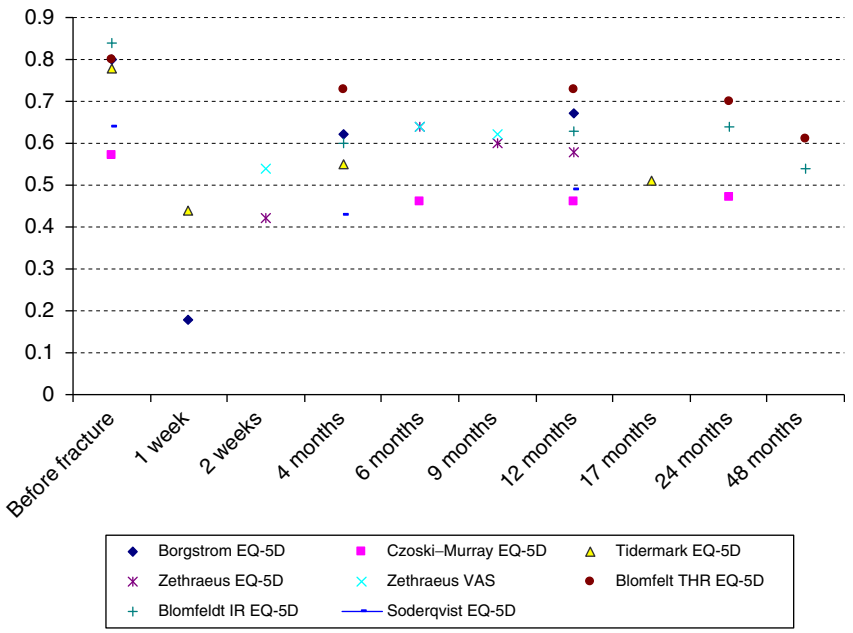


Figure 1 EQ-5D and EQ-VAS for hip fracture over time. Reproduced from Peasgood, T., Herrmann, K., Kanis, J. A. and Brazier, J. (2009). An updated systematic review of Health State Utility Values for osteoporosis related conditions. *Osteoporosis International* **20**, 853–868, with permission from Springer.

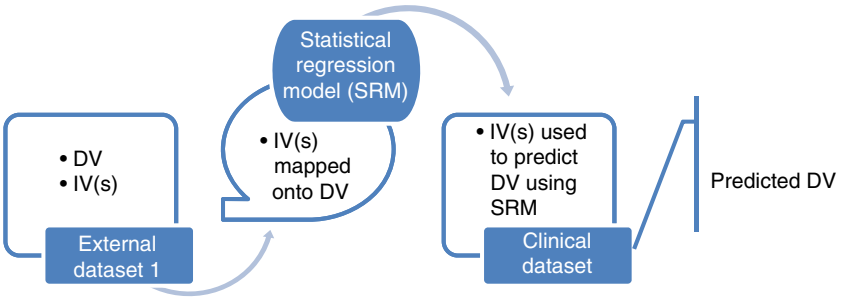


Figure 2 Mapping or crosswalking exercise.

vertebral fracture from 0.31 to 0.8. This leaves considerable scope for discretion in the selection of values for an economic model. The variation was partly due to differences in methods. In this example, the values were limited to EQ-5D for populating the cost-effectiveness model because the submission was for NICE. The values still varied considerably between studies. This may have been due to the different source countries, with much of the data coming from Sweden. It may also have been due to the very low response rate in some studies. There has been little research into the synthesis of HSUVs using techniques similar to those used for clinical efficacy including simple pooling or metaregression, but such work is at an early stage and the number of studies available for given conditions tend to be too small and heterogeneous. For this reason, current practice often involves selecting the study, which provides the most relevant values.

In practice, there may be little or no relevant HSUVs available for the cost-effectiveness model, but there may be trials or observational datasets that have collected HRQL or clinical data on relevant patients. The next section considers an increasingly used solution to this problem of mapping the relationship between the HRQL or clinical measure and the required preference-based measure.

Predicting Health State Utility Values When Preference-Based Data are Not Available

When the required preference-based utility measure is not collected in the clinical effectiveness studies or any relevant observational source, a mapping exercise can be undertaken to predict the required values (e.g., EQ-5D) from an alternative HRQL or clinical measure collected in the key study or studies. This exercise (Figure 2) requires an external dataset, which includes both the preferred preference-based data (the dependent variable (DV)) and at least one other variable (the independent variable (IV)) that is also available from the key clinical effectiveness or observational study. The data in the external dataset are used to obtain a statistical relationship, known as a statistical regression model, which can then be used to predict the required preference-based utility scores using the data available from the clinical effectiveness study.

The statistical regression model can take many different forms depending on the relationship between the variables and the underlying distributions of the data. The simplest model is a straight linear function ($y=\alpha+\beta x+\varepsilon$) where y is the DV (the preference-based HSUVs), α is the intercept, β is the vector of coefficients for the IVs, and ε is the error term. These regression models can be used to predict the DV in any datasets, which include the IVs. If some of the IVs are missing from the second dataset, the mean values from the external dataset used to obtain the statistical relationship can be used as proxies.

Using Clinical Variables and Progressive Conditions

Statistical regression models are also used to determine relationships between clinical variables and preference-based utility values when the cost-effectiveness models are driven by clinical variables, which represent stages or progression in the primary health condition. In these instances it may be that, although the clinical effectiveness study collects the required preference-based data, the distribution of patients across disease severity is such that the subgroup sizes are too small to determine HSUVs for each of the individual stages of the condition. For example, ankylosing spondylitis is a chronic progressive condition, and the severity of the condition is described using two clinical measures: the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Bath Ankylosing Spondylitis Functional Index (BASFI). Both measures range from 0 to 10, which represent no disease activity or functional impairment and maximum disease activity or functional impairment, respectively. Figure 3 shows how the preference-based utility values (the EQ-5D) vary by BASDAI and BASFI scores using the function: EQ-5D = 0.9235 − 0.0402 ∗ BASDAI − 0.0432 ∗ BASFI, which was obtained using ordinary least square regressions.

Figure 4 shows the BASDAI/BASFI profile (primary *y*-axis) and the corresponding EQ-5D values (secondary *y*-axis) plotted over time (*x*-axis) as would be used in a cost-effectiveness model. The figure shows individuals enter the model with average BASDAI/BASFI scores of seven units at baseline (time = 0). They initially respond to treatment, and their BASDAI/BASFI score improves to an average score of 4. After 4 years they stop responding to treatment and their BASDAI/BASFI scores revert to the baseline score of 7. These scores gradually worsen as the condition progresses until reaching the maximum possible score (BASDAI/BASFI equals 10) at 17 years. The BASDAI/BASFI scores remain at these levels until the patient dies (time = 26 years). Using the function described earlier to predict EQ-5D values from the BASDAI and BASFI scores, the predicted EQ-5D values are 0.241 (0.544, 0.241, − 0.062, and 0) at baseline (4–7 years, 17 years, 26 years, and after 26 years).

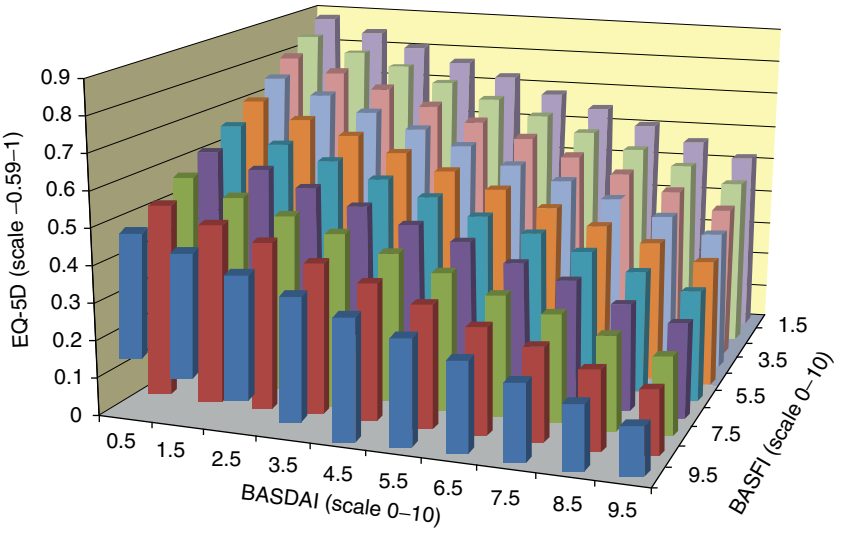


Figure 3 Plot of EQ-5D against BASDAI and BASFI.

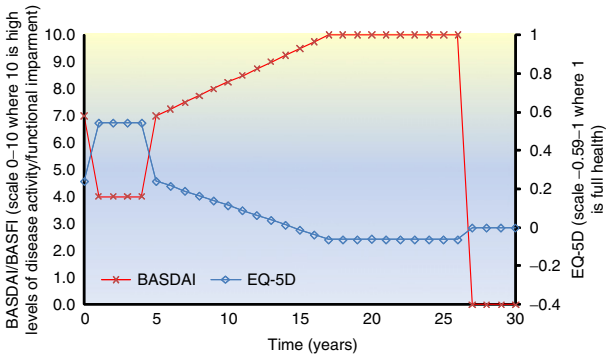


Figure 4 Patient's BASDAI/BASFI profile and associated EQ-5D scores over time.

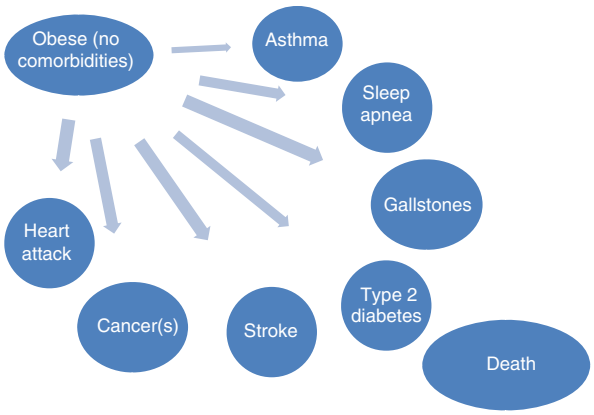


Figure 5 Possible health states in a cost-effectiveness model in obesity.

Multiple Health States

For a simple cost-effectiveness model involving few health states, the mean (and variance) preference-based utility values for each of the health states may be sufficient to describe the average HRQL and associated uncertainty for the health condition. However, when the cost-effectiveness model includes numerous distinct health states and additional predictors of health status, a statistical regression model and associated covariance matrix can be used to ensure correlations between preference-based utility values and are maintained when exploring uncertainty in the probabilistic sensitivity analyses.

One example is a cost-effectiveness model exploring the potential benefits of pharmaceutical interventions used to induce weight loss in obese patients. Obese patients are at increased risk of comorbidities (e.g., type 2 diabetes, cancer(s), heart attacks, strokes, etc. Figure 5) and the effectiveness of interventions are quantified in terms of changes in body mass index. To model this, analysts would need HSUVs for each of the comorbidities differentiated by body mass index and potentially age and/or gender. It is unlikely that this level of detailed information for each of the different subgroups would be available from clinical effectiveness studies. In this case, a statistical regression model obtained from a large external dataset could be used to predict the values required for each of the health states in the cost-effectiveness model.

Double Mapping

There are occasions when it is not possible to obtain an external dataset which includes both the required preference-based utility measure and one or more of the variables collected in the clinical study. In these instances, although not ideal, it is possible to obtain preference-based utility values using a process known as 'double mapping'. Double mapping involves the use of two external datasets and one statistical regression model obtained from each of these.

For example, in patients with psoriatic arthritis, a chronic progressive condition, the clinical study did not collect HRQL data but did collect information on demography (age, gender, and current and previous pharmaceutical treatments). In

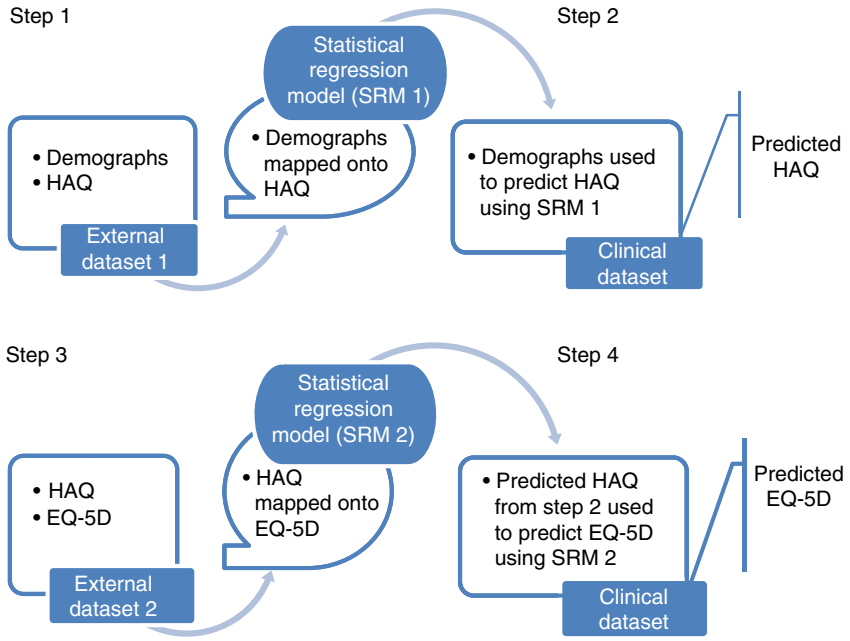


Figure 6 Double mapping exercise in psoriatic arthritis.

the cost-effectiveness model, the Health Assessment Questionnaire (HAQ: range 0–3, 3 = worse) was used to describe both the initial benefits of treatment and the long-term progression of the condition. Two external datasets were available (Figure 6). The first dataset (external dataset 1) had data on demography (age, gender, and current and previous pharmaceutical treatments) and HAQ but did not have any HRQL data. The second dataset (external dataset 2) had HAQ and the required preference-based data (EQ-5D) but did not have data on demography. The cost-effectiveness model required a relationship, which would link HRQL data to HAQ, the clinical variable, which would describe the benefits of treatment and long-term progression of the condition.

The process used to predict EQ-5D scores in the cost-effectiveness model is described in Figure 6. Step 1: External dataset 1 was used to obtain a statistical regression model 1 mapping demography (age, gender, and pharmaceutical treatments) onto HAQ. Step 2: The statistical regression model 1 was used to predict HAQ using the data on demography (age, gender, and pharmaceutical treatments) in the clinical study. Step 3: External dataset 2 was used to obtain the statistical regression model 2 mapping HAQ onto EQ-5D. Step 4: The predicted HAQ scores from the clinical study were used to predict EQ-5D in the cost-effectiveness model using the statistical regression model 2.

Predictive Ability

Ideally, any statistical model would be validated in an external dataset before use in a cost-effectiveness model. However, in the majority of cases, regressions are performed because the actual data are not available in a particular group, and therefore it is not possible to validate results in this way. Regression models, which have HRQL measures as the DV, typically underestimate and overestimate values at the top and bottom of the index, respectively. Consequently, it is important to demonstrate the accuracy in the predicted values across the full range of the index. If the objective of the regression is to obtain a model to predict values in cost-effectiveness model, then it is also useful to assess the ability of the regression model to predict incremental values accurately. The predicted values are typically assessed using the mean absolute error and root mean squared error. However, these summary scores can mask inaccuracies at the extremes of the index, and the predicted values should be assessed by subgrouping across the range of actual values.

Applying Health State Utility Values in Cost-Effectiveness Models

Baseline or Counterfactual Health States

Decision-analytic models in healthcare typically assess the benefits of interventions in terms of the incremental QALY gain associated with alleviating a health condition or avoiding a clinical event. To calculate this, in addition to requiring the HSUVs associated with the condition or event, the analyst will also need the baseline or counterfactual HSUVs to represent the HRQL associated with not having the particular health condition or event. For example, if modeling the benefits of introducing a screening program for breast cancer, analysts would require the mean HSUVs from a cohort with a history of breast cancer (including longer term data to model any potential changes in HRQL as the condition progresses) and the mean HSUVs for patients who do not have breast cancer. Similarly, when modeling an intervention that has the potential to avoid

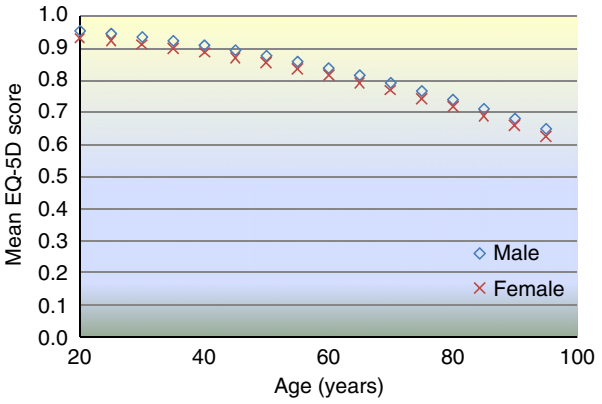


Figure 7 HSUVs by age and gender from the general population.

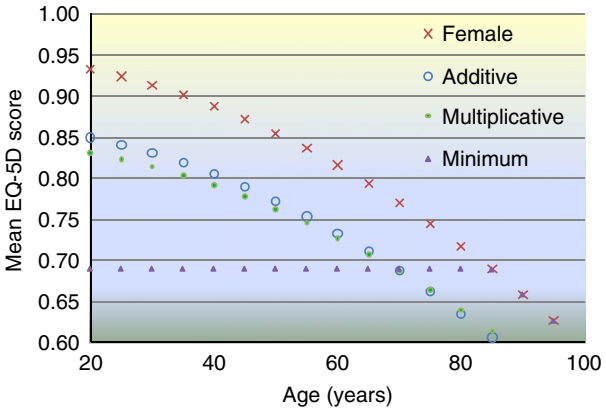


Figure 8 Combining HSUVs using the additive, multiplicative, and minimum methods.

subsequent cardiovascular events in patients with acute coronary syndrome, for example, a stroke, the analyst would need to know the mean HSUV for patients who have experienced a stroke and the mean HSUVs for individuals who have not experienced a stroke but have a history of acute coronary syndrome.

A patient without a particular condition is unlikely to have an HSUV of one. A better approach would be to use a normative dataset. Furthermore, the values of those with a condition are likely to change over time. HSUVs from the general population, for example, show a negative relationship with age (i.e., as age increases, the average HRQL decreases, Figure 7). This is due to several factors such as general decline in health directly attributable to age and an increase in prevalent health conditions, which are in general correlated with age. As many cost-effectiveness models use a lifetime horizon to accrue the costs and QALYs associated with interventions, it is reasonable to assume that the baseline or counterfactual HSUVs within the model may not remain constant over the full horizon modeled. Although there is a substantial volume of HSUVs in the literature describing the HRQL for specific health conditions, corresponding data for individuals without a specific health condition are more difficult to obtain without access to huge datasets. Unless the health condition is particularly prevalent, or unless it has a substantial effect on HRQL, removing a cohort who has a specific health condition will not have a substantial effect on the mean HSUVs obtained from the general population. In many instances, if the condition-specific baseline data are not available, it is possible to use data from the general population as proxy scores to represent the baseline or counterfactual HSUVs in the decision-analytic model.

Adjusting or Combining Health States

Healthcare decision-analytic models depict the typical clinical pathway followed by patients in normal clinical practice. As such they can become quite complex involving multiple health states, which represent the primary health condition with additional health states representing either comorbidities (e.g., when an additional condition exists concurrently alongside the primary condition), or an adverse event associated with the intervention or treatment (e.g., nausea is a side effect of treatments for cancer, whereas patients receiving aspirin for hypertension are at increased risk of hemorrhagic strokes). Ideally, each individual health state within a decision-analytic model would be populated with HSUVs obtained from cohorts with the exact condition defined by the health state. For example, it has been demonstrated that statins, which are typically given to manage cholesterol levels in patients with or at high risk of cardiovascular disease, have a beneficial effect on inflammation, thus may provide an additional benefit in patients with rheumatoid arthritis. To assess the benefits of statin treatment in a cohort with both cardiovascular disease and rheumatoid arthritis, the analyst would need HSUVs obtained from patients with both these conditions. However, many clinical effectiveness studies use very strict exclusion criteria relating to comorbidities and/or concurrent medications. As a consequence, the people who represent typical patients with comorbidities are excluded from studies, and analysts frequently combine the mean data obtained from cohorts with the single conditions to estimate the mean HSUVs for a cohort with more than one condition.

The methods used to combine the data can have a substantial effect on the results generated from decision-analytic models, and it has been shown that the result can vary to such an extent that they could potentially influence a policy decision, which is based on a cost per QALY threshold. There are a number of different ways to estimate the mean HSUV for the combined health condition using the mean HSUVs from the single health conditions. Traditional techniques include the additive, multiplicative, and minimum methods. The first two apply a constant absolute and relative effect respectively, whereas the latter ignores any additional effect on HRQL associated with the second health condition, using the minimum of the mean HSUVs obtained from cohorts with the single conditions as shown in Figure 8. Additional methods that have recently been tested include exploring the possibility of regressing the mean HSUVs from cohorts with single conditions onto the mean HSUVs from cohorts with comorbidities using ordinary least square regressions. Although this research is in its infancy, the early results look promising. However, based on the current evidence base, researchers recommend that the multiplicative method is used to estimate HSUVs for comorbidities, using an age-adjusted baseline as a minimum when calculating the multiplier used.

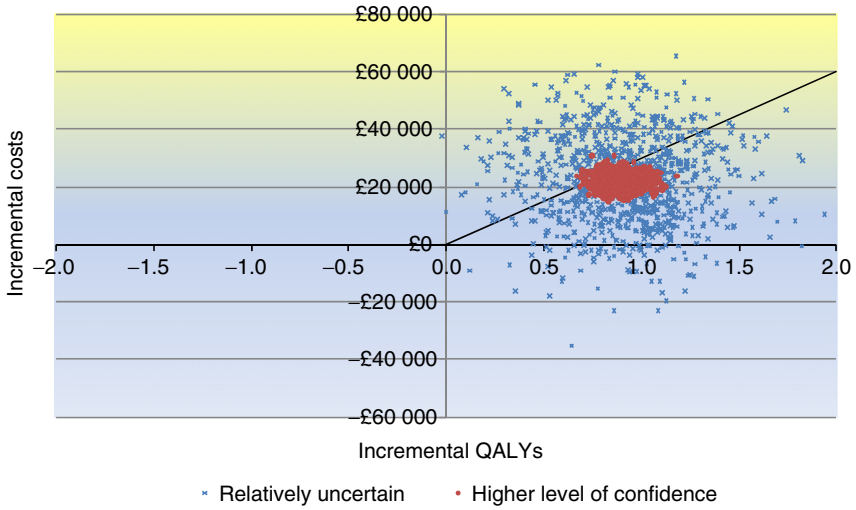


Figure 9 Cost-effectiveness scatter plane.

Worked Example

Females with condition A have a mean EQ-5D score of 0.69 and a mean age of 73 years, and females with condition B have a mean EQ-5D score of 0.70 and a mean age of 80 years. Using the data from the general population (Figure 8) as the baseline, these data are combined to determine what the EQ-5D score is for females with both condition A and condition B. Using data from the general population, at the age of 73 years and 80 years, the mean EQ-5D score for females is 0.7550 and 0.7177, respectively. The multipliers for conditions A and B are 0.9138 (=0.69/0.7550) and 0.9754 (=0.70/0.7177). The baseline data are then adjusted using these multipliers to estimate the age-adjusted EQ-5D score for the combined conditions A and B as shown in Figure 8.

Adverse Events

When considering adverse events for inclusion in cost-effectiveness models, it is important to distinguish between acute events and chronic sequelae. Although the inclusion of decrements on HRQL associated with grade 3–4 adverse events is particularly important, the cohort used for the main HSUVs may have included a proportion of patients who had experienced grade 1–2 adverse events and care should be taken to ensure these are not double counted. As in the preceding section, treating the decrement associated with the adverse event as a constant value may be inappropriate and based on the current evidence, the HSUVs should be multiplied (adjusting for age wherever possible) when combining these data.

Uncertainty

All results generated from cost-effectiveness models used to inform policy decision making in healthcare are subject to uncertainty. The uncertainty is examined and reported using sensitivity analyses. One-way sensitivity analysis is a procedure in which the central estimates for key parameters in the model are varied one at a time (generally using the 95% confidence intervals) and inform readers which variables drive the results generated by the model. Probabilistic sensitivity analysis is a method of varying all variables simultaneously to assess the overall uncertainty in the model. The individual Monte Carlo simulations (e.g., 5000) are generated using random numbers to sample from the distributions of the parameters. New results are generated by the model and each of the 5000 results stored. The recorded results are then used to illustrate the overall variability in the model results.

Figure 9 shows a scatter plot of the incremental costs (y-axis) and incremental QALYs (x-axis) generated from a cost-effectiveness model. The red points represent the individual results generated when there is relatively little uncertainty in the parameters used in the model. The blue symbols represent the individual results generated when there is considerable uncertainty and thus cover a broader area. The mean results (£24 500 per QALY) are the same in the results that are relatively uncertain and the results that are associated with a higher level of confidence. Using a cost per QALY threshold of £30 000 per QALY (the diagonal line), 41% of results from the model, which has a high level of uncertainty, are greater than this threshold, compared to just 7% of results from the model with a smaller level of uncertainty.

When looking at the uncertainty associated with the HSUVs, the distribution used to characterize the variables for the probabilistic sensitivity analyses should be chosen to represent the available evidence as opposed to selected arbitrarily. HRQL data, in particular the preference-based utility data, are generally not normally distributed. They are typically skewed, bimodal or trimodal, bounded by the limits of the preference-based index, and can involve negative values representing health states consider to be worse than death. Despite this, in the majority of decision-analytic models, the uncertainty in the mean HSUV can be adequately described by sampling from a normal distribution. Exceptions to this rule include when conducting patient-level simulation models using data from cohorts with wide variations in HSUVs and a relatively low or high mean value. In these cases an alternative approach would be to describe the

utility values as decrements from full health (i.e., 1 minus the HSUV) and then sample from a log normal or gamma distribution, which would give a sampled utility decrement on the interval $(0, \infty)$. If a lower constraint is required (i.e., -0.594 for the UK EQ-5D index), the standard beta distribution could be scaled upwards using a height parameter (λ) producing a distribution on a $(0, \lambda)$ scale.

Conclusions

The use of HSUVs in cost-effectiveness models has not received much attention in the literature. However, there are often no relevant HSUVs to be found in the literature, observational sources, or even trials. This article has provided practical guidance to those seeking to build cost-effectiveness models. In the near future, it is expected that there will be further developments in the field including methods of mapping, the synthesis for HSUVs across studies, and in the measures themselves. Policymaker's requirements may also change over time.

See also: Health and Its Value: Overview. Multiattribute Utility Instruments: Condition-Specific Versions. Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies

Further Reading

Ara, R. and Brazier, J. E. (2010). Populating an economic model with health state utility values: Moving toward better practice. *Value in Health* **13**(5), 509–518.

Ara, R. and Brazier, J. E. (2011). Using health state utility values from the general population to approximate baselines in decision analytic models when condition-specific data are not available. *Value in Health* **14**(4), 539–545.

Ara, R. and Wailoo, A. (2011). The use of health state utility values in decision models. Decision Support Unit, Technical Support Document 12. Available at: www.nicedsu.org.uk/Utilities-TSD-series(2391676).htm (accessed 01.02.13).

Brazier, J., Ratcliffe, J., Saloman, J. and Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation* (1st ed.). Oxford: Oxford University Press.

Brazier, J. E., Rowen, D., Tsuchiya, A., Yang, Y. and Young, T. (2011). The impact of adding an extra dimension to a preference-based measure. *Social Science and Medicine* **73**(2), 245–253.

Longworth, L. and Rowen, D. (2011). The use of mapping methods to estimate health state utility values. Decision Support Unit, Technical Support Document 10. Available at: http://www.nicedsu.org.uk/Technical-Support-Documents (1985314).htm (accessed 01.02.13).

Papaioannou, D., Brazier, J. and Paisley, S. (2010). The identification, review, and synthesis of health state utility values from the literature. Decision Support Unit, Technical Support Document 9. Available at: http://www.nicedsu.org.uk/ Technical-Support-Documents(1985314).htm (accessed 01.02.13).

Papaioannou, D., Brazier, J. and Parry, G. (2011). How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A Systematic Review. *Value in Health* **14**(6), 907–920.

Cost–Value Analysis

E Nord, Norwegian Institute of Public Health, Oslo, Norway, and The University of Oslo, Oslo, Norway

Introduction

Cost–value analysis (CVA) is a type of formal economic evaluation that can be used to inform decision makers in a public health service about the value to the public of different health technologies and what ought to be the public health service's maximum willingness to pay for them. In estimating value and limits to willingness to pay, CVA takes into account that in most countries with a public health service, citizens and societal decision makers hold concerns for both efficiency and equity. The concern for efficiency means that value – and thus willingness to pay – increases with the size of the health benefit provided by the technology – measured, for instance, in terms of the number of quality-adjusted life-years (QALYs) produced. Equity concerns may, for instance, mean that for a given health benefit, value and willingness to pay increase with the severity of the condition that is addressed. Other equity concerns may also be relevant (see History and Value Basis).

CVA has features in common with cost–utility analysis. Costs are estimated in the same way, and health benefits are expressed in QALYs. The difference is that concerns for equity are included in the determination of value. The replacement of the term 'utility' with the term 'value' in the name of the analysis serves to emphasize this difference. Whereas 'utility' refers to individuals' personal valuations, 'value' in 'cost–value analysis' refers to a broader societal concept. The basic premise of CVA is that simple aggregations of QALYs do not yield reliable estimates of citizen's overall valuation of different programs, because concerns for equity are not included in such simple aggregations.

Example

In CVA, the value of a program can either be expressed in equity-weighted QALYs (EQALYs) or in a public health care service's willingness to pay for QALY gains, given the context in which the gains occur and the characteristics of the patients who receive them. A simple example is as follows: Assume a scale of individual utility of health states from 0 to 1. Assume that intervention 'A' takes one type of patient from utility level 0.4 to level 0.6 for 1 year at a cost of EUR 10 000, whereas intervention 'B' takes another type of patient from level 0.8 to level 1.0 for 1 year at the same cost. The two interventions are equally cost effective (because the QALY gain and the cost is the same). But the societal appreciation (value) of the 0.2 QALYs in intervention A may be, say three times as high as that in intervention B, given the much greater severity of the pre-intervention condition in A and thus the much greater need in this type of patient. The cost–value ratio of intervention A would then be better than that of intervention B, namely $10\,000/(0.2 \times 3) = €16\,700$ EUR per EQALY versus $10\,000/0.2 = €50\,000$ EUR per EQALY, which suggests that A should be given priority among the two if a choice had to be made. To put it differently, it suggests that, in a society where such a concern for severity prevails, the public health care system should have a three times higher willingness to pay for intervention A than for intervention B, in spite of B producing the same amount of QALYs. CVA thus supports context-dependent, graded willingness to pay for QALYs.

History and Value Basis

The term 'cost–value analysis' was first introduced by Nord in 1993. It may be used in a general sense, that is, about any evaluation that takes into account relevant concerns for fairness (equity) in the weighting of individual benefits, whatever these concerns may be. However, in the development of CVA hitherto, some concerns have been treated as particularly salient. Based on a review in 1999 of existing materials in Australia, the Netherlands, New Zealand, Norway, Spain, Sweden, the UK, and the USA, Nord suggested that ethicists' and policy makers' reflections, and results from public preference measurements, seem to converge on the following points:

A. Society demands that medical interventions satisfy a minimum requirement of effectiveness for resource use to be justified.
B. Society's appreciation (valuation) of medical interventions increases strongly with increasing severity of the patients' condition. (This is often referred to as a 'concern for the worse off.')
C. Life saving or life extending procedures are particularly highly valued, and significantly more highly than interventions even for patients with severe chronic conditions.
D. When the minimum requirement of effectiveness is satisfied (see point A), society worries less about differences in the size of the health benefits provided by treatment programs for different patient groups, the underlying attitude being that people are entitled to realizing their potential for health, whether that be large or moderate, given the state-of-art in different areas of medicine.
E. As a special case of point D, society in most cases does not wish to discriminate between people with different potentials for health in decisions about life saving or life extension. For instance, society regards the prevention of premature death in people with chronic disease as equally worthy of funding as the prevention of premature death in otherwise healthy people. (Life extending interventions for people in vegetative states or states of very low subjectively perceived quality of life is an exception from this rule.)

Work on CVA hitherto has aimed at incorporating the specific ethical concerns above in formal valuation models (Nord *et al.*, 1999; Nord, 2001). The term 'cost–value analysis' is thus mostly used in this specific operational meaning rather than in the more general sense noted earlier.

Preference Measurements

To incorporate the above concerns in a numerical valuation model, data are needed on the strength of preferences for equity. The strength of societal concerns for severity and realization of potentials has been studied in samples of the general public in several ways. The most widely used technique is the person trade-off, which was introduced by Patrick, Bush, and Chen in 1973 under the name 'equivalence of numbers' and was given its present name by Nord in 1995. Typically, samples of the general public are presented with pairs of programs targeting two groups of patients that differ on one characteristic. The subjects are presented with numbers of beneficiaries in the two programs and asked to judge at what ratio between the numbers of beneficiaries they find the two programs equally worthy of funding. For instance, program A provides an improvement in utility from 0.4 to 0.6 for 100 people, whereas program B provides an improvement from 0.8 to 1.0 for a larger number of people. How large must the latter number be for the two programs to be deemed equally worthy of funding? The stronger the concern for the worse off (those in program A), the higher will the stated 'equivalence number' in program B be.

Person trade-off responses that take into account special concerns for severity may be represented by values for health states on the 0–1 scale from dead to full health used for QALYs. For instance, a program A prevents death in 10 people and allows them to live in full health. Program B averts an illness that leads to nonfatal state S. Assume that people consider 100 averted cases in program B to be equally worthy of funding as 10 averted deaths in program A. The value of S is then given by $1-(10/100)=0.9$. Person trade-off-based values for health states are typically higher than utilities obtained for the same states by techniques ordinarily used in the QALY field.

Other possible approaches to measuring public preferences for equity include questions about willingness to pay and questions formulated by Paul Dolan about how large a health benefit for one group of patients needs to be relative to a given health benefit for another group of patients for the two benefits to be deemed equally valuable from a societal perspective.

Modeling

Technically, there are various ways of incorporating data about concerns for equity in formal evaluation models. They may all be seen as modifications of the QALY approach that lead to evaluation in terms of EQALYs.

One modification, suggested by Nord *et al.* in 1999, is to count as one all gained life years, even if they are in less than full health, as long as they are good enough to be desired by the individuals concerned. The purpose of this is to avert discrimination against the chronically ill or disabled in valuations of interventions that extend life (confer (cfr) point E in the section 'History and value basis'). A second modification proposed by the same researchers is to place less weight than the QALY approach does on the duration of health benefits in comparisons of programs for patients with different life expectancies (cfr point D in the section 'History and value basis'). This may, for instance, be done by discounting distant health gains more strongly than at the 3–5% annual rate that is customary in conventional cost-effectiveness analysis, or by disregarding benefits that lie beyond a certain point in time. A third modification is to multiply utility gains as estimated by conventional QALY tools with explicit equity weights reflecting the severity of the preintervention condition and the degree to which health potentials are realized (cfr points B–D in the section 'History and value basis'). Alternatively, one may transform conventional utilities into societal values as illustrated in Figure 1. A transformation curve that is convex to the *Y*-axis and has strong upper end compression can, in principle, accommodate concerns for both severity and realization of potentials. For instance, in the figure the curve transforms conventional utilities of 0.4 and 0.7 to societal values of 0.8 and 0.95. If one replaces utilities from the *X*-axis with the values from the *Y*-axis, the value, for instance, of a cure of A relative to B increases from 2:1 to 4:1 (concern for severity), whereas the value, for instance, of taking someone from A to B relative to from A to healthy increases from one-half to three-fourth (concern for reduced potential).

Tentative transformation functions of the kind in Figure 1 were published by Nord in 2001 for utilities from various multiattribute utility instruments commonly used in QALY calculations. Table 1 contains the same type of information. Based on meta-analysis of policy documents and public preference measurements in several countries, the table shows a set of values for health states that purports to reflect the structure of societal concerns for severity and realization of potentials, using limitations in mobility as an example. The table is included as a potentially helpful analytical tool in guidelines for pharmacoeconomic evaluations in Norway.

Consider first the columns 1–3 in the table. The examples of states on the 8-level scale in column 1 were chosen with a view to making any one step move upwards on the scale to be roughly of the same importance from the viewpoint of affected individuals. In other words, the scale purports to be an equal interval scale in terms of individual utility. This suggests an even distribution of the 8 levels over the 0–1 utility space, i.e., utility scores for the various levels, roughly as in column 2. The numbers in column 3 are societal values. Concerns for

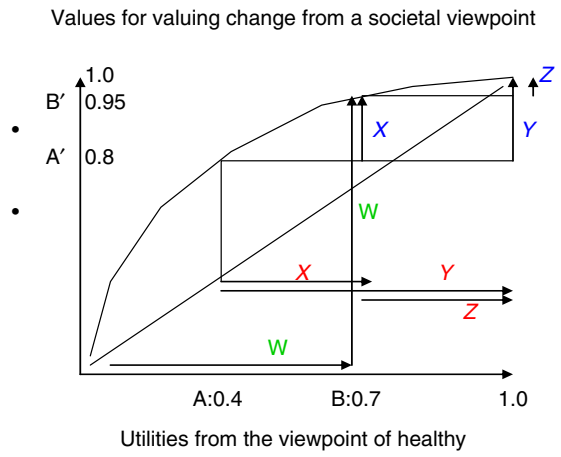


Figure 1 Utilities versus societal values for priority setting.

Table 1 Health state values encapsulating concerns for severity and realization of potential. Implied public willingness to pay (WTP) assuming WTP of €10 000 EUR for saving a life year

1. Problem level	*2. Utility (approximate)*	*3. Societal value (approximate)*	*4. Value of raise to level 1 for 1 year*		*5. Limit to willingness to pay (euro) for raise to level 1 for 1 year*	*6. Implied willingness to pay for a QALY, based on column 4(a) and 5*
			(a) Utility	*(b) SV*		
1. Healthy	1.00	1.00				
2. Slight problem	0.86	0.995[a]	0.14	0.005	500	3 500
3. Moderate	0.72	0.98[a]	0.28	0.02	2 000	7 000
4. Considerable	0.58	0.92	0.42	0.08	8 000	19 000
5. Severe	0.44	0.80	0.56	0.20	20 000	36 000
6. Very severe	0.30	0.65	0.70	0.35	35 000	50 000
7. Completely disabled	0.15	0.40	0.40	0.60	60 000	70 000
8. Dead	0.00	0.00	1.00	1.000	100 000	100 000

[a]Values adjusted after original publication.

Abbreviation: QALY, quality-adjusted life year.

Source: Adapted from Nord, E., Pinto, J. L., Richardson, J., Menzel, P. and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programs. *Health Economics* **8**, 25–39.

Examples at levels 2–7:

2. Can move about anywhere, but has difficulties with walking more than 2 km.
3. Can move about without difficulty at home, but has difficulties in stairs and outdoors.
4. Moves about without difficulty at home. Needs assistance in stairs and outdoors.
5. Can sit. Needs help to move about – both at home and outdoors.
6. To some degree bedridden. Can sit in a chair part of the day if helped by others.
7. Permanently bedridden.

severity are reflected in the fact that movements one step upwards on the scale are assigned more value the lower the start point. Concerns for not discriminating too strongly against groups with reduced potentials for health are reflected in the fact that from any given start point, improvements of different size (i.e., consisting in different numbers of steps on the scale) do not differ as much in value as they do in terms of individual utility gains calculated from column 2.

Health state values with a pattern as that in column 3 may be used to weight life years and improvements in health status in the same way as is done in QALY calculations. But valuations of outcomes are then in terms of EQALYs rather than conventional ones. They may be related to costs in cost–value ratios that in theory indicate value for money of different interventions in a broader way than cost–utility ratios do.

An alternative to calculating EQALYs by using numbers like those in column 3 is to keep QALYs themselves untouched and instead practice context-dependent willingness to pay for QALYs. Consider columns 4–6 in **Table 1**. The figures in columns 4(a) and 4(b) follow from columns 2 to 3, respectively. The figures in column 5 presuppose an anchoring value for willingness to pay. If, for instance, the willingness to pay to save a life year in normal health is €100 000 EUR, the rest of the figures in column 5 follow by rescaling the figures in column 4(b) by a factor of 100 000. The column shows that willingness to pay increases much more than proportionally to the severity of the start point. As a consequence, the willingness to pay for a QALY increases with the severity of the start point (column 6).

This can be developed further. One may, in principle, construct a hierarchical set of priority classes that takes into account various equity concerns that society deems relevant in priority setting. For each class, a maximum societal willingness to pay for a QALY is decided, such that the higher the priority class, the higher is the willingness to pay. Any outcome in terms of QALYs is assigned to its appropriate class, which will be higher in the hierarchy the more the outcome has equity concerns counting in its favor. The cost of the QALY gain will then be compared to the maximum willingness to pay for a QALY in that class. For instance, QALYs gained in people with severe conditions will, all else equal, be placed in higher classes than QALYs gained in people with moderate conditions and thus justify higher costs. An approach of this kind is considered for implementation in the Netherlands, with a social willingness to pay for a QALY ranging from roughly €10 000 EUR to 80 000 depending on preintervention severity.

Although technically different, a scheme consisting of priority classes and context-dependent willingness to pay is in its actual content equivalent to a system in which QALYs themselves are weighted and compared to a uniform willingness to pay for a QALY. In both approaches, judgments need to be made regarding how much weight the QALYs in question deserve to be given. In one approach, the chosen weight is connected to willingness to pay by assignment to priority class, in the other approach the same weight is connected to the QALY gains themselves and thus indirectly to willingness to pay. Preference data that have been elicited by means of the person trade-off or other methods in order to determine equity weights for QALYs may thus also be relevant in determining the gradient of willingness to pay in a hierarchy of priority classes. To judge whether the cost per QALY of a given intervention is within the willingness to pay for QALYs in the priority class in question may thus be seen as a variant of CVA in the general sense of the term.

Alan Williams suggested in 1997 that QALYs should be assigned more value the more the beneficiaries' expected

health over the whole life time falls short of a normal amount of health (including longevity) over a whole life. This fair innings approach is essentially a proposal to include a societal concern for equity in the formal economic evaluation. The fair innings approach to weighting QALYs for equity may thus be seen as yet another variant of CVA in the general sense of that term.

Issues

Population preference data to support CVA are presently not satisfactory. Data on what would be reasonable separate equity weights are almost nonexisting. This also applies to the fair innings approach. For the values in Table 1, column 3, the empirical basis in preference measurements is substantial, but the values are the result of an informal meta-analysis of the relevant preference literature conducted by one researcher. As noted in a review by Shah in 2009, other researchers could reach different conclusions.

Another current limitation is that Table 1 refers to health problems in terms of reduced mobility. This is because so much of the existing societal preference data pertain to this particular dimension. To apply the numbers to other kinds of health problems, one needs to know where they belong on the severity scale of Table 1. This may be judged by judging the effect on quality of life of those other problems compared to the effects on quality of life of the various mobility problems indicated in the table. Alternatively one may regard columns 2 and 3 as roughly indicating the relationship in general between individual utilities and societal values. So for instance, if one has utilities from the multi-attribute utility instrument EQ-5D columns 2 and 3 may be used to roughly estimate corresponding societal values.

One common criticism of societal value numbers is that people's responses to numerical preference questions in mailed questionnaires are unreflective and unreliable. This is to some extent true. However, researchers have also collected preference data in more high quality ways, for instance, in focus groups that discuss ethical issues carefully before each participant gives their responses to specific quantitative questions.

Finally, the idea of incorporating concerns for fairness in a numerical valuation model is controversial. Some researchers, for instance, Dolan and Olsen (2003), are concerned that such incorporation may overload the model and perhaps makes it more difficult to understand and less reliable. The alternative is to leave it to decision makers to take concerns for fairness into account informally when dealing with the results of cost-effectiveness analyses. This is an important practical issue for continued debate. It is also a theme for further research. At the end of the day, it is an empirical question whether decision makers feel helped or not by CVA, or feel more helped when provided with such analyses in addition to conventional cost-effectiveness analyses.

See also: Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Quality-Adjusted Life-Years. Valuing Health States, Techniques for. Willingness to Pay for Health

References

Dolan, P. and Olsen, J. A. (2003). *Distributing health care: Economic and ethical issues*. Oxford: Oxford University Press.

Nord, E. (2001). Utilities from multi attribute utility instruments need correction. *Annals of Medicine* **33**, 371–374.

Nord, E., Pinto, J. L., Richardson, J., Menzel, P. and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programs. *Health Economics* **8**, 25–39.

Further Reading

CVZ (2006). *Pakketbeheer in de praktijk*. Diemen: CVZ Rapport.

Dolan, P. (1998). The measurement of individual utility and social welfare. *Journal of Health Economics* **17**, 39–52.

Nord, E. (1993). The trade-off between severity of illness and treatment effect in cost-value analysis of health care. *Health Policy* **24**, 227–238.

Nord, E. (1995). The person trade-off approach to valuing health care programs. *Medical Decision Making* **15**, 201–208.

Nord, E. (1999). *Cost-value analysis in health care: Making sense out of QALYs*. Cambridge: Cambridge University Press.

Patrick, D., Bush, J. and Chen, M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research* **8**, 228–245.

Shah, K. K. (2009). Severity of illness and priority setting in healthcare: A review of the literature. *Health Policy* **93**, 77–84.

Williams, A. (1988). Ethics and efficiency in the provision of health care. In Bell, J. M. and Mendus, S. (eds.) *Philosophy and medical welfare*, pp. 111–126. Cambridge: Cambridge University Press.

Williams, A. (1997). Intergenerational equity: An exploration of the 'fair innings' argument. *Health Economics* **6**, 117–132.

Cross-National Evidence on Use of Radiology

NR Mehta, Riddle Hospital, Media, PA, USA, and University of Pennsylvania, Philadelphia, PA, USA
S Jha and AS Wilmot, University of Pennsylvania, Philadelphia, PA, USA

Introduction

The specialty of radiology, diagnostic imaging, has revolutionized the practice of medicine across the globe. No other form of diagnostic medicine has had such a dramatic impact on disease detection and mapping progression of treatment in the preceding decades. In a 2001 survey of physicians, magnetic resonance imaging (MRI) and computed tomography (CT) scanning ranked number 1 amongst 30 medical innovations of the last 25 years, beating cholesterol-lowering HmG-CoA reductase inhibitors (statins), coronary arterial bypass graft, and newer generation antibiotics (Fuchs and Sox, 2001).

With the diagnostic imaging technological revolution has come the inherent increased costs of the technology itself. With CT scanners and MRI scanners costing upward of $3 million (US), the utilization of these machines at an ever increasing pace has helped drive up the medical bills of patients everywhere.

One of the benefits to having diagnostic imaging technology disseminated throughout the world is to provide a window into how the differing health care delivery systems tackle this issue of managing cost and utilization in the face of limited resources. In this article, four countries are studied: the US, the UK National Health Service (excluding Scotland), Canada, and Japan. The US provides a window into their blend of private and government-sponsored health care systems. The UK and Canada allow a glimpse into two variants of government-run health care. Japan allows for an analysis of their social insurance health care system, which has the highest per capita number of CT and MRI scanners of the comparison countries. As can be seen from **Tables 1** and **2**, these countries differ substantially in their numbers of advanced diagnostic equipment (CT and MRI scanners) as well as radiologists per capita. This article will document these differences and provide some suggestions of possible contributing factors. More rigorous analysis of determinants of cross-national differences in technology uptake and their effects on health outcomes remains an important subject for future research.

United States

Health care in the US is a mix of private and government-sponsored methods of financing and care delivery. Insurance coverage largely depends upon age, income, and employment. For the majority of the adult population under the age of 65, private insurance is obtained through the workplace. Employer-sponsorship of health insurance takes advantage of tax preferences, facilitates contract negotiation for employees, and creates an insurable pool of enrollees. Those who are not employed or who do not have employer-sponsored health care (sole business owners, independent contractors), can buy insurance directly from insurance companies in what is known as the individual market. Much of the *Patient Protection and Affordable Care Act of 2010* is devoted to reforming this individual market, such as removing preexisting condition exclusions, setting medical-loss ratios for insurance companies, and creating health insurance exchanges to provide information and subsidies to individuals who purchase these policies.

For senior citizens over the age of 65, there is government-sponsored Medicare. This program, which is administered by private carriers, sets provider payments for hospitals and physicians nationally, including reimbursement for radiology. The program is funded by a combination of payroll taxes on workers, general revenues and premiums paid by beneficiaries. Finally, a subset of people below the poverty line are eligible for Medicaid. Medicaid is government-sponsored by both Federal and state governments. Provider payments are set on a state-by-state basis and the program is funded via taxes. People who fall outside of these public and private programs remain uninsured, except for minor additional programs. Private insurance programs (employer-sponsored and individual) for those under the age of 65 tend to follow the national fee structure provided by Medicare.

Among the four countries considered, the US has the most radiologists per capita. In addition, the US has the second highest number of MRI and CT scanners compared with the other countries. The high number of scanners can in large part be attributed to the fee-for-service system, a system that rewards doing more per patient. The majority of the country has no limits regarding the number of scans performed or the number of scanners in operation, with only a few state-based exceptions where a certificate of need is required prior to the purchase of a scanner. For every scan performed, a fee is collected, and thus the incentive to perform higher volume of scans. The higher volume of scans translates to a higher volume of scanners.

Payment for imaging services in the US is, in general (driven by Medicare), split into two categories: technical fee and professional fee. The technical fee is that which goes to the owner of the imaging equipment. The professional fee goes to the radiologist for interpretation of the study. Typically, the professional component is much less than the technical component, reflecting the relatively high equipment costs. In 2011, for example, a CT scan of the head carried a professional fee around $40, as compared to the technical fee of around $150.

A major legislation undertaken by the Federal government to curb cost and growth in imaging was enacted in the Deficit Reduction Act of 2005 (DRA 2005). This legislation reduced the technical fee payment for contiguous body part scanning. Hence, a CT scan of three contiguous body parts, such as the chest, abdomen, and pelvis, where the reimbursed technical fee was 100% for each, became 100% for the chest and 50%

Table 1 Data on MRI and CT in US, England, Canada, and Japan from OECD

OECD data	*Total health care expenditure (THE) as % of GDP*	*Radiology as % of THE*[d]	*Per capita spending on radiology*[c]	*MRI units per million of population*	*MRI exams per 1000 of population*	*Yearly utilization per MRI scanner (calculated)*	*CT units per million of population*	*CT exams per 1000 of population*	*Yearly utilization per CT scanner (calculated)*
US	17.6 (2010)	4.9	$403.42	31.6 (2010)	97.7 (2010)	3091.8	40.7 (2011)	265.9 (2010)	6533.2
UK	9.6 (2010)	1.4	$48.06	5.9 (2011)	38.6 (2009)	6542.4	8.9 (2011)	72.8 (2009)	8179.8
Canada	11.4 (2010)	1.2	$55.29	8.6 (2011)	47.7 (2010)	5546.5	15 (2011)	126.9 (2010)	8460.0
Japan	9.5 (2009)	5.2	$157.82	43.1 (2008)	65.4 (2002)[b]	1518.3[a]	97.3 (2008)	155.3 (2002)[b]	1596.0[a]

[a]Japan yearly utilization calculated based on 2002 data, where there were 92.62 CT units per million of population and 35.32 MRI units per million of population.
[b]Kandatsu, 2002
[c]Calculated based on OECD per capita health care expenditure in combination with percentages from 1st and 2nd columns. Per capita health expenditure for the US and the UK is 2010. Per capita health expenditure for Canada is 2011 (estimated). Per capital health expenditure for Japan is 2009.
[d]Percentages from text.
Note: Data for UK are based on hospital numbers, as ambulatory numbers were unavailable.

Table 2 Radiologists per million of population with data obtained calculated as described

Country	*Radiologists per million of population*	*CT scans per radiologist per year*[d]	*MRI scans per radiologist per year*[d]
US	100 (2009)[a]	2279.0	912.0
UK	45 (2012)[b]	1693.0	897.7
Canada	67 (2011)[c]	1871.6	641.8
Japan	36 (2004, OECD)	1725.5[e]	727.1[e]

[a]American College of Radiology, practice of radiology in the US, 2009.
[b]Center for Workplace Intelligence, 2012.
[c]Based upon 2294 Canadian radiologists (Canadian Medical Association- Number and percent distribution of physicians by specialty and sex, Canada 2011).
[d]Based on assumption of stable number of radiologists over short period of time and based on calculations from **Table 1**.
[e]Corrected for 40% of scans interpreted by radiologists.

for the abdomen and pelvis (Moser, 2006). In 2012, Medicare further reduced payments to radiologists by decreasing the professional fee on a second body part for patients scanned on the same day by 25%. Although these changes primarily impact Medicare patients, insurance carriers tend to follow Medicare rates, giving this legislation tremendous impact. Indeed, Medicare rates indirectly serve as a 'national fee schedule.' As a result of the DRA 2005, imaging volumes in radiology offices decreased 2.0% between 2006 and 2007, as compared to yearly increases of 8.4% between 2002 and 2006 (Levin *et al.*, 2009).

The Organization for Economic Development and Cooperation (OECD) data indicate that the US, as compared to UK, Canada, and Japan, has the highest total health expenditure on imaging as a percent of gross domestic product (GDP) (**Table 1**). The US ranks second among this group in terms of number of CT and MR scanners per million, with Japan taking the top spot. The US also has by far the highest number of scans – both MRI and CT – per capita population, implying that there is a relatively high level of access to imaging technology in the US. However, when utilization per scanner is estimated (number of scans per scanner), the country falls to the second to last in terms of MR and CT utilization, indicating a relative under utilization of imaging equipment compared to other countries. Indeed, the US and Japan are the only high-income countries in the world, which allow for essentially unrestricted acquisition of high-technology scanners in a fee-for-service environment (Cutler and Ly, 2011). It is, therefore, not surprising that both of these countries have more scanners and lower utilization of scanners compared to the UK and Canada.

The number of radiologists practicing in the US is around 100 per million population (**Table 2**), the highest of the analyzed countries. In contrast to the relatively low scanner utilization in the US, the radiologist utilization is the highest, indicating that the US radiologist is reading more studies per year than their peers in other countries. Thus the overall evidence shows that the US has a relatively high number of scanners, radiologists, and scans per capita, which is consistent with it having relatively few controls on investment in new equipment and on licensure of new radiologists.

According to the Center for Medicare and Medicaid Services and Blue Cross Blue Shield Association, expenditure on diagnostic imaging has been approximately 5% of total expenditure on health care. The total amount of money spent on diagnostic imaging in the US in the year 2000 was approximately $75 billion. In 2000, the national health expenditure was $1.377 trillion, making imaging costs 5.4% of total health care expenditure. The total cost of diagnostic imaging for 2005 was estimated to be $100 billion. In 2005, the national health expenditure was $2.029 trillion, making imaging costs approximately 4.9% of total expenditure on health care.

Between 1998 and 2005, the annual growth rate in diagnostic imaging in the Medicare population was 4.1%. This has slowed down in recent years, likely as a result of a combination of cost-containment strategies from the government as well as the economic slowdown. Between 2005 and 2008, the annual growth rate of imaging in the Medicare population was 1.4% (Levin *et al.*, 2011).

United Kingdom/England

England has a universal public health care system (National Health Service, NHS) with a supplementary private insurance

system. Taxes are used to fund the NHS, where most care is provided at no cost to the patient at the point of service. Patients register with and go to a general practitioner (GP) who then serves as a gatekeeper between them and the hospitals/specialists, including radiologists who normally are employed in radiology departments within hospitals. Supplementary private insurance is purchased by about 12% of the population. It mostly pays for quicker access to specialists and elective surgeries, which may be performed in private hospitals or private beds in NHS hospitals. Anecdotally, private insurance provides a greater degree of access to imaging than the NHS. The private system is staffed largely by the same physicians who serve in the NHS.

In general, over the preceding decades, radiology in the NHS has been characterized by limited quantity of radiology equipment, limited number of radiologists, and waiting lists for patients. These issues have been tackled and have steadily improved.

While the density of radiology equipment per capita in England remains far lower than in the US, there has been a substantial upgrading of imaging equipment in England over the past decade. According to the UK Department of Health, during 2000–07 the NHS spent £564 million (£80 million per year) on CT, MRI, and LINAC (linear accelerator, for radiotherapy) machines in inflation-adjusted currency. The estimated cost to replace this equipment over the next decade is £1 billion, noting yearly NHS annual budgets of around £100 billion.

In 2001, there were 1586 consultant radiologists. In 2010, the number of full-time equivalent radiologists was 2194, representing an approximate 38% increase over the decade. This translates to approximately 45 full-time equivalent radiologists per million of population. Despite this increase, it is still below the Royal College of Radiologists recommendation of eight full-time equivalent radiologists per 100 000 of population, according to a December 2012 Center for Workplace Intelligence report. Universal evening and weekend coverage is not prevalent as is the case in the US. There is a drive toward longer hours, and 12–14 h days per radiologist, working 7 days per week has been implemented at Royal Sussex County Hospital in Brighton with reported success. In order to provide around the clock coverage, 24 h a day and 7 days per week, the number of radiologists would need to increase to 6000, which implies roughly doubling the current number.

In addition to high case volume, radiologists in England face additional work pressures. The NHS Cancer Plan requires that a radiologist be present at multidisciplinary meetings, which have increased in duration and frequency since 2007. These, on average, occupy 10% of the radiologists' clinical time. In contrast, this is not a requirement in countries such as the US, where it is occasionally provided as a voluntary effort. This results in additional radiologist time taken away from reading films, exacerbating shortages. As in the US, an aging population and increasingly complex imaging examinations with an increased number of images per study, have also increased the clinical burden on radiologists. A Center for Workplace Intelligence report from August 2011 reports on burn out resulting in radiologists leaving the work force for sick leave or early retirement, as well as an increased rate of mistakes such as overlooked lung cancers on radiographs. A study from the Royal College of Surgeons of Ireland in March 2011 surveying Irish radiologists describes understaffing issues in a system in which radiologist numbers are centrally controlled by government agencies. The authors argue that current methods of determining radiologist productivity are outdated and do not give adequate weighting to responsibilities such as teaching, procedures, double reading, and interpreting outside films.

Private practice radiology does exist on a more limited scale than the US, providing 10–15% of radiology services, as per a July 2002 Audit Commission report. Fees for diagnostic exams vary from provider to provider, but in general align with fees charged in the US.

England has made progress in terms of patient wait times for imaging. An audit commission report in 2002 found the average wait time for outpatient MRI services was 20 weeks, while for CT this was over 6 weeks. In 2004 the NHS contracted with an independent sector radiology provider, Alliance Medical, to provide 635 000 MRI scans to assist with MRI backlogs. This served as a short-term solution to the waiting lists. However, there is concern from within the NHS radiology departments as to direct competition with the independent sector for limited NHS funds. According to the Department of Health, as of 2009, wait times over 6 weeks for CT and MRI have been essentially eliminated.

The OECD data show that as of 2012, health care spending in England is lower than in the US, accounting for 9.8% of GDP compared to 17.4% in the US. However, rising health care costs have led to recent reforms in the NHS. As per the Department of Health spending review, the budget of the NHS for 2011 is £103.8 billion, and the current budget provides a 0.4% increase in real terms through 2015. Overall planned cost cutting include £20 billion in efficiency savings and a 33% decrease in administrative costs. Specific to radiology, there will be an expected £8 million in savings annually to be achieved by having some plain radiographs interpreted by radiographers (nonphysicians) rather than radiologists.

In 2008/2009, £1.1 billion was spent on radiology services, equating to 1.4% of the NHS budget. This is a smaller percentage when compared to the US (Grant *et al.*, 2012). About 38.8 million imaging examinations were performed in England in 2010, including 4 million CTs, 2.1 million MRIs, and 22.2 million radiographs, as per the Center for Workplace Intelligence. This volume amounts to approximately 73 imaging examinations per 100 population per year. There has been a rapid increase in volume of imaging in the UK, and between 1996 and 2010 there has been a 445% increase in MRI, 279% increase in CT, 94% increase in ultrasound, and a 16% increase in radiographs. As in the US, the increase has primarily involved the more advanced and expensive imaging modalities. Based upon calculations from the OECD health data, the US, in comparison, has had an increase of 208% in MRI, and 262% in CT.

Tables 1 and 2 show fewer CT and MRI scanners in the UK relative to the US. When accounting for the total number of scans performed, on average the UK seems to have a higher utilization of their imaging equipment. On a per radiologist basis, despite the aforementioned concerns of high case load and clinical burden, radiologists read on average less number

of CT and MRI cases per year than their US counterparts. The difference between the countries might in part be attributable to the differential payment structure of radiologists. In the US, there is a financial gain for reading more studies, while in the UK there is no such overt financial benefit in their salaried model.

Canada

Canada has a single-payer universal health care system paid for through taxation. Cost containing strategies, such as patient copayments, are effectively prohibited for 'medically necessary services' by federal mandates. The roots of the Canadian health care system date back to the Federal Health Insurance and Diagnostic Services Act of 1957. The act provided that provinces funded 50% of the health care cost with a federal match of 50%. Federal funding was contingent upon the provinces providing medically necessary care, portability of coverage, and universal coverage. In 1977, the open-ended federal funding of health care was replaced with a federal per capita block grant, meaning a fixed amount of money would be provided to provinces every year, initially with indexing to the GDP. In the early 1990s, the federal contribution was frozen at 1989 levels, making the provinces responsible for all growth of spending. By 1999, the federal share of health care costs had fallen to between 10% and 20%. Until 2005, the Canadian system banned private insurance from providing services covered by public health insurance. In 2005, the Quebec Supreme Court ruled in Chaoulli versus Quebec that Quebec's prohibition of private medical insurance in the face of long wait times for public federally mandated care violated 'rights to life' and 'security of person'in Quebec's charter. The provincial government has so far responded to this ruling by managing waiting times, rather than encouraging growth of private insurance.

The provincial contribution to health care is generally from income or payroll taxes that are not earmarked specifically for health care, and hence the amount of money individuals pay for health care is not obvious to the taxpayer. From 2001 to 2010, the rate of health care spending has increased at greater than three times the rate of inflation. Health care spending is projected to equal or exceed 50% of all revenue in 6 of 10 Canadian provinces by 2017 (Skinner and Rovere, 2011).

Every year the provincial government negotiates annual global budgets with the hospitals. The fixed budget covers all operating costs and is based on estimated volume of patients (occupied beds). New capital expenditures are allocated separately. There is a theoretical disincentive on the part of the hospitals to provide expensive services, unless this would result in increased revenue, which would typically only happen with a lag.

Physicians are primarily in solo practices (about 50% are GPs), and collect their revenues via a fee-for-service system but subject to an annual aggregate spending limit. Provinces and medical associations determine a uniform fee schedule that typically applies throughout the province. Expenditure and income caps per physician are put into place (varying from province to province), which are intended to prevent overutilization. After achieving a certain level of income (total fees), the physician is paid only a percentage of the remaining fees. Radiologists, in particular, are facing such 'clawbacks' proposed by provincial governments. Once total billings reach a certain level, the clawback reduces payment of subsequent services by a fixed percentage. A 2012 Ontario proposal reduces payment by 5% for billings above $400 000, 10% over $750 000, 25% for billings over $1 million, and 40% for billings over 2 million. This reduced marginal benefit attempts to balance the incentive of reading too many scans. The concern of the clawback scheme is the potential exacerbation of current waiting lists.

Canadian radiologists are paid primarily in a fee-for-service system. Based on data from the 2010 National Physician survey, 80% of diagnostic radiologists who responded received greater than 90% of their income from fee-for-service, while 10% received income from a blended source (which can include fee-for-service, salary, capitation, contract, on-call remuneration, etc.).

A small number of private radiology clinics do exist in Canada. As of 2007, there were 42 for-profit MRI/CT clinics in Canada. Traditionally these clinics have performed scans as a fee-for-service out of pocket payment and radiologists at these sites do not work in the public sector. Rates for scans in Alberta range between $500 and $800 per scan, while rates per scan in British Columbia range between $500 and $2200. To help combat public sector wait lists, they are now being used to help increase imaging capacity in the provinces via contracting with the public health service (Mehra, 2008).

As per 2010 data from the OECD, Canada spends 11.4% of its GDP on health care costs (**Table 1**). Estimated per capita spending for radiology in Canada is $55.29. This is closer to the spending of the UK, and considerably less than that of the US. According to the Canadian Association of Radiologists in 2012, costs of medical imaging in Canada (including maintenance of equipment and physician payment) is approximately $2.14 billion (US). Total health care expenditure is 11.4% of a GDP of $1.6 trillion (US), or $180.8 billion. Based on this, diagnostic imaging costs in Canada are approximately 1.2% of total health care expenditure. However, it should be noted that this does not include capital costs of scanner purchase. Taken at face value, the percentage is on par with the UK share of 1.4%, however much lower than the US share of 4.9%.

While a majority of Canadian citizens and physicians have an overall positive impression of the Canadian health care system, the system is not without criticisms. One criticism of the Canadian health care system has been with regard to long wait times. There is a low density of physicians in Canada that serves as one potential rate-limiting step with regard to overall health care spending. This holds true in radiology, with 67 radiologists per one million population, compared to 100 radiologists per million in the US. In addition to lower manpower availability, the density of expensive medical equipment such as MRI and CT scanners is also lower in Canada, potentially limiting access and resulting in long wait times. According to one survey released in 2009, the wait list for urgent MRIs ranged from 24 h to greater than 1 month, and the wait list for elective MRIs ranged from 28 days to 3 years. Other criticisms that have been raised in the past decade relate to slower adoption of new technology, which may in

some cases and in some parts of the country lead to patients undergoing diagnostic and interventional procedures performed with less modern equipment than would be possible with more resources. Furthermore, due to the single-payer system, diagnostics and procedures that are reimbursed at a low rate or not at all by the public health system may be difficult for patients to obtain. As noted in **Table 2**, the number of CT and MRI scans interpreted per radiologist is less than their US counterparts, whereas Canada ties with the UK in having the highest number of scans per scanner. This suggests that availability of scanners or budget allocation to pay for scans are on average more often the limiting factors on patient access, rather than manpower.

Comparison between the UK and Canada, both with single-payer health care systems, shows similarities in the percent of radiology expenditure as a share of total health expenditure, as well as the per capita expense of radiology services. There are also strong similarities in scanner utilization. These similarities are in place despite the difference in payment models to radiologists, with Canada being fee-for-service and the UK being a salary model. It might be surmised that radiologists are not, therefore, in the driver seat of imaging utilization, and that it is the organization of the health care system that plays a more critical role.

Japan

Japan has a universal health insurance system. Health insurance is mandatory, with individuals receiving insurance either via employer-sponsored plans or via one of several government-sponsored health insurance plans. Health care spending as a percent of GDP is low in Japan relative to other industrialized nations – largely a result of the government's tight regulation of health care prices. The system operates via a national fee schedule, reviewed biennially, which determines government reimbursement for all health care services. This single payment system has served as a remarkable control mechanism for costs.

Despite tight government control over reimbursement for imaging, there is no central government control on the installation of high-technology scanners in Japan. Japan has the most MRI and CT scanners per capita. The high density of MRI and CT scanners in Japan is an interesting phenomenon, because reimbursement for imaging in Japan is far lower compared to the US. For instance, the reimbursement for a CT scan in Japan in 2008 was equivalent to $80, with the reimbursement for an MRI equivalent to $155–180, these prices being approximately one-fifth to one-tenth of the reimbursement for the same studies in the US (Ehara *et al.*, 2008).

The low level of reimbursement begs the question as to why Japanese hospitals and clinics would purchase so many scanners and perform such a high volume of imaging, and whether imaging in Japan is profitable. The answer in part is cultural and relates to the expectation for rapid diffusion of medical technology in the Japanese society, which is quick to believe in its benefits even in the absence of clinical effectiveness data. Medical imaging in Japan is in fact not typically profitable, yet hospitals reportedly seek high-technology scanners so as to maintain their prestige and competitive edge. The prestige of having an MRI scanner may attract more patients and increase profits indirectly from margins on other services. Furthermore, while the government reimburses imaging at a low rate, it does provide subsidies for purchasing imaging equipment by major public hospitals and academic medical centers (Ikegami and Campbell, 2005). Outside of major academic centers, private sector imaging providers who do not receive any government support tend to operate with lower cost Japanese-made scanners. The cost of imaging equipment in Japan is significantly lower than in the US, which also helps to explain the high density of scanners. Toshiba, Hitachi, and Shimadzu produce less expensive models of imaging equipment for sale to Japanese providers (Kandatsu, 2002).

A 2001 survey of scanners in Japan by the Japan Radiological Society revealed that approximately 30% of MR scanners were high-field 1.5 T scanners. A survey from 2005 showed that 53% of installed MRI scanners were 1.0 T or less. In comparison, a 2006 IMV market research survey of the US reported 90% of MRI scanners at 1.5 T field strength or greater. The strength of the magnetic field in MRI is measured in Tesla units, and higher Tesla scanners are stronger scanners. This increased magnetic field strength in MRI results in higher signal-to-noise ratio (SNR) in the resulting image. With high SNR, smaller structures and finer details are more easily visualized, which theoretically improves diagnostic accuracy. When comparing costs between superconducting scanners, in 2004, a 0.3 T scanner can cost around 70 million Yen (approximately $753 000), while a 1.5 T scanner can run 120 million Yen (approximately $1.3 million) (Hayashi *et al.*, 2004).

Japanese fee schedules have been adjusted over time to reflect the increased use of MR and CT. As the volume of imaging increases, the government decreases reimbursement to control overall expenditures. For instance, in 2002 the reimbursement for an MRI brain exam was decreased from 16 600 Yen ($180 in US dollars, using early 2013 exchange rate) to 11 400 Yen ($124), an approximate 30% decrease. Over the past decade, there has been recognition that the higher cost of operation and the higher quality of imaging provided by higher field strength MRI scanners and multidetector CT deserves higher levels of reimbursement.

Other issues that distinguish the practice of radiology in Japan from that in the US, Canada, and England include the prevalence of interpretation of images by nonradiologists. In 1996, the government began offering higher reimbursement for studies interpreted by board-certified radiologists. While the proportion of studies interpreted by radiologists has increased since that time, only 40% of imaging examinations were interpreted by radiologists as of 2003 (Nakajima *et al.*, 2008).

Of all of the countries included in this article, Japan has the lowest density of radiologists, with 36 per one million of population as of 2004. Japanese radiologists worked an average of 63.3 h per week in 2006. Cases read per radiologist, or radiologist utilization, are on par with the US when accounting for the 40% radiology interpretation rate. A 2002 survey from the European Society of Radiologists of 14 European countries showed that essentially all CT and MR examinations are reported by radiologists.

Japan's total health expenditure is on the lower end of the spectrum when compared to the other countries analyzed in this article. In 2003, radiology costs were estimated to be approximately 5.2% of total health care expenditures (Imai, 2006). This is closer to the radiology share of spending in the US (4.9%), than in the UK and Canada.

Conclusion

Comparison of the four countries used in this study demonstrates important cross-national differences in the utilization of diagnostic imaging, both absolutely and as a percent of total health care spending.

On the side of total spending, the US and Japan have the highest percentage of total health expenditure utilized for radiology, at 4.9% and 5.2%, respectively. By contrast, the UK and Canada have the lowest percentage of total health expenditure utilized for radiology, at 1.4% and 1.2%. However, note that data for Canada do not include costs of scanner purchase, only operational costs. One of the major differences between these groups of countries is that the former (US and Japan) are not single public payer systems. And although the latter group (UK and Canada) do have a degree of private practice running alongside the single public payer, the public system is by far the dominant mode of health care delivery. Publicly owned providers are fundamentally not designed to make a profit on the delivery of care.

From the provider reimbursement standpoint, fee-for-service versus salaried model of radiologist pay does not, with this limited glance, account for significant differences. Canada is a fee-for-service system, while the UK is a salaried model, and both systems achieve a relatively low percentage of total health expenditure utilized for radiology.

In terms of access, Japan has the most scanners per capita but ranks second, after the US, in number of scans per capita. Utilization of equipment numbers, however, indicates that the UK and Canada use their equipment more intensively than the US and Japan, which is perhaps unsurprising given the former two countries' lower number of scanners per capita. Access to imaging is related to the percentage of total health expenditure utilized for radiology.

Ultimately, it might be surmised that an 'if you build it, they will come' mentality exists within health care, and that single-payer models serve as a better mechanism to limit both imaging access and costs. Both the UK and Canada have government budget constraints that can tightly control number of scanners in the market. And while Canada pays the radiologist a fee-for-service model for interpretation of the scan, the performance of the scan is not reimbursed in this manner. Therefore, a potential strategy for countries attempting to reign in radiology expenditures is the elimination of technical fee-for-service, while preserving current mechanisms of radiologist interpretation reimbursement. Simply reducing the technical component fee may not be enough, as Japan has shown with its reduced fee schedule. The market response in Japan has been to utilize lower cost scanners, and the country has continued high radiology costs as a percentage of total health expenditure.

Further research is needed into whether technical fee-for-service reimbursement is a causative factor for higher costs, not just for medical imaging, but also for health care as a whole. The removal of technical fee-for-services, not merely the reduction of fees, in laboratory services, surgical and clinical services, in addition to imaging services could serve as a future direction of health care cost containment and health care policy.

See also: Diagnostic Imaging, Economic Issues in. Health Insurance Systems in Developed Countries, Comparisons of

References

Cutler, D. M. and Ly, D. P. (2011). The (paper)work of medicine: Understanding international medical costs. *Journal of Economic Perspectives* **25**(2), 3–25. Spring.

Ehara, S., Nakajima, Y. and Matsui, O. (2008). Radiology in Japan in 2008. *American Journal of Radiology* **191**, 328–329.

Fuchs, V. and Sox, Jr., H. C. (2002). Physicians' view of the relative importance of thirty medical innovations. *Health Affairs* **20**(5), 30–42.

Grant, L., Appleby, J., Griffin, N., Adam, A. and Gishen, P. (2012). Facing the future: The effects of the impending financial drought on NHS finances and how UK radiology services can contribute to expected efficiency savings. *The British Journal of Radiology* **85**(1014), 784–791.

Hayashi, N., Watanabe, Y., Masumoto, T., et al. (2004). Utilization of low-field MR scanners. *Magnetic Resonance in Medical Sciences* **3**(1), 27–38.

Ikegami, N. and Campbell, J. (2005). Medical care in Japan. *New England Journal of Medicine* **333**(19), 1295–1299.

Imai, K. (2006). Medical imaging: It's medical economics and recent situation in Japan. *Igaku Butsuri* **26**(3), 85–96. Article in Japanese.

Kandatsu, S. (2002). *Modalities in Japan*. Special Report. Japanese Radiological Society. Available at: www.radiology.jp (accessed 26.07.13).

Levin, D. C., Rao, V. M., Parker, L. and Frangos, A. J. (2009). The disproportionate effects of the Deficit Reduction Act of 2005 on radiologists' private office MRI and CT practices compared with those of other physicians. *Journal of the American College of Radiology* **6**, 620–625.

Levin, D. C., Rao, V. M., Parker, L., Frangos, A. J. and Sunshine, J. H. (2011). Bending the curve: The recent marked slowdown in growth of noninvasive diagnostic imaging. *American Journal of Roentgenology* **196**, W25–W29.

Mehra, N. (2008). Eroding public Medicare: Lessons and consequences of for-profit health care across Canada. Ontario Health Coalition. Available at: www.web.net/ohc/ (accessed 26.07.13).

Moser, J. W. (2006). The Deficit Reduction Act of 2005: Policy, politics, and impact on radiologists. *Journal of the American College of Radiology* **3**, 744–750.

Nakajima, Y., Yamada, K. and Imamura, K. (2008). Radiologist supply and workload: International comparison. *Radiation Medicine* **26**, 455–465.

Skinner, B. J. and Rovere, M. (2011). *Canada's Medicare bubble: Is government health spending sustainable without user-based fundings?* Fraser Institute. Available at: www.fraserinstitute.org/uploadedFiles/fraser-ca/Content/research-news/research/publications/canadas-medicare-bubble.pdf (accessed 26.07.13).

Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties

L Bojke and M Soares, University of York, York, UK

Glossary

Covariate A variable that is possibly related to the outcome under study.

Credible interval An estimate of the range of values possible within a specified degree of credibility, usually 95%.

Elicitation Method to obtain subjective beliefs from an individual.

Heuristics Experience-based techniques for problem solving, learning, and discovery such as rules of thumb.

Introduction

Decision-modeling is increasingly used or required by health technology funding/reimbursement agencies as a vehicle for economic evaluation. The process of developing and analyzing a decision analytic model as part of a health technology assessment (HTA) involves many uncertainties. Some relate to the assumptions and judgments regarding the conceptualization and structure of a model, others to the quality and relevance of data used in the model. Where data are absent or inadequate to inform model uncertainties, the decisionmaker is faced with the options of using whatever data are available, or commissioning and/or waiting for further research. Delaying a decision is not without negative consequences, however, as patients may not receive what is actually the most cost effective intervention and population health will be negatively affected. As an alternative to delaying decisions, eliciting expert opinion can be useful to generate or complement the missing evidence.

Elicitation can transform the subjective and implicit knowledge of experts into quantified and explicit data. Characterizing experts' uncertainty over the elicited values of parameters further used within a decision model, and assessing the consequential impact on decision uncertainty, is particularly important in HTA. It is also useful in exposing disagreements and different degrees of uncertainty among experts. By specifying the 'current level of expert knowledge' as distributions, these can be used to generate estimates of the value of conducting further research to resolve these uncertainties.

There are many possible uses for elicitation in HTA (Box 1). In general, it is relevant where otherwise less informed, implicit or explicit assumptions have to be made. Expert knowledge can, therefore, help to characterize uncertainties that otherwise might not be explored.

Techniques for eliciting uncertain quantities have received a lot of attention in Bayesian statistics. However, it is a relatively new technique in HTA and there are few examples of its use. This article attempts to distill a large literature so as to outline the methods available and their applicability to HTA, using relevant examples from the field. It is not intended to be a comprehensive summary but is instead a general guide with further reading for those wishing to dig deeper.

The stages of an elicitation are divided into: the design of the exercise, its conduct, methods for synthesizing data from multiple experts, and assessments of adequacy of the exercise.

Box 1 Uses of elicitation in HTA decision-modeling

The possible uses of elicitation in HTA decision-modeling include:

- Generating an appropriate set of comparators.
- Identifying appropriate patient pathways and relevant events.
- Describing parameters and their associated uncertainty.
- Quantifying the extent of bias, or improving generalizability from one context to another.
- Characterizing structural uncertainties either through generating differential weights for scenarios or by eliciting distributions of parameterized uncertainties.
- Validating or calibrating model estimates.

The Design Process

Decisions on what quantities to elicit and how to do it should be determined by the intended purpose. There are a number of issues to consider, and these can be categorized as: whose beliefs to collect, what and how to elicit, and specificities of elicit complex parameters such as beliefs regarding correlation.

Whose Beliefs?

There is a large literature on the selection of experts. The criteria range from citations in peer reviewed articles to membership of professional societies. There is no consensus on the best approach. It is generally agreed that an expert should be a substantive expert in the particular area. However, the issue of whether an expert should possess any particular elicitation skills (e.g., previous experience of elicitation) is less clear and will depend on the complexity of the task. Experts with statistical knowledge may be required for elicitation of quantities such as population moments or parameters of statistical distributions, though most experts can be assumed to provide reasonable estimates of observable quantities, such as proportions. In selecting experts, ideally only those without competing interests should be chosen so as to reduce motivational bias. Once the analyst has selected the expert group, one needs to decide how many experts to include in an exercise. Generally, multiple experts will provide more information than a single expert; however, there is a lack of guidance regarding the appropriate number of experts.

 doi:10.1016/B978-0-12-375678-7.01406-1

What to Elicit?

Although previous elicitations have often sought to elicit probabilities or numbers of events, costs, quality of life weights, and views on relative effectiveness can also be elicited. Once the analyst has decided on the parameters to elicit, the methods of doing so come to the fore. There are several methods available. When eliciting, for example, a transition probability, experts can be asked to indicate their beliefs regarding the probability itself, the time required for x% patients to experience the event, or the proportion of patients who would have had experienced the event after y amount of time. In other words, conditional on particular assumptions, evidence on each of these aspects can inform the same parameter. In selecting an appropriate method, there is a need to consider the compatibility of the format with that of other evidence in the model to be used jointly with the elicited judgments. Where multiple parameters are to be elicited, the analyst may promote some homogeneity in the quantities used, avoiding, for example, seeking judgments on transition probabilities by using proportions of patients for some parameters and the time required for x% patients having had experienced the event for others. It is also generally accepted that experts should neither be asked regarding unobservable quantities nor regarding moments of a distribution (except possibly, the first moment, the mean) or coefficients for covariates.

How to Elicit?

After choosing which quantities to elicit, the expert needs to be able to express his/her uncertainty over each. Previous applications of elicitation techniques have found that nonnumerical expressions of uncertain quantities can be useful. However, obtaining quantitative rather than qualitative judgments on the level of uncertainty is required in a decision model. This is usually done by asking experts to specify their beliefs over a manageable number of summaries characterizing their uncertainty surrounding the quantity of interest. Ideally, the focus should be on eliciting summaries with which the experts are familiar and it is generally agreed that experts do not perform well when asked directly to provide estimates of variance. It can also be useful to elicit quantities that are conditional on observed or hypothetical data.

Experts can be asked to reveal credible intervals directly (the range of values that an expert believes to be possible within a specified degree of credibility, usually 95%) or other percentiles of the distribution. Variable interval methods can be used, where percentiles are prespecified and the expert is asked to indicate intervals of values in accordance with their beliefs regarding the particular parameter. Alternatively, the fixed interval method, which is also based on percentiles, requires the analyst to specify a set of intervals that a specific quantity X can be contained within. The expert then gives the probability that X lies within each interval.

A method that has been applied previously in HTA is the histogram technique or probability grid. This is a graphical derivation of the fixed interval method where the expert is presented with possible values (or ranges of values) of the quantity of interest, displayed in a frequency chart on which he/she is asked to place a given number of crosses in the intervals or 'bins'. Histograms are appealing to even the least technical of experts (see **Box 2** for an example of this method in practice).

Eliciting Complex Parameters

Complex parameters include joint and conditional quantities, regression parameters, and correlation, and transitions in a multistate model (e.g., a Markov model). Perhaps the most common challenge arising with parameters that are interdependent is that a joint distribution may need to be elicited. The analyst can assess the model's sensitivity to variations in the correlation coefficient, or estimate the correlation as part of the elicitation exercise. There are a number of methods for eliciting correlations but no consensus regarding the most appropriate method. The methods include descriptions of likely strength of correlation, direct assessment, and the specification of a percentile for quantity X contingent on a specified percentile for quantity Y. However, the complexity of eliciting probability distributions that is conditional on other probability distributions is likely to be too cognitively difficult for many experts. In these circumstances, it may be appropriate to adopt a second best approach and elicit distributions conditional on means or best guesses. This was the approach used by Soares *et al.* (2011), where experts were first asked to record the probability (and uncertainty) of a patient's pressure ulcer being healed when they received treatment with hydrocolloid dressing. For experts who believed that the effectiveness of other treatments was different from the hydrocolloid dressing, the distribution of the relative treatment effects was elicited by asking experts to assume that the value they believe best represented their knowledge about the effectiveness of the comparator treatment, hydrocolloid dressing, was true (reference value). The reference value was the mode (or one of multiple modes, selected at random).

Conducting the Exercise

Explaining the Concept of Uncertainty

Eliciting measures of uncertainty can be complicated, particularly because one wants to ensure that data reflect uncertainty in the expected value rather than its variability or heterogeneity. This is largely a question of the format of the exercise; however, it can also be useful to present contrasting examples of uncertainty and variability to help the expert understand the key distinctions. Visual aids (such as the histogram) can be useful for the elicitation exercise and can help to reduce the burden on experts. It is also helpful to train them, especially when they have limited experience of elicitation. Experts will often respond better to questions and give more accurate assessments if they are familiar with the purpose and methods used in the elicitation exercise. Frequent feedback should also be given during the process and, if possible, experts should be allowed to revise their judgments.

Understanding the Impact of Bias and the Impact of Heuristics

It can be useful to understand how experts judge unknown quantities, in particular, whether they use specific principles or methods in order to make the assessment of probability

Box 2 Application of the histogram method (Soares *et al.*, 2011)

The histogram method is a fixed interval method. The range of values that the quantity may take is partitioned into intervals, and for each interval, information is collected on the probability of observing values. In an empirical application where uncertain quantities were elicited to inform a cost effectiveness model of negative pressure wound therapy for severe pressure ulceration, 23 nurses elicited 18 uncertain quantities. All uncertain quantities elicited were probabilities, thus a common scale was used (from zero to 100). A snapshot of the instrument used, to display the questions, is represented in **Figure 1**.

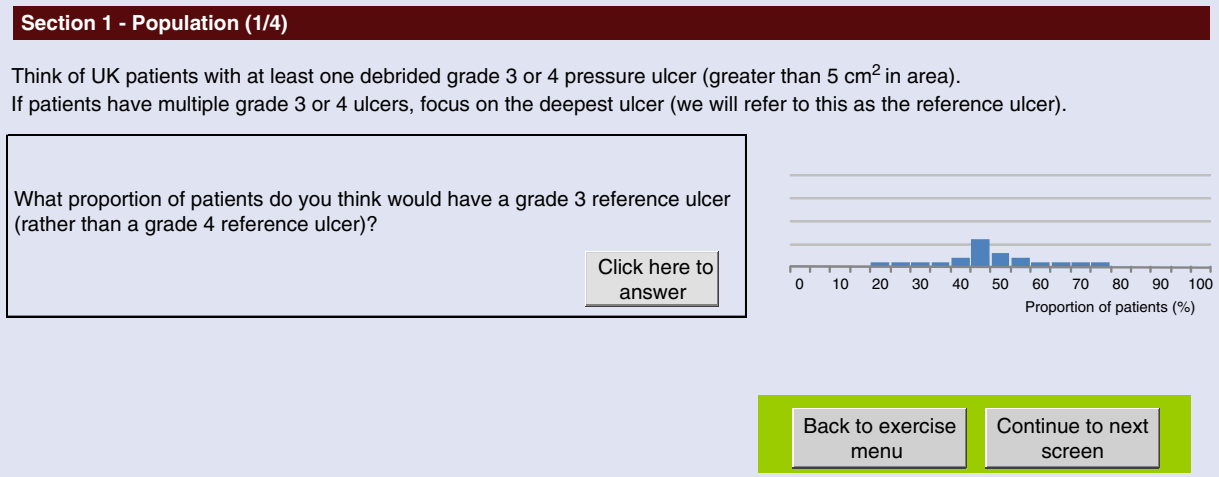


Figure 1 Graphic set-up of the instrument used in the elicitation exercise.

For each uncertain quantity, individual experts were asked to place 21 crosses on a grid defined to have 21×21 cells (**Figure 2**). Note that, for ease, the possible values that the quantity could take were made discrete (i.e., 0, 5, 10, …, 100). By placing the 21 crosses in the grid, the expert is effectively attributing a probability mass to each of the possible values, where each cross represents 4.765% probability. The expert can either express certainty by stacking all of the crosses in the same value (vertical column) or express the full certainty that a value is not possible by not attributing any crosses to it. By attributing one cross to each possible value, the expert is expressing the view that any value could be possible, i.e., full uncertainty.

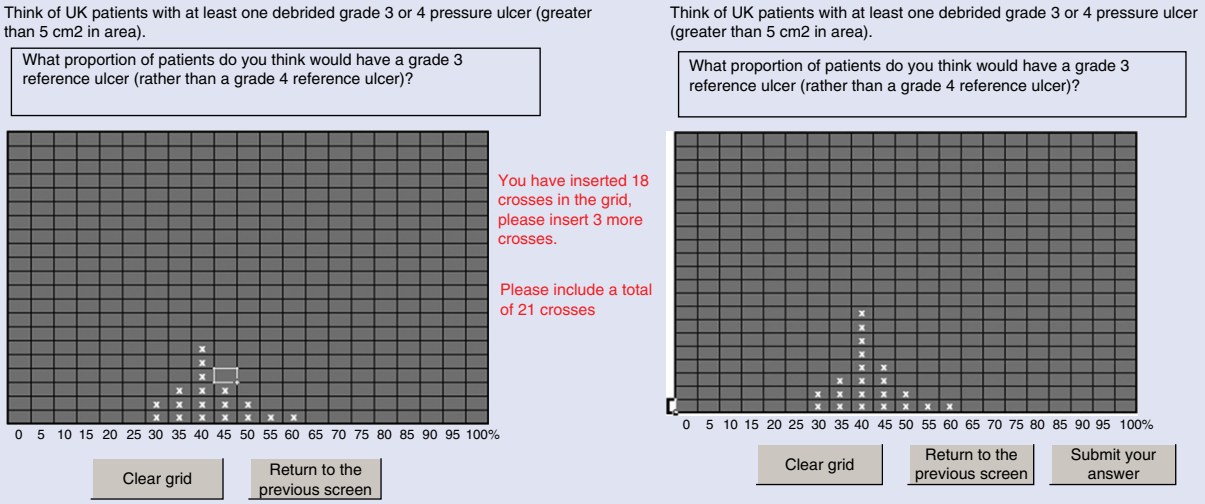


Figure 2 Graphic set up for the data capture histogram.

simpler. These heuristics are useful but can sometimes lead to systematic errors. Garthwaite *et al.* (2005) described the following heuristics: judgment by representativeness, judgment by availability, judgment by anchoring and adjustment, conservatism, and hindsight bias. All these issues should be considered when eliciting probabilities, as each can bias the assessments derived from experts, although the direction of bias is unlikely to be known. In addition, any motivational biases, bias from operational experience, and confirmation biases must be considered and appropriate measures taken to address their implications. Examples of biases in elicitation are described in Box 3.

Box 3 Examples of biases in elicitation

Biases in elicitation can include:

△ Biases associated with experts:

→ Motivation biases: for example, when experts have an incentive (e.g., financial) to reach a certain conclusion.

→ Cognitive biases: these commonly involve the use of heuristics to help reach decisions, solve problems, or form judgments quickly. Examples are:

Conjunction fallacy: When the probability of conjunction (combined) events is judged to be more likely than either of its constituents.

Availability: Where easy to recall events (like natural disasters) are judged to have high probabilities of occurring.

Hindsight bias: The tendency to overestimate the predictability of past events.

Anchoring effect: The tendency to rely on an anchor value that does not provide any information regarding the actual value.

△ Biases associated with elicitation methods:

→ Structuring elicitation questions: biases may arise from how the question is framed, for example, if relevant events have been omitted, experts are unlikely to consider them in replying. But biases can also occur when scales are used; for example, contraction bias occurs when the full range of a scale has not been presented to the expert.

→ Elicitation medium (e.g., interview or email survey) or aggregation method. Experts in group meetings (typically conducted when consensus aggregation methods are applied) tend to adopt a stronger position often resulting in overconfident statements.

Although it is not clear from the literature how most biases can be reduced/avoided, it is good practice to provide experts with an appropriate and comprehensive training session, which may make it clear what biases they might exhibit. The analyst can also attempt to avoid bias in designing the elicitation task, and avoid motivation biases in the selection of experts.

Synthesizing Multiple Elicited Beliefs

When judgments from several experts are required, it is often desirable to obtain a unique distribution that reflects the judgments of all of them. There are two broad methods for achieving this: behavioral and mathematical.

Behavioral approaches focus on achieving consensus. A group of experts is asked jointly to elicit its beliefs, as if it were a single expert, through the implicit synthesis of opinion and without aggregating individual opinions. In this approach, experts are encouraged to interact in order to achieve a level of agreement for a particular parameter. There are a number of behavioral aggregation techniques. The Delphi technique is probably the best known of these and it has been frequently applied to decision-making in healthcare. It involves sequential questionnaires interspersed by feedback and has characteristics that distinguish it from conventional face-to-face group interaction, namely, anonymity, iteration with controlled feedback and statistical response. The Nominal Group Technique is another popular consensus method. Here individuals express their own beliefs to the group before updating these on the basis of group discussion. The discussion is facilitated either by an expert on the topic or by a credible nonexpert. The process is repeated until a single value (or distribution) is produced.

However, there are problems with group consensus. First, consensus may not be easily achieved, and in some circumstances, there may be no value that all experts can agree on. Second, dominant individuals may so lead a group that they effectively determine the view of the whole group. Perhaps most importantly, however, is that a focus on achieving consensus means that behavioral approaches miss the inherent uncertainty in experts' beliefs regarding a parameter. There is a tendency for the group to be overconfident when reaching consensus regarding an unknown parameter.

Mathematical approaches to synthesizing multiple beliefs do not attempt to generate a consensus. Rather, they focus on combining individual beliefs to generate a single distribution using mathematical techniques. Aggregating individual experts' estimates into a single distribution is the preferred approach in applied studies. However, some studies have also used individual experts' assessments as separate scenarios for exploration. Synthesis of data from multiple experts often involves two steps: fitting probability distributions and combining probability distributions.

Fitting Probability Distributions

Fitting probability distributions to elicited data can be undertaken by the analyst either post elicitation or by asking the experts to assess fitting as part of the elicitation exercise. Parametric distributions can be fitted if an expert's estimates can be represented in such a way. The choice of parametric distribution is usually governed by the nature of the elicited quantities. If elicited priors are to be updated with sample information, then choosing conjugate distributions is advantageous for analytical simplicity. However, the development of computational methods has made it possible to choose nonconjugate distributions (i.e., distributions not from the same statistical family). Nonparametric methods can also be used. These do not assume that the data structure can be specified *a priori*; in effect, they have an unknown distribution.

Combining Probability Distributions

There are two main methods for combining probability distributions: weighted combination and Bayesian approaches. Weighted combination is referred to as opinion pooling, more specifically either linear opinion pooling or logarithmic opinion pooling. If $p(\theta)$ is the probability distribution for unknown parameter θ, in linear pooling, experts' probabilities are aggregated using the simple linear combination: $p(\theta) = \sum_i w_i \cdot p_i(\theta)$, where w_i represents a weight assigned to expert i. In logarithmic opinion pooling, averaging is undertaken using multiplicative averaging. These two methods can differ greatly, with the logarithmic method typically producing a narrower distribution for the parameter, implying less uncertainty in the estimate.

An example of the use of linear pooling is described by White *et al*. (2005), they have elicited expert opinion on treatment effects and the interaction between three trials. Experts are asked to assign a weight of belief (up to 100) to intervals of annual event rates. Experts' weights were

then combined by taking the arithmetic mean of individual assessments (linear pooling with equal weighting of experts).

More recently, there has been a move toward using Bayesian models for combining probabilities. Aggregation in a Bayesian model uses the experts' probability assessments to update the decisionmakers' own prior beliefs regarding an uncertain parameter. These methods have not yet been applied in HTA and the need for the decisionmakers' input is likely to be difficult to implement in practice.

If experts have been asked to express their beliefs regarding the value of an unknown quantity using a histogram, number of options are available for aggregation. Linear opinion pooling and Bayesian models can be used to aggregate parametric distributions, fitted to each expert's histogram. Alternatively, the empirical distributions derived can be combined to generate one overall empirical distribution.

Interdependence of Experts

Regardless of the method used to combine experts' probability distributions, an additional level of complexity is introduced when the assumption that experts provide independent beliefs is not sustainable. This is more likely if experts are chosen from the same professional organization or base their beliefs on shared experience or information. In this case, joint distributions should be used, incorporating the covariance matrix for the experts' assessments.

Assessing Adequacy

Four alternative measures have previously been described in the literature for assessing the adequacy of an elicitation: internal consistency, fitness for purpose, scoring rules, and calibration.

Internal Consistency

Internal consistency is particular relevant when eliciting probabilities. An expert's assessment of one (or more) unknown parameters should be consistent with the laws of probability. Achieving coherence may, however, involve more complex reasoning and, in the presence of such complexity, either incoherent judgments are transformed for further use or the exercise is constructed in order to minimize or eliminate incoherence. Qualitative feedback can also be useful in assessing internal consistency. Any discrepancies can be fed back to the experts and appropriate adjustments to assessments can be made.

Fitness for Purpose

Inevitably, some degree of imprecision will remain in elicited beliefs and their fitted distributions. Sensitivity analysis can be useful in discovering whether the ultimate results of the analysis change if alternative (but also plausible given the expert's knowledge) distributions are used. A commonly used sensitivity analysis in a Bayesian framework explores alternative prior distributions. If results do not change appreciably, then the distributions can be said to represent the experts' knowledge and are thus fit for purpose.

Scoring Rules

For parameters that are known or subsequently become known to analysts, comparisons can be made between elicited distributions and those known distributions. This provides an opportunity for assessing the 'closeness' of the elicited and actual distributions. The 'scoring rule' then attaches a reward (a score) to an expert using some measure of accuracy, with those gaining higher scores being regarded as performing better. Commonly used scoring rules are the quadratic, logarithmic, and spherical methods. In the example from Chaloner *et al.* (1993), elicitation was used to inform a model using the intermediate results of a randomized trial. On completion of the trial, comparisons were made between elicited estimates and those based on actual data. It was concluded that the elicitation exercise, although producing some thought-provoking results, did not necessarily predict trial outcomes with much accuracy. Although not done explicitly as part of the exercise, it would have been possible to score experts' beliefs retrospectively, possibly with a view to combining these with the experimental data.

Calibration

The most commonly used method for assessing the adequacy of elicitation is to measure experts' performance through calibration. The basic premise of calibration is that a perfectly calibrated expert should provide assessments of a quantity that are exactly equal to the frequency of that quantity. By asking experts to provide estimates of known parameters, their performance, in terms of distance between their estimates and the true value, can be determined. Unlike scoring rules, measures of performance such as calibration can then be used to adjust estimates of future unknown quantities. Alternatively, a recent example by Shabaruddin *et al.* (2010), used the mean number of relevant patients to derive a weighting for each expert. This was then used to generate weighted means in the linear pooling.

Discussion

Formally elicited evidence to parameterize HTA decision models is yet to be used widely. However, it has huge potential. Compared with many other forms, elicitation also constitutes a reasonably low cost source of evidence. However, the potential biases in elicited evidence cannot be ignored, and due to its infancy in HTA, there is little guidance to the analyst who wishes to conduct a formal elicitation exercise.

This article has summarized the main choices that an analyst will face when designing and conducting a formal elicitation exercise. There are a number of issues, of which the analyst should be particularly mindful, especially the need to characterize appropriately the uncertainty associated with model inputs and the fact that there are often numerous

parameters required, not all of which can be defined using the same quantities. This increases the need for the elicitation task to be as straightforward as possible for the expert to complete.

There are numerous methodological issues that need to be resolved when applying elicitation methods to HTA decision analysis. In choosing to use more complex methods of elicitation, it is also important to note that the complexity of many HTA decision models and the need to capture experts' beliefs, as inputs into these, creates a tension between generating unbiased elicited beliefs and populating a decision model with usable parameters. However, where experimental evidence is sparse, controversial, and difficult to collect, as far as emerging technologies, the need to explore the added value of elicited evidence seems particularly pressing.

See also: Adoption of New Technologies, Using Economic Evaluation. Economic Evaluation, Uncertainty in. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Infectious Disease Modeling. Information Analysis, Value of. Observational Studies in Economic Evaluation. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Problem Structuring for Health Economic Model Development. Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies. Synthesizing Clinical Evidence for Economic Evaluation. Value of Information Methods to Prioritize Research

References

Chaloner, K., et al. (1993). Graphical elicitation of a prior distribution for a clinical trial. Special Issue: Conference on practical Bayesian statistics. *Statistician* **42**(4), 341–353.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**(470), 680–700.

Shabaruddin, F. H., et al. (2010). Understanding chemotherapy treatment pathways of advanced colorectal cancer patients to inform an economic evaluation in the United Kingdom. *British Journal of Cancer* **103**, 315–323.

Soares, M. O., et al. (2011). Methods to elicit experts' beliefs over uncertain quantities: Application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration. *Statistics in Medicine* **30**(19), 2363–2380.

White, I. R., Pocock, S. J. and Wang, D. (2005). Eliciting and using expert opinions about influence of patient characteristics on treatment effects: A Bayesian analysis of the CHARM trials. *Statistics in Medicine* **24**(24), 3805–3821.

Further Reading

Bojke, L., et al. (2010). Eliciting distributions to populate decision analytic models. *Value in Health* **13**(5), 557–564.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. New York: Oxford University Press.

Jenkinson D. (2005) The elicitation of probabilities – A review of the statistical literature. *BEEP Working Paper*. Department of Probability and Statistics, Sheffield: University of Sheffield.

Leal, J., et al. (2007). Eliciting expert opinion for economic models. *Value in Health* **10**(3), 195–203.

O'Hagan, A., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester: Wiley.

Ouchi F. (2004) A literature review on the use of expert opinion in probabilistic risk analysis. *World Bank Research Working Paper 3201*. Available at: http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2004/04/15/000009486_20040415130301/additional/115515322_20041117173031.pdf (accessed 06.08.13).

Demand Cross Elasticities and 'Offset Effects'

J Glazer, Boston University, Boston, MA, USA, and Tel Aviv University, Tel Aviv, Israel
TG McGuire, Harvard Medical School, Boston, MA, USA

Glossary

Offset effects When use of one service 'offsets' or reduces use of another.

Sufficient statistic In welfare analysis, when a sufficient statistic is available, no data are informative about a welfare effect.

Introduction

The typical analysis of health insurance and service use considers coverage for a single aggregate commodity, 'health care.' It is natural to extend the analysis to more than one service, raising a number of issues in health insurance design. Fundamentally, two covered services can be substitutes or complements. 'Offset effects,' a term common in the empirical literature, refers to the substitute case, when use of one service 'offsets' or reduces use of another. The main insight regarding optimal insurance with multiple services is straightforward: When one service substitutes for another covered service, the increase in demand from insurance generates an efficiency gain from the decreased use of the other covered service. The reason for this is that the other service is itself insured and therefore to a degree 'overused.' The under appreciated subtlety in this result is the role of coverage for the 'other' service. Without coverage and overuse, there is no efficiency gain/loss with a change in demand for the other service. The role of coverage emerges in the analysis of multiple services, and has important implications for the way 'offset effects' should be measured and interpreted.

Concern about multiple services and substitutability and complementarity in insurance design need only be concerned with relationships with other covered services. Other services, if these are not part of the insurance plan even if they are health care services, are irrelevant for questions of optimal insurance. For example, suppose coverage for a certain prescription drug for pain offsets use of over-the-counter analgesics. Because these are not insured, there is no inefficiency associated with their use, and any 'offset' in the use of over-the-counter drugs is irrelevant for insurance design.

Coverage for the 'other good' plays a role in the empirical literature studying cross effects in demand. A large literature in health economics and health services research tests for 'offset effects.' The most active area for current research is on the cross effect of coverage for prescription drugs. Drug coverage is relatively new and variable. Furthermore, effective drug treatment for many, particularly chronic illnesses, might reasonably be expected to prevent/offset the need for other forms of care.

A related question is insurance coverage for 'prevention,' health care that affects the probability of illness. The argument for coverage for preventive services is similar to the offset argument, and rests on the presence of coverage of the service for the illness that would be prevented.

The article begins with a brief review of some of the empirical literature on offset effects, and then considers the issue from the standpoint of welfare economics and insurance design.

Empirical Literature Cross Elasticities

Much of the empirical research on cross elasticities in health care has focused on drugs. Ellison *et al.* (1997) studied cephalosporins, a class of anti-infectives, using IMS monthly time series data from 1985 to 1991, and found significant elasticities between some therapeutic substitutes. More recently, Ridley (2009) investigated cross-price elasticities for antiulcer drugs and drugs to treat migraines using data for 3 million people from a large pharmacy benefit manager (PBM) in the early 2000s. He found large effects on demand when drugs differed in the co-payment from other drugs in their class.

A particularly interesting case of a cross elasticity has emerged in statins, used to treat high cholesterol. In June 2006, the second largest-selling statin, Zocor, became available as generic simvastatin. Statin drugs had very high sales. In 2004, Zocor was the fifth largest selling drug worldwide in terms of dollar sales, and another statin, Lipitor, was the worldwide leader among all drugs from any class greater than $12 billion of sales annually. In response to the availability of generic simvastatin, managed care plans moved Lipitor to higher (less favorable) tiers (Aitken *et al.*, 2008 p. W157). One PBM moved Lipitor to tier 3 in January, 2006 in anticipation of generic simvastatin, and saw more than 40% of patients switch from Lipitor to a lower-tier statin (Cox *et al.*, 2007). Among those with co-payment differences of $21 or more, 80% switched.

It is typical in this literature to measure the 'offset effect' by the effect on total spending not just covered or plan spending on the 'other service.' For example, Shang and Goldman (2007) use Medicare Current Beneficiary Survey (MCBS) data from 1992 to 2000 to show that extra spending, measured by plan plus consumer medical costs, on drugs use induced by Medigap coverage, is more than offset by reductions in total health care spending. Hsu *et al.* (2006) compared medical spending for Medicare beneficiaries with a cap on drug coverage to those without a cap at Kaiser Permanente of Northern California before Medicare Part D. Drug spending was 28% less in the capped group but other

categories of expenditures were higher and total spending for all care was not significantly different between the groups, implying a near dollar-for-dollar offset in total costs. Gaynor *et al.* (2007) studied the effect of increases in co-payments charged for drugs among private employees on total (plan plus consumer) spending. Increases in nondrug spending, largely in outpatient care, offset $0.35 of each dollar saved in drug costs. An exception to the singular focus on total spending is the paper by Chandra *et al.* (2010), finding that the savings in costs due to higher co-payments for drugs were partly offset by higher spending on hospital services among retired state employees in California. They tracked offsets by payer because a primary (Medicare) and secondary (employer-provided supplemental) shared in offsets unequally. Approximately 20% of the cost savings from higher cost sharing for physician services and drugs was 'offset' by higher costs of hospitalization overall, with the offset concentrated among those with a chronic illness. Interestingly, as the authors point out, in the CalPERS case, this offset largely takes the form of a negative fiscal externality from the CalPERS supplemental policy (which saves from the elevated co-payments) to Medicare (which pays most of the costs of hospitalization).

The implicit logic in offset papers is that if total medical costs fall due to an increase in coverage, then the change in coverage is welfare improving (i.e., 'pays for itself'). This article argues that change in total medical spending, meaning the sum of plan and patient out-of-pocket spending, is not the right measure of the economic value (or cost) of a change in insurance coverage due to offset effects. Rather, changes in health plan costs alone measure the economic value of savings due to reductions in the use of other services. Applying methods reviewed by Chetty (2009) and Glazer and McGuire (2012) showed that a 'sufficient statistic' for evaluating the welfare effect of change in coverage for one that is good is the change in total plan-paid costs less the change in costs transferred to/from consumers. They derived an elasticity rule for when the offset effects of an improvement in coverage increases welfare.

A simple argument shows why total costs are not the right welfare measure of an offset effect. Suppose the plan covers just one service, 'health care,' and an increase in coverage of health care increases a consumer's total expenditures on health care. The consumer budget constraint implies that spending on some other noncovered services has to fall. This 'offset' says nothing about efficiency because coverage expansions are always exactly 'offset' in this trivial sense. What if the other affected spending were on another form of health care that was minimally covered in the plan, say for 1% of costs with consumers paying 99%? Logically, token coverage cannot imply that the full spending change as an offset should be counted.

A Model of Offsets in Health Insurance

Suppose a health plan covers services 1 and 2. Quantity of each received by a representative individual in the plan is x_1 and x_2 measured in dollars. Benefits to the individual are $B(x_1,x_2)$, where $B_i \geq 0$, $B_{ii} < 0$, $i=1,2$, with subscripts indicating partial derivatives. Letting c_i denote the co-payment charged for each unit of service i, then the individual demands service i to satisfy:

$$B_i(x_1,x_2) = c_i \quad i = 1,2 \qquad [1]$$

Let R denote the plan premium paid by the enrollee. Assuming the plan makes zero profit, the premium is

$$R(c_1,c_2) = (1-c_1)x_1 + (1-c_2)x_2 \qquad [2]$$

where (x_1,x_2) are given by eqn [1].

The individual's total utility from the plan is thus

$$U(c_1,c_2) = B(x_1,x_2) - c_1x_1 - c_2x_2 - R(c_1,c_2) \qquad [3a]$$

where (x_1,x_2) are from eqn [1] and R is from eqn [2]. Substituting for R to recognize that the individual pays for services by a combination of the cost sharing and the premium:

$$U(c_1,c_2) = B(x_1,x_2) - x_1 - x_2 \qquad [3b]$$

Consider now what happens to utility (welfare) eqn [3b] if the plan were to change the co-payment for service 2:

$$\frac{\partial U(c_1,c_2)}{\partial c_2} = (B_1-1)\frac{\partial x_1}{\partial c_2} + (B_2-1)\frac{\partial x_2}{\partial c_2}$$

$$= (c_1-1)\frac{\partial x_1}{\partial c_2} + (c_2-1)\frac{\partial x_2}{\partial c_2} \qquad [4]$$

The second equality follows from eqn [1].

Suppose co-payment for service 2 is reduced. If $\partial x_1/\partial c_2 > 0$, there is an offset effect and consumption of x_1 falls with this change. What happens to welfare? Equation [4] tells us how to value the offset. Reversing the sign of eqn [4] to get an expression in terms of plan shares, when co-payment for service 2 goes up (down), utility of the individual goes up (down) if and only if eqn [5] holds:

$$\left[\underbrace{(1-c_1)\frac{\partial x_1}{\partial c_2}}_{\text{Offset effect}} + \underbrace{(1-c_2)\frac{\partial x_2}{\partial c_2}}_{\text{Own-price effect}}\right] < 0 \qquad [5]$$

The intuition for this result is the following: The second term on the left-hand side of the inequality captures the inefficiency in consumption induced by the reduction in co-payment for service 2. With health insurance, the marginal benefit of health care is less than the marginal cost ($B_2 = c_2 < 1$), and the extra consumption of x_2 due to the reduction in co-pay creates additional welfare loss. In the conventional analysis of optimal health insurance, this welfare loss is weighted against the risk spreading gain to find the optimal co-payment, c_2. The first term on the left-hand side in eqn [5] is the offset effect due to the change in consumption (in this case reduction) of x_1. Just as with the own-price effect, benefits and costs both matter in valuing welfare of any offset effect. The $1(\partial x_1/\partial c_2)$ part is the reduction in total cost from the change in x_1 and, because $B_1 = c_1$, the $-c_1(\partial x_1/\partial c_2)$ part is the loss in benefits. Thus, the net welfare measure of offset effects is plan's savings: $(1-c_1)\partial x_1/\partial c_2$.

Changes in (consumer's) welfare to changes in plan costs can now be related. From eqn [2] it is known that when co-payment for service 2 changes, the change in the plan costs is

given by

$$\frac{\partial R(c_1,c_2)}{\partial c_2} = (1-c_1)\frac{\partial x_1}{\partial c_2} + (1-c_2)\frac{\partial x_2}{\partial c_2} - x_2 \qquad [6]$$

Equation [4] for changes in welfare, and eqn [6] for changes in plan costs, are the same except for the presence of x_2, the cost shifting effect of a change in c_2, a transfer ultimately paid by the consumer in any case. Using eqns [4] and [6] a rule for a welfare change, in terms of changes in plan-paid costs, can be stated.

Rule for Welfare Effects

The welfare effect of a change in coverage is equal to minus the change in plan costs net of the cost-shifting effect of the coverage change.

Proof. From eqns [4] and [6] the result is

$$\frac{\partial U(c_1,c_2)}{\partial c_2} = -\frac{\partial R(c_1,c_2)}{\partial c_2} - x_2 \qquad [7]$$

This rule for welfare effects constitutes, in Chetty's (2009) term, a 'sufficient statistic' for welfare evaluation of health insurance changes. The measure, change in plan costs less cost shifting, is equal to the welfare change, and thus yields an 'if and only if rule': Welfare goes up if and only if plan costs less transfers go down.

The rule brought out in this article can be used to interpret the existing logic of the offset literature which focuses on total costs, plan paid plus patient paid, and concludes that an improvement in coverage for good 2 is worthwhile if it 'pays for itself' in savings on good 1. Consider a reduction in c_2 that decreases use of covered good x_1 (an offset effect). Suppose the improvement in coverage for x_2 'pays for itself' in the sense that the reduction in the total cost of x_1 exceeds the increase in plan costs for x_2. This rule tells us that this condition is neither necessary nor sufficient for an increase in welfare. It is not necessary because the cost-shifting effect of the change in c_2 is disregarded for welfare. It is not sufficient because it is not total costs that measure the value of the offset, but plan-paid costs. Instead of looking for a coverage improvement to 'pay for itself,' the following simple rule, expressed in terms of demand elasticities for when an improvement in coverage improves welfare via an offset effect, is proposed.

A Simple Rule for When Offsets Increase Welfare

Welfare goes up with a decrease in c_2 (improvement in coverage) when the partial derivative in eqn [4] is negative, or alternatively

$$(1-c_1)\frac{\partial x_1}{\partial c_2} > -(1-c_2)\frac{\partial x_2}{\partial c_2} \qquad [8]$$

Putting this in elasticity form and dividing through by $-\varepsilon_{22}$ (a positive number), the criterion for a welfare improvement with a decrease in c_2 becomes

$$-\frac{\varepsilon_{12}}{\varepsilon_{22}} > \frac{(1-c_2)x_2}{(1-c_1)x_1} \qquad [9]$$

In eqn [9], ε_{12} is the cross and ε_{22} is the own-price elasticity with respect to c_2. The RHS of eqn [9] is positive and equal to the ratio of plan paid costs for service 2 to service 1. The following rule can now be stated: For a decrease in c_2 to improve welfare, the goods must be substitutes ($\varepsilon_{12} > 0$); and the ratio of the absolute values of the cross to the own-price elasticity must exceed the ratio of the plan paid costs for the two services.

The offset rule for welfare is simple to apply. Suppose it is known that the own-price elasticity of drugs is -1.0 and the cross-price elasticity for hospital services is $+0.2$. If the plan paid drug costs are less than 20% of the plan-paid hospital costs, an improvement in coverage for drugs improves welfare.

Attention to plan rather than total cost can change the tenor of the policy implications of offset effects, particularly for drug coverage where plan shares are relatively small. Turning to some results in significant recent offset papers illustrates the quantitative importance of the plan-cost perspective. Comparing the change in total costs for drugs and hospitals, Chandra *et al.* (2010, p. 208) found that a decrease in coverage for drugs reduced total drug costs by $23.06 per member per month, but increased total hospital costs by only $7.23 – the offset amounted to only a 1:3 ratio of hospital cost increases to drug cost savings, and in the authors' judgment was 'unlikely to be enough' to reverse the perceived value of the co-payment increase. However, taking the plan rather than total cost perspective it can be said that because drugs are covered at roughly 50% and hospital cost at 100%, the offset ratio doubles, to approximately 2 to 3. It should be noted here that the California change studied in Chandra *et al.* (2010) also involved increases to outpatient co-pays, which are ignored in this illustrative example. These increases also saved money, making the offset ratio 1:5. By ignoring this other benefit change in this discussion, it is, in effect, assumed that it is the drug coverage change that causes the offset.

Final Comments

In applied policy research, offset effects played an important role in the discussion about the design of optimal health insurance for mental health treatment, and more recently they do so in the case of coverage for drugs. Most public and private plans cover drugs, but the coverage is partial in the sense that a drug formulary typically excludes many drugs, and for those drugs that are covered, the percent paid by the plan is much less than for other health care services. Interestingly, the co-payment for generic drugs is often so high that it exceeds the acquisition cost to the health plan. The ideas in this article about valuing offset effects have the most current direct application to the question of coverage for drugs. If health insurance markets worked perfectly, competition would maximize welfare of the representative consumer, implying the efficiency issues discussed here would be taken care of in competitive equilibrium. Health insurance markets are fraught with sources of market failure, however, such as moral hazard, adverse selection, imperfect competition, externalities due to the participation of multiple insurers, as well as concerns about equity. In many cases there can be little assurance that

market forces alone will lead to optimal coverage, leaving a role for calculations of the type illustrated here.

The major limitation of this rule for offsets and model setup generally, stems from the assumption that quantity is determined by the equality of marginal benefit to the consumer/patient and patient co-payment. Although the standard demand model is widely applied in theoretical and empirical health care research, it is also seriously questioned as a basis for describing the outcome of patient–provider interactions. Effective physician agency on behalf of the patient would be consistent with this approach, but it is acknowledged that the marginal benefit–marginal cost equality is still a strong assumption. Relatedly, health economists doubt whether consumer demand should be interpreted as marginal benefit when assessing the efficiency of changing coverage. Perspectives from 'value-based insurance design' and behavioral economics both question the conventional welfare framework for assessing the efficiency cost of added coverage for a service.

See also: Efficiency in Health Care, Concepts of. Evaluating Efficiency of a Health Care System in the Developed World. Resource Allocation Funding Formulae, Efficiency of. Value-Based Insurance Design

References

Aitken, M., Berndt, E. and Cutler, D. (2008). Prescription drug spending trends in the United States: Looking beyond the turning point. *Health Affairs* **28**(1), W151–W160.

Chandra, A., Gruber, J. and McKnight, R. (2010). Patient cost-sharing, hospitalization offsets in the elderly. *American Economic Review* **100**(1), 193–213.

Chetty, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics* **1**, 451–487.

Cox, E., Klukarni, A. and Henderson, R. (2007). Impact of patient and plan design factors on switching to preferred statin therapy. *The Annals of Pharmacotherapy* **41**, 1946–1953.

Ellison, S., Cockburn, I., Grilichres, Z. and Hausman, J. (1997). Characteristics of demand for pharmaceutical products: An examination of four cephalosporins. *RAND Journal of Economics* **28**(3), 426–446.

Gaynor, M., Li, J. and Vogt, W. B. (2007). Substitution, spending offsets, and prescription drug benefit design. *Forum for Health Economics and Policy* **10**(2), 1–31.

Glazer, J. and McGuire, T. G. (2012). A welfare measure of 'offset effects' in health insurance. *Journal of Public Economics* **96**, 520–523.

Hsu, J., Price, M., Huang, J., et al. (2006). Unintended consequences of caps on Medicare drug benefits. *New England Journal of Medicine* **354**(22), 2349–2359.

Ridley, D. (2009). Payments, promotion and the purple pill. Fuqua School of Business, Duke University, unpublished.

Shang, B. and Goldman, D. P. (2007). Prescription drug coverage and elderly medicare spending, NBER working paper 13358. Available at: http://www.nber.org/papers/w13358 (accessed 26.07.13).

Further Reading

Duggan, M. (2005). Do new prescriptions pay for themselves? The case of second-generation antipsychotics. *Journal of Health Economics* **24**(1), 1–31.

Gibson, T. B., Mark, T. L., Axelsen, K., et al. (2006). Impact of statin copayments on adherence and medical care utilization and expenditures. *American Journal of Managed Care* **12**, SP11–SP19.

Goldman, D. P., Joyce, G. F. and Karaca-Mandic, P. (2006). Varying pharmacy benefits with clinical status: The case of cholesterol-lowering therapy. *American Journal of Managed Care* **12**(1), 21–28.

Demand for and Welfare Implications of Health Insurance, Theory of

JA Nyman, University of Minnesota, Minneapolis, MN, USA

Introduction

Importance of the Theory

Why do consumers purchase health insurance? To purchase anything, the consumer must give up something, and in the case of health insurance, that 'something' is the premium payment. Although the nature of the premium payment is clear (to both consumers and economists), what is not clear is the nature of the benefits that consumers receive in return. This represents the central objective and the challenge of health insurance theory: to describe just what it is that consumers receive in return for the premium. If this is known, then why consumers purchase health insurance will be known.

This is an important question because it can affect consumer welfare in fundamental ways. From the perspective of the insurance firm, if insurers knew precisely what it is that people value in insurance, they would be able to design more competitive insurance contracts, contracts that provide more of what consumers want to purchase. From a public policy perspective, policy makers would be able to design more efficient and effective government health insurance programs, implement more equitable subsidies and taxes, and encourage more efficient behavior with regard to the types and amount of health care insured consumers purchase. From a larger social perspective, if it were known why consumers purchase health insurance, politicians would better know the value of health insurance relative to other goods and services, and thereby better understand the importance of health insurance programs compared to all the other programs that government could sponsor.

Complexities of Health Insurance

Although this might seem like a relatively straightforward exercise, it is not. Insurance contracts have a number of complexities that make them difficult to analyze. Here is a listing of the most important ones. It should be noted that many of these complexities were identified by Kenneth Arrow in his famous 1963 paper on the characteristics of the medical care portion of the economy that make the sector unusual.

First, there is the uncertainty with regard to illness itself: not everyone becomes ill during the contract period and many of the benefits that consumers derive from paying a premium occur only if they become ill. Payoffs that are contingent appear in many types of contracts so they are not unusual, but they always make things more complex because they require the consumer to think about what might happen in the future.

Second, because illnesses vary, there is uncertainty with regard to the cost of treating illness. Some illnesses require health care expenditures that are relatively affordable to the typical consumer, but other illnesses are catastrophically expensive. Not only do the costs of different illnesses vary, but also the resources available to individuals if they were to remain uninsured and had to pay for health care themselves. That is, some consumers who become ill are rich and some are poor. On top of that, the diseases and the procedures used to treat them may also reduce the budget if the consumer is no longer able to work and make income. The variation in economic circumstances of consumers interacts with the variation in the cost of the illness, and both conspire to make a large portion of health care expenditures unaffordable to a substantial segment of the population. This complexity must also be accounted for in the theory.

Third, uncertainty also occurs with regard to the effectiveness of the health care in treating the disease. Sometimes the health care cures the disease, and sometimes it does not. Indeed, sometimes the health care is represented only by the palliative care during the short period before death. Although the variability in the effectiveness of the health care is a consideration in the purchase of insurance, it is clear that modern health care is often effective and for that reason, can be very valuable to the consumer. Thus, the value of the health care covered by the insurance benefit is a consideration in determining why consumers purchase insurance. This is especially true in light of implication of the second complexity that sometimes the health care would not be affordable and thus accessible to the consumer without insurance.

Fourth, the contingent benefit of insurance is based on the consumer transitioning from a state of being healthy to a state of being ill. The change in health state clearly affects how one values medical care – what 'healthy' person would value chemotherapy or a leg amputation enough to 'consume' it? Sometimes, the change in health state can also affect how consumers value the other goods and services that can be purchased. For example, some illnesses can be in the form of a broken bone or a minor respiratory disease, where it is clear that one is feeling poorly on a temporary basis and the state of illness represents largely an inconvenience. Other illnesses, however, may have severe symptoms in terms of pain and ability to function normally, be chronic, or threaten the lives of the individuals suffering from them. Thus, when thinking about the value of all the benefits of an insurance contract, the consumer would likely need to consider how they would regard the benefits of insurance if they were filtered through the perspective of being in an ill state. In the ill state, consumers may appreciate the various aspects of life – both the medical care and the income to spend on entertainment, travel, and other consumer goods – differently than in a healthy state, and this would bear on how the benefits of insurance are perceived and evaluated. Theorists who desire to model why people purchase insurance would need to acknowledge this change in perspective in order to produce a complete theory.

Fifth, health insurance contracts are not perfect. Although we may think about illness as an exogenous event that we have no control over, in actuality, we have a great deal of control over whether we become ill. For example, whether we develop heart disease is associated with a number of discretionary

 doi:10.1016/B978-0-12-375678-7.00906-8

behavioral choices – whether we smoke, are overweight, exercise, eat cholesterol-laden foods, etc. Insurance contracts (so far) do not distinguish between illnesses that are brought on by the behavior of the insured and those that are caused by factors beyond the control of the individual. The problem this creates for insurance is that sometimes being insured might alter the extent to which a consumer acts to avoid disease. 'Moral hazard' is the term that those in the insurance business use to describe the changes that occur in behavior of the insured and '*ex ante* moral hazard' is the term used by economists to describe the type of behavioral change where the probability of becoming ill increases when an individual becomes insured.

Sixth, most health insurance contracts simply pay for the sick consumer's health care. As a result, the amount of the insurance benefit when ill is not fixed in advance of becoming ill (nor is the benefit even totally dependent on becoming ill). Insurers often pay for more health care than the ill consumer would pay for if they had remained uninsured. '*Ex post* moral hazard' is the term used to describe the type of behavioral change where once insured persons become ill, they purchase more health care and incur greater expenditures than they would if they were not insured and were paying for the care themselves.

And finally, the basic idea behind insurance is that many people who are not ill pay into a pool in order to benefit the few members of the pool who become ill during the period of insurance coverage. This means that one of the fundamental incentives for prospective purchasers of insurance is to try to join the pool 'after' one becomes ill, in order to avoid paying premiums during the years when one is not ill. This phenomenon is called 'adverse selection' and is represented by the tendency of those who purchase insurance to be sicker or more prone to becoming sick, and therefore more costly to insure, than the average person. If the insurer does not catch this bias and charge these people higher premiums, the firm would pay out benefits that are greater than the premiums it takes in. Again, health insurance contracts are not perfect.

Modern health insurance plans often provide other benefits – the ability to bargain down producer prices, the evaluation of new technologies for effectiveness, the screening of physicians and other providers for quality – that add to the complexity, but those that are listed above represent the major complexities associated with the *quid pro quo* of the traditional insurance contract. In the discussion that follows, we consider how improvements in our understanding of insurance have coincided with increases in the benefits that are recognized to derive from insurance. We begin, however, with the conventional theory that the demand for health insurance is simply related to the avoidance of the uncertainty associated with illness and the loss of income that paying for one's own health care would entail.

Conventional Insurance Theory

The Gain from Certainty

The conventional theory of demand for 'health insurance' was originally borrowed from the theory of the demand for 'insurance,' which was concerned primarily with a type of indemnity policy where the consumer possesses a certain asset for which they desired protection from loss. For example, a homeowner might want protection from fire. The consumer has the choice between remaining uninsured and accepting the chance that the asset and its value might be lost to fire, or paying a premium for an insurance contract that would pay the consumer a lump-sum payment equal to the value of the asset if the asset were lost. Assuming that there is no difference between the premium payment and the expected loss if uninsured – that is, assuming that the insurance premium is actuarially fair and nothing extra is included in the premium to cover the administrative costs of the insurer – the consumer is better-off with insurance.

The insurance decision for this type of loss was laid out in 1948 by Milton Friedman and L. J. Savage in what has come to be regarded as the seminal article in the health economics literature (Friedman and Savage, 1948). **Figure 1** shows the fundamental relationship that economists assume exists between utility, on the one hand, and either income or wealth, on the other. Utility increases with income or wealth, but at a decreasing rate. The shape of this curve, U, derives from that intuitively appealing principle that consumers would gain more utility from a given amount of additional income or wealth (that is, consumers would value or appreciate it more) if they were poor than if they were rich. For example, a consumer with \$20 000 in wealth gains more utility from an additional \$1000 than he would if he had started out with \$100 000 in wealth.

The gain from purchasing insurance can be demonstrated using **Figure 1**. A consumer starts out with assets (or income, but for simplicity, the discussion will use assets) of \$100 000 and is faced with a 50% chance of becoming ill and incurring a \$80 000 loss due to the need to purchase a medical procedure. The utility function, U, indicates the utility of \$100 000 is $U(\$100\,000)$ and the utility of \$20 000 is $U(\$20\,000)$. Without insurance, the expected value of the consumer's assets is \$60 000 because he starts out at \$100 000, but loses \$80 000 with a 50% probability, so the expected loss is \$40 000. Similarly, with regard to utility, without insurance, the consumer starts out at utility of $U(\$100\,000)$ but falls to $U(\$20\,000)$ with a 50% probability, so the expected utility is $EU(\$60\,000)$ as in **Figure 1**. Thus, point A represents the expected position of the uninsured consumer facing a loss of \$80 000 with a 50% chance.

Assume that the insurer charges the actuarially fair premium, one that reflects only the expected payout and none of the administrative costs or profits. The actuarially fair premium is \$40 000 because that is the amount that the insurer expects to payout for each person that is insured for this illness (that is, \$80 000 payout times the 0.5 chance of illness, for each person who is insured). If the consumer pays such a premium and purchases insurance, she will have \$60 000 regardless if healthy or ill. If the consumer stays healthy, she would start out with \$100 000 in assets, would have no health care expenditures and receive nothing in payout from the insurer, but would pay a \$40 000 premium, leaving \$60 000 in assets. If the consumer becomes ill, she would start out with \$100 000 in assets, would incur health care expenditures of \$80 000, would receive \$80 000 from the insurer, but must pay a \$40 000 premium, again leaving \$60 000 in assets. Thus,

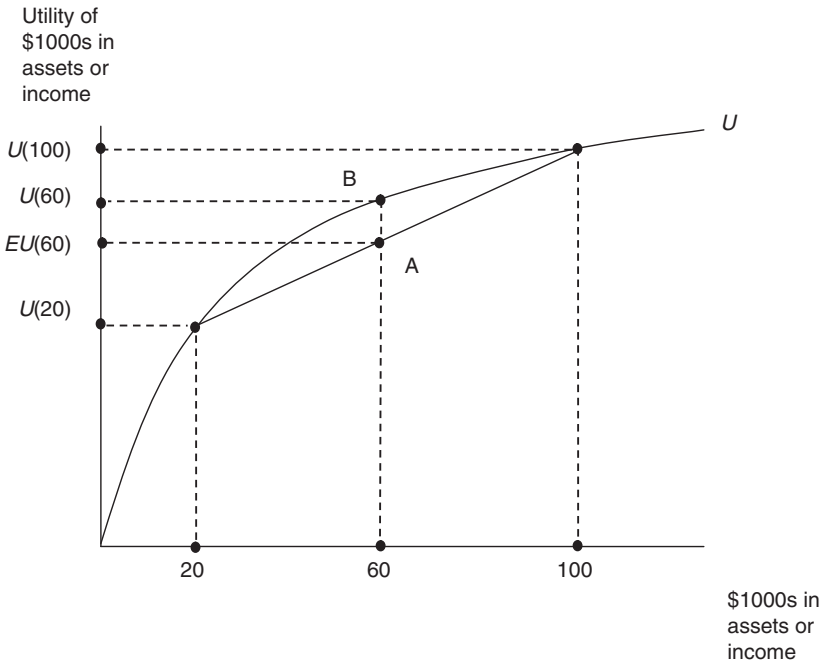


Figure 1 Gain from insurance under conventional theory.

regardless of whether the consumer stays healthy or becomes ill, if she purchases this insurance, she has $60 000 in assets. The utility of $60 000 with certainty is determined by the utility function as U($60 000), and so with insurance, the consumer would be at point B in Figure 1. The gain in utility from insurance is measured by the vertical distance between points B and A, or the difference between U($60 000) and EU($60 000) on the vertical axis. This difference in utility is the welfare gain from buying health insurance under the conventional theory, and represents the sole reason for purchasing it under this theory.

To this theory was added the complexity of loading fees (the additional amount that the insurer includes in the premium to cover administrative costs and profits), but the basic source of the gain remained the same. Friedman and Savage interpreted this gain as satisfying the consumer's preference for certainty, as opposed to uncertainty, and many have viewed the benefits of health insurance from this perspective. Based on this theory and the utility gain from the certainty that health insurance contracts provide, Arrow concluded in his 1963 article that the case for health insurance was 'overwhelming.' This is the theory that has been used over the years to explain why consumers purchase health insurance.

Limitations of the Theory

The theory, however, has a number of limitations. First, the theory would only apply to those medical procedures that are affordable. This is because there is no uncertainty if the loss cannot occur, and this would most likely be the case if the cost of the procedure is so high that the ill consumer cannot pay for care. It is possible that the consumer might be able to borrow the additional resources, but an uncollateralized loan for a risky procedure would be difficult to obtain and so this option is limited at best. Saving for the procedure is also possible, but saving when ill may be out of the question because of the ill consumer's diminished earning capacity and the limitations on time available. Thus, this theory does not recognize that many procedures and health care episodes may be too expensive to be financed privately, save for insurance. This is an important omission because, given that about half of all health care expenditures in the US are incurred by the top 5% of spenders (Stanton and Rutherford, 2006) and that those under 65 in the lowest quartile of the income distribution in the US have virtually no net worth and those in the second lowest quartile of the income distribution have net worths that average close to their annual income (Bernard *et al.*, 2009), procedures that are too expensive for consumers to afford to purchase privately make up a substantial proportion of health expenditures in the US.

Second, the 'loss' in this theory is the income or assets lost due to the spending on the medical care. In contrast to the simple destruction of an asset (e.g., a house burning down), the spending on medical care is not really a loss, but part of *quid pro quo* transaction where the consumer spends income or wealth to obtain medical care. The medical care that the consumer obtains in return for this 'loss' may be very valuable, but the value of the medical care does not appear in the model.

Third, the model assumes that the utility that the consumer gains from income or assets when ill is the same as the utility when healthy. For example, it assumes that $100 000 in assets is just as valuable when healthy and being spent on restaurant meals, gas for the car, etc., as it would be when ill and being spent on restaurant meals, gas for the car, and a $50 000 medical procedure that saves the consumer's life. In fact, this model implicitly assumes that the utility from income is

derived 'only' from the nonmedical care purchases that one can make with income, and that becoming ill does not alter at all the utility that is derived from these purchases. And as was noted, the utility from income that can be used to purchase medical care when ill simply does not enter the model.

Fourth, as mentioned, the motivation for purchasing insurance under this model was interpreted by Friedman and Savage to reflect the consumer's natural preference for certain ones over uncertain ones and that this preference for certain losses summarized the reason why consumers purchase health insurance. Whether consumers actually do have a preference for certain losses over uncertain ones has been tested by Kahneman and Tversky. In a series of experiments that led to the formulation of prospect theory (and to a Nobel prize in economics for Kahneman), these researchers found that consumers generally prefer uncertain losses to certain ones of the same expected magnitude, the opposite of what the conventional insurance theory asserted (Kahneman and Tversky, 1979). If this preference for uncertain losses is generally true of consumers, as the experiments appeared to show, then the demand for health insurance cannot be attributed to a preference for certain losses.

Fifth, the payoff in this theory is in the form of a lump-sum transfer of income to the insured. Although such a policy is possible and actually exists for some types of insurance, such as personal accident insurance (e.g., policies that pay $50 000 for the loss of sight in one eye), most health insurance policies pay off by paying for care (or a portion of it after some copayment by the insured). Moreover, spending (that is, the loss) with and without insurance is assumed to be the same in this simple model. As a result, this model does not allow for moral hazard.

Moral Hazard Welfare Loss

Of all the limitations of this risk avoidance model, the one that was seized on initially was the lack of recognition of moral hazard – but not all moral hazard, only *ex post* moral hazard. As mentioned earlier, economists distinguish between two types of moral hazards. *Ex ante* moral hazard occurs when the consumer takes less care to avoid losses if insured than if not insured. For example, because health expenditures are covered, a consumer might have an increased probability of illness if insured, compared with if uninsured. *Ex post* moral hazard was defined originally as the additional spending that occurs after one becomes ill, insured versus uninsured. Recently, some economists have suggested that *ex post* moral hazard is represented only by the portion of the change in this behavior that is due to a response to prices, but that was not the original view. This distinction has come about only recently, because for a long time it was thought that *ex post* moral hazard was 'only' a response to prices.

In a 1968 comment on Arrow's (1963) article, Pauly wrote what was to become 'one of the,' if not 'the,' most influential articles in the health economics literature. Pauly's article led to almost a 'preoccupation' among American health economists with the notion that the basic problem with the high health care costs in the US was the consumption of too much care (and, implicitly, not the high prices of health care). This perspective, in turn, led to important policy initiatives in the US over the next 30 or 40 years that focused on reducing the quantity of care: The introduction of copayments into insurance policies, the adoption of managed care, and the promotion of consumer-driven health care (where policies with large deductibles are paired with health savings accounts). Indeed, some economists argued during this period that high prices of medical care were beneficial because they choked off demand by making coinsurance rates more effective.

Pauly's argument recognized that health insurance policies paid off not by paying a lump-sum amount when the consumer became ill, as the Friedman and Savage model assumed, but by paying for any health care that the individual consumed. Thus, the impact of insurance on the consumer's behavior was essentially to reduce the price of health care, to which the consumer responded by demanding a greater quantity of care. **Figure 2** shows the observed or Marshallian demand for health care, D, by the individual consumer and the quantity of health care consumed, m_u, if uninsured and if 1 is the price of a unit of medical care, m. If the consumer becomes insured under a contract where the insurer pays for a percentage of care represented by $(1-c)$ with c representing the coinsurance rate, then the price of care that the consumer faces effectively drops to c and the consumer purchases m_i quantity of health care. So, *ex post* moral hazard is represented by the increase in consumption from m_u to m_i.

The problem with moral hazard according to Pauly's model is that the additional care is worth less than the cost of the resources used to produce it. If the health care market is competitive, then the market price of health care, 1, would also represent the marginal cost of the resources used to produce the care, that is, the value of the goods and services that the same resources could have been produced in their next most valuable use. The marginal cost curve represents the cost of producing each of the units of health care, given the assumptions of the model. The value of health care is measured by the willingness to pay for it, as shown by the height of the demand curve at each level of m. For example, according to the demand curve, the willingness to pay for the m_u unit of medical care is just equal to 1, the market price. If the price were to drop to c because of insurance, the additional health

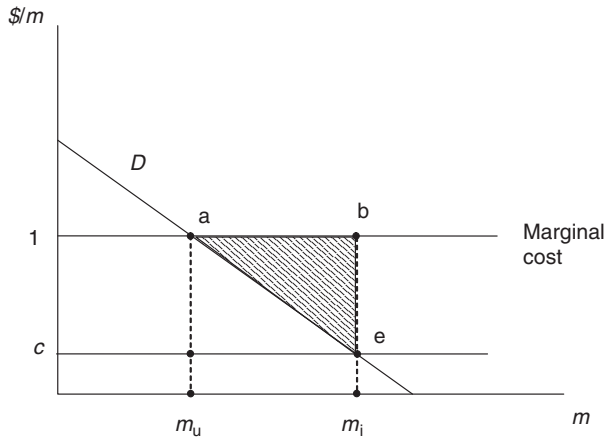


Figure 2 Welfare loss from moral hazard under the conventional theory.

care consumed, that is, the moral hazard, is $(m_i - m_u)$. The value of this additional care is represented by the area under the demand curve, area aem_im_u. The cost, however, is the area under the marginal cost curve, or abm_im_u. Costs exceed the value by the area *abe*. This area, then, represents the welfare loss associated with moral hazard.

Empirical and Professional Support

With the publication of Pauly's paper, the conventional theory of the demand for health insurance was now set. The demand for health insurance was represented by the gain from averting the risk of loss, but it was necessary to subtract from this gain the welfare loss from *ex post* moral hazard. Pauly thought that the loss was potentially so important that the net effect, 'could well be negative' (Pauly, 1968, p. 534), implying that insurance could make the consumer worse-off, especially if the government mandated its purchase. In 1973, Martin Feldstein empirically estimated the net gain from health insurance in the US based on conventional theory and concluded that "the overall analysis suggests that the current excess use of health insurance produces a very substantial welfare loss" (Feldstein, 1973, p. 275). Feldstein argued that raising the coinsurance rate to 67% across the board would improve welfare. This view persisted over the remainder of the century and into the next. In 1996, for example, Willard Manning and Susan Marquis found that low coinsurance rate health insurance policies also resulted in a net welfare loss based on conventional theory and concluded that a coinsurance rate of approximately 45%, also across the board and with no limit on out of pocket spending, would be optimal.

During the same period, a health insurance experiment – the most costly social experiment ever performed in the US – was also conducted by the RAND Corporation. The RAND Health Insurance Experiment randomly assigned some participants to receive free care and others to care with some form of cost sharing. As was expected, those assigned to free care consumed more medical care – both physicians services and hospital admissions – than those who had to pay for a portion of the cost of their care, but more importantly, aside from better correction of vision problems, there was no significant improvement in health for those who received more care (Newhouse, 1993). Thus, the influential findings of the RAND health insurance experiment fit the Pauly's model like a glove: Insurance generated additional care, but the additional care was not very valuable because it did not result in any important improvements in health.

Why Pauly's focus on *ex post* moral hazard caught on among American economists is not clear: after all, two other sources of inefficiency in health insurance contracts – *ex ante* moral hazard and adverse selection – were also broadly recognized at the time. *Ex ante* moral hazard would have generated a similar welfare loss from the reduction in purchase of efficient health preservation services and the increase in the purchase of inefficient health recovery services once ill (medical care), because the prices of the recovery services were made to be artificially low relative to the prices of the health preservation activities. The inefficiency associated with adverse selection (the nonpurchase of insurance by those who would have purchased insurance were it not for the high premiums caused by adverse selection) was also broadly recognized at the time, but this inefficiency did not rise to the level of a component of the basic theory. Although the confirmatory studies by influential economists were clearly a factor, perhaps even more important for its appeal was that it underscored the importance of competitive prices, which was consistent with the prejudices of economists. Moreover, its diagrammatic argument was accessible, elegant, and easily taught.

Alternative Theory

The Gain from an Income Transfer When Ill

Recently, an alternative theory has been suggested that incorporates all the factors that were limitations to the conventional theory (Nyman, 2003). The basic notion is that health insurance represents a *quid pro quo* contract where the consumer pays an actuarially fair premium to the insurer when healthy in order to receive a lump-sum income payment if the insured were to become ill during the period of time covered by the insurance contract. If the insured consumer does not become ill, the contract holder simply relinquishes the insurance premium. An actuarially fair health insurance contract is therefore purchased because the utility gained from the additional income if ill exceeds the utility lost from paying the premium if the consumer remains healthy.

This theory is fundamentally different from the Friedman and Savage theory because it does not incorporate a designated loss when ill as part of the insurance decision. That is, there is no loss of assets or income from illness recognized by the theory. As a result, there is no 'preference for certainty' in this model and no 'smoothing of income' across the states of the world, as some have interpreted the Friedman and Savage approach to imply. The only loss of income that occurs in the alternative model is the loss of the insurance premium if the insured person remains healthy. Because the theory does not incorporate a designated loss, the income payment when ill can be any amount and does not need to reflect the spending that would occur without insurance.

Advantages over Conventional Theory

This theory has a number of advantages over conventional theory. First, the theory is not limited to explaining the demand for insurance coverage for only that portion of medical care that the consumer could otherwise purchase if uninsured (the portion that would generate a loss of income and/or wealth due to such spending), but it also explains why consumers purchase insurance coverage for medical care spending that would exceed the consumer's resources. Indeed, the access that the insurance payoff provides to that medical care that would otherwise be unaffordable is one of the main reasons why insurance is purchased under this alternative theory.

Second, the value of insurance is directly linked to the value of the medical care that the consumer can purchase as a result of being insured and receiving an income payoff when ill. As was mentioned, some modern medical care is

ineffective, but much of it is very effective and can generate large health improvements, both in terms of limiting the negative effects of illness and expanding life expectancy. The health improvements derived from this medical care can be very valuable to consumers, and there is often no alternative (private) means for obtaining this care other than to purchase insurance. This value, entirely missing from the conventional model, is emphasized in the alternative model.

Third, this model recognizes that consumer preferences can be altered when the consumer becomes ill by specifying two utility functions for both consumer commodities and medical care: one when healthy and another when ill. This allows for the consumer to incorporate a different evaluation of consumer goods and services in the two states. For example, is spending on traveling or home improvements as valuable when ill as when healthy? But, more importantly, it allows for a different evaluation of medical care by the consumer in the two states. For example, is spending on a new heart valve or leg amputation as valuable when healthy as when ill? It recognizes that illness changes preferences so that a coronary bypass procedure or course in chemotherapy now becomes valuable, whereas it would reduce utility if purchased when healthy. Under this theory, insurance is the mechanism by which an increase in income occurs at precisely the same time as the onset of illness generates a change in preferences, making it possible to purchase the medical care services that would not be valued or purchased, given preferences when healthy.

Fourth, rather than trying to explain the purchase of insurance by claiming that consumers generally exhibit a preference for certain losses over uncertain losses of the same expected magnitude – a claim that has been thoroughly discredited and indeed proved to be diametrically opposed to the preferences of most consumers by the empirical studies underlying prospect theory – the alternative theory suggests that preferences for certainty are not part of the demand for health insurance at all. Uncertainty exists in life, clearly, but insurance cannot do anything about it other than to coordinate the uncertain occurrence of illness with an equally uncertain payment of income.

Fifth, the conventional theory focuses on a welfare loss from *ex post* moral hazard, all of which is deemed to be welfare decreasing because it is generated by a reduction in price and a subsequent movement along the consumer's demand curve with a payment of income. It is as if a hospital suddenly announced a sale on coronary bypass procedures and additional shoppers flocked to take advantage of the bargain, whether they were ill and needed a bypass operation or not. With the alternative theory, the price reduction is the vehicle by which income is transferred from those who purchase insurance and remain healthy to those who purchase insurance and become ill. As a result, the price reduction applies only to those who are ill enough to need an important health care intervention and the income transfer within the price reduction works to shift out the demand curve of those who are ill. It is as if a hospital suddenly announced a sale on coronary bypass operations and those additional patients who now flocked to the hospital are only those who suffered from coronary artery disease and could not afford to purchase the procedure at the existing market prices.

Welfare Implications of Moral Hazard

Actually, the moral hazard response to the price reduction under the alternative theory requires some additional explanation because it can be partly a response to the price decrease that is used to transfer income and partly due to the income transfer itself. Indeed, this is one of the important implications of the new theory: Some of the additional spending due to insurance (moral hazard) is efficient and due to the income transfer, and some is inefficient and due to using the price reduction to transfer income. It is the efficient moral hazard that represents one of the most important reasons for purchasing insurance. At the same time, inefficient moral hazard also exists, but it is not quite the same as described by Pauly (1968). A short explanation is required.

As described earlier, conventional theory suggests that the response to insurance can be described as a movement along the observed or Marshallian demand curve. In **Figure 2**, at the market price, 1, a certain amount of medical care, m_u, is demanded. If insurance was purchased, the price of medical care faced by the consumer is c, then m_i would be purchased. Thus, conventional theory uses the Marshallian demand curve to show the response to insurance. With insurance, however, the price does not simply drop due to exogenous market forces as would be consistent with the Marshallian demand, but instead, the price reduction must 'be purchased' by paying the premium for an insurance contract. Moreover, the greater the price reduction or lower the coinsurance rate specified in the contract, the greater the premium that must be paid. The payment of the premium reduces the amount of income remaining that can be used to purchase medical care after insurance is purchased, and thus reduces the amount of care that is purchased at the lower insurance price. (Medical care is a 'normal good' implying that less would be purchased if the consumer had less income.) For example, for a family of 4 making $40 000, an 80% reduction in the price that occurred as a result of market forces would generate a greater increase in the quantity of medical care purchased than would an 80% reduction in the price which the family had to pay for with a $20 000 health insurance premium. This implies that the insurance demand curve is steeper than the Marshallian demand curve used by Pauly, and that the actual moral hazard welfare loss is smaller than would be the case if evaluated by a movement along the Marshallian demand curve.

More importantly, however, the price reduction is the mechanism used in the insurance contract to transfer income out of the insurance pool to the consumer who has become ill. For example, without insurance, a consumer who contracts breast cancer would spend $20 000 of her own money on a mastectomy. If she purchased an insurance contract for $6000 that lowered her price to 0, she would purchase the $20 000 mastectomy, plus the $20 000 breast reconstruction and two extra days in the hospital to recover for $4000, all paid for by the insurer. The additional $24 000 in spending on the breast reconstruction and the two extra days in the hospital represents the moral hazard. Although the price has fallen to 0 to the consumer, the price of the care that the hospital and physicians provide has not changed, and $44 000 must come out of the insurance pool to pay for her care. Of that amount, $6000 represents the premium that she paid originally, but the

rest, $38 000, represents the premiums that others paid into the pool and that were used to pay the providers on her behalf. These payments represent a transfer of income to her. If the insurance contract was such that this income transfer were paid directly to the consumer upon becoming ill, it would cause the consumer to purchase more medical care than if uninsured, and thus generate a portion of the moral hazard.

Indeed, by comparing the total moral hazard under a standard insurance contract to the moral hazard under a contract that paid off with a lump-sum equal to the same income transfer, one can distinguish the efficient moral hazard from the inefficient moral hazard. If the insurer had paid off by writing a check to the consumer for $44 000 upon the diagnosis of breast cancer, this additional income may have caused the consumer to purchase the $20 000 breast reconstruction, but not the two extra days in the hospital for $4000. If this were the case, then the $20 000 breast reconstruction would represent efficient moral hazard because the consumer could have used the additional income to purchase anything of her choosing. So, if she chooses to purchase the medical care, one can assume that the additional income has shifted the preinsurance demand curve outward and that the willingness to pay now exceeds the cost of producing the care. The $4000 for the extra hospital days is inefficient and consistent with Pauly's original concept.

Conventional versus Alternative Theories of Moral Hazard Welfare Compared

The alternative theory can now be compared directly with the conventional theory of the moral hazard welfare loss. In Figure 3, the Marshallian demand curve D shows the response to an exogenous change in the price for the consumer who has become ill. At a medical care price of 1, the consumer, if uninsured, would consume m_u medical care. If the price had fallen to c exogenously, m_e would be purchased, but that would not represent the response to 'purchasing of a price of c' through an insurance contract. Purchasing a price of c through an insurance contract would have generated a smaller demand response because income in the amount equal to the premium payment is no longer available to use in purchasing medical care at the lower insurance price, c. The effect is to make the insurance demand steeper and to reduce spending from m_e to m_i. And as increasingly lower insurance prices (cs) are purchased, the difference between the Marshallian demand and the insurance demand would increase, because of the increasingly greater insurance premiums charged for lower and lower coinsurance rates. At the same time, the effect of the income transfer would shift the Marshallian demand curve to the right, D^i, exhibiting this shift directly for all prices above 1, but for prices below 1, both the price and income transfer effects together would be manifested as a simultaneous movement along an increasingly steeper demand curve and a shifting of that portion of the curve to the right.

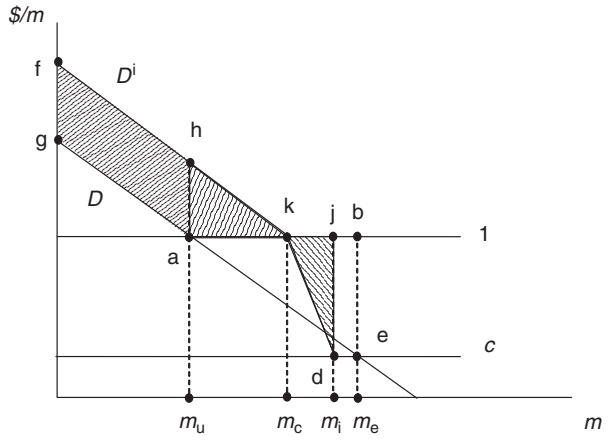


Figure 3 Net welfare gain under the alternative theory.

If a price of c were purchased with the insurance contract, the additional medical care that would be purchased because of using a price reduction to transfer income is represented in Figure 3 as $(m_i - m_c)$. The welfare loss from this purchase can be represented by triangle *kjd*. The shifting out of the demand curve caused by the income transfer to D^i would result in $(m_c - m_u)$ additional medical care purchased, relative to the amount that would have been purchased if uninsured. This additional medical care has a welfare value, that is, an increase in the consumer surplus equal to triangle *hka*. In addition, the transfer of income through insurance would increase the willingness to pay for all the care that was being purchased without insurance, resulting in an increase in the consumer surplus of area *fhag*. In contrast, under the conventional theory, there would only be a welfare loss defined by a movement along the Marshallian demand and equal to area *abe*.

Implications of the Alternative Theory

The implications of the alternative theory are far-reaching, and contrast dramatically to the implications of the conventional theory. Here are some of them.

First, not all moral hazard is welfare decreasing. Some moral hazard purchases are efficient and some are inefficient, and the challenge for policy is to distinguish one from the other in order to apply cost sharing only to the inefficient moral hazard. Thus, the theory is consistent with the concept of value-based insurance design which attempts to apply coinsurance rates only to those areas of insurance coverage that are to be discouraged, and not to others. Contrast this to the policies supported by conventional theory to apply high coinsurance rates to all types of medical care across the board, and with no limit on out-of-pocket spending, in order to reduce all moral hazard spending.

Second, health insurance is more valuable than has been deemed so under conventional theory because of the explicit recognition that insurance provides access to expensive health care that would otherwise be unaffordable and for which there would be no alternative way to access privately. That is, insurance is valuable precisely because of the additional care that it allows the ill consumer to purchase. Indeed, it has been argued that the RAND Health Insurance Experiment was biased by attrition and that the attrition accounts for the lack of a health effect from the reduction in health care use, especially hospitalizations, among the participants assigned to the cost-sharing arm. This means that, far from being welfare decreasing, insurance is welfare increasing, and

government programs designed to insure the uninsured represent beneficial public policy.

Third, an insurance policy that pays-off by paying for care represents a stand-in for a *contingent claims* insurance policy that would pay off by making a lump-sum income payment upon diagnosis. Although there may be a moral hazard welfare cost from the prevalent use of the standard policy, it is likely that the welfare cost of a contingent claims policy would be higher. For example, before a claim could be paid, the insurer would need to hire physicians or other health professional to review each claim and verify that the claimant actually had the claimed diagnosis. Moreover, to specify the various payment adjustments that would be required in the event of the various complications or adverse events that could occur with a diagnosis and its treatment, the insurer would need to hire a number of lawyers, actuaries, economists, accountants, and others to write the contracts and to keep them updated in light of scientific advances, price increases, and other changes that would necessitate adjustments in the payoff. If the moral hazard welfare costs in a standard insurance policy represent the transactions costs of transferring income to those who become ill and if the level of these costs in the standard policy is the lowest of any type of policy, then these costs can essentially be ignored as a necessary inefficiency.

Fourth, by focusing on the moral hazard welfare loss, conventional theory led economists to focus on solutions to the health care cost problem in the US that were related to reducing the quantity of medical care, rather than reducing the price of care: applying coinsurance rates and deductibles, moving to managed care and promoting consumer-driven health care insurance arrangements. These policies seemed to work. Using recent Organization for Economic Cooperation and Development statistics for the Group of 7 (G7) countries (Canada, France, Germany, Italy, Japan, the UK, and the US), it can be shown that Americans went to the doctor about half as often and spent half as many days in the hospital as citizens of the other G7 countries. Nevertheless, the US spent over twice as much per capita as the comparable average for the rest of the G7 countries. One interpretation of this is that by focusing on the moral hazard welfare loss, conventional theory misled economists to focus on the solutions that would reduce the quantity of health care consumed, when the more important source of the health care cost problem in the US was high prices that were generated by the monopoly power of providers.

See also: Access and Health Insurance. Health Insurance and Health. Health Insurance in Developed Countries, History of. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in the United States, History of. Health Insurance Systems in Developed Countries, Comparisons of. Health-Insurer Market Power: Theory and Evidence. Moral Hazard. Performance of Private Health Insurers in the Commercial Market. Risk Selection and Risk Adjustment. Value-Based Insurance Design

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**, 941–973.

Bernard, D. M., Banthin, J. S. and Encinosa, W. E. (2009). Wealth, income, and the affordability of health insurance. *Health Affairs* **28**, 887–896.

Feldstein, M. S. (1973). The welfare loss of excess health insurance. *Journal of Political Economy* **81**, 251–280.

Friedman, M. and Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy* **66**, 279–304.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–292.

Newhouse, J. P. and the Insurance Experiment Group (1993). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press.

Nyman, J. A. (2003). *The theory of demand for health insurance*. Stanford, CA: Stanford University Press.

Nyman, J. A. (2007). American health policy: Cracks in the foundation. *Journal of Health Politics, Policy and Law* **32**, 759–783.

Pauly, M. V. (1995). When does curbing health care costs really help the economy? *Health Affairs* **14**, 68–82.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**, 531–537.

Stanton, M. W. and Rutherford, M. K. (2006). *The high concentration of U.S. health care expenditures*. Rockville, MD: Agency for Healthcare Research and Quality.

Demand for Insurance That Nudges Demand

MV Pauly, University of Pennsylvania, Philadelphia, PA, USA

Introduction

The primary benefit from health insurance for risk averse people is to spread the risk of high expenses. But it can also affect the use of medical care. Although insurance coverage can then do harm to efficiency by distorting consumer/patient demand for medical services, it can also provide potential benefit by offsetting existing distortions. An important further consideration in some countries and in some settings, however, is whether such corrections will be offered by competitive insurers and accepted by buyers in voluntary insurance markets. Will market insurance coverage, that nudges people to do the right thing, be supplied and purchased?

The answer, to be discussed in this article, is the usual answer in welfare economics: 'It depends.' Depending both on the source of the distortion and the parameters of buyer preferences, corrective insurance may sometimes be strongly demanded, might or might not be expected to occur, or be unlikely to happen in unregulated insurance markets.

For example, the idea that individuals can be incentivized to change behaviors that are harmful to them is at the core of the normative appeal of behavioral economics (Della Vigna and Malmendier, 2006). Beginning with the observation that people with given incomes faced with a set of prices for different goods and services sometimes make mistakes, and sometimes do so in predictable ways, economists have developed normative analyses to show how incentives including prices can be changed to nudge, push, or drive people away from these consistently irrational acts into behaviors that will end up making them better off, at least in some way and along some dimensions (Thaler and Sunstein, 2008; Kahneman, 2011).

Both healthcare and health insurance have, not surprisingly, been prime candidates for nudging. Because information about illness and medical treatment is imperfect (for consumers, but also for providers of care and suppliers of insurance), there are many cases in which choices are made that turn out to be wrong later. More relevantly, there is also suspicion that consumers have less information than the maximum amount available, and so may make choices that do not maximize expected net benefit, either for them or for society. As a result, there is interest in changing how health insurance is designed and sold in order to improve matters (Chernew *et al.*, 2007; Fendrick and Chernew, 2009; Fendrick *et al.*, 2012). The most prominent (but not the only) example: if consumers do not have a full evidence-based understanding of the benefit from some treatment, and systematically use less or more than the amount that would maximize expected net benefit, might cost sharing for that treatment be changed in ways that help (Pauly and Blavin, 2008)?

In much of the analysis, the identity of the agent who is going to be doing this incentive changing is not specified; it is enough to show that 'we' could change things in such a way as to make 'us' better off. In some of the specific applications to such things as employee retirement benefits (Madrian and Shea, 2001), the implicit incentive-planner would plausibly seem to be the employer, though the proof that doing things that make workers better off will also make the stockholders of the firm better off is usually absent. In the largest share of this literature, it is government, broadly imagined as an entity interested in maximizing ex post economic welfare, that appears to be the intended customer for the normative advice (Bernheim and Rangel, 2009). What is probably least common is a serious investigation of the question whether or when voluntary markets in behavioral change might emerge and function efficiently.

This article investigates under what circumstances consumers might choose to change the health insurance incentives they face in order to bring about behavior which is likely to make them better off. (What exactly 'better off' means will be important.) Although attention will be paid to the risk reduction benefits of insurance, it is also worth noting that the main tradeoff in insurance – pay a higher fixed amount initially (the premium) in order to reduce the price at point of use later (the coinsurance) – is the same structure that has been studied for health clubs, great book clubs, and other examples of devices to bring about behavioral change.

In this model consumers can be perceptive about their failings; they are assumed to be able to understand that sometimes, for various reasons, they may not choose behavior which is ex post optimal for them, or that something needs to be done, either to information stock or user prices, to improve choices. Therefore, it is necessary to investigate whether it is possible that insurance that covers its costs and corrects such failings that will be demanded. It is assumed that supply is competitive so insurers will supply the kind of changes consumers might demand.

A main finding is the likelihood that consumers will voluntarily agree to be nudged depends critically on the reason why their behavior was nonoptimal in the first place, and even then, on the values of key variables in the problem. Sometimes there will be demand for nudging, and sometimes not.

Outlined here is a simple model of economically efficient cost sharing when consumers might underestimate the marginal benefit of some kinds of medical care; also indicated here is the voluntary insurance and medical care choices these consumers would make in such a situation, compared to what they would choose if they correctly estimated benefit. Then it is asked whether and when consumers would be willing to choose something different from this choice of both insurance and medical care, is there something else that they would prefer and which would make them better off? One fairly tautological model is provided where the outcome of voluntary efficiency-increasing nudging does occur with competitive insurers. That model is compared as a benchmark with other stories that raise serious issues of whether people will voluntarily demand the incentive-changing mechanisms that will make them better off, and whether insurers will supply them if

they are demanded. It is shown that under not implausible assumptions there are some cases in which voluntary demand will not materialize in the ideal way, and it is explained why. At the end the question addressed is whether institutional arrangements alternative to voluntary insurance markets, like public sector interventions, can do better, and show that government in a democratic setting might be subject to similar problems.

The Core Model

The model is one of competitive insurers choosing to offer policies with possibly different levels of cost sharing for different services. Two kinds of services, 'preventive' and 'treatment,' come to mind. The distinction is that a 'preventive' service affects the probability of future health or illness states, whereas a treatment service provides only short-term (if valuable) benefit when an illness strikes. Thus a preventive service both provides benefit in the form of improved future health and potentially lower demand for treatment if illness is avoided; the concept includes both what is usually labeled prevention but also the great majority of other health services in the first stage or early onset of some illness that affects what happens to health later.

In the absence of insurance, demand for either kind of service bought in a competitive market by fully informed consumers would be (presumably) first best optimal, at the point where marginal benefit equals marginal cost. For treatments of this condition, this just equates the (money) value of current period health benefits to price, assumed to be equal to marginal cost. For preventive services, both current period cost and future 'cost offsets' are part of the full marginal cost, whereas the value of expected future health (if care were costless) is the measure of marginal benefit. Alternatively, the value of marginal future health could be combined with cost offsets as a measure of benefit, to be equated to the current period price or cost of preventive care.

A simple version of the first order condition for optimal preventive care use would be:

$$P_S = \Delta\Pi\left(\frac{\Delta U_H}{\lambda} + C\right) \qquad [1]$$

where $\Delta\Pi$ is the change in probability of future illness due to consumption of one more unit of the preventive service, ΔU_{H} is the marginal utility of future health (comparing health in the illness state vs. health in the healthy state), λ is the marginal utility of money, and C is the cost of treating the future illness.

If there is no insurance coverage, consumption of both services will be at the optimal level. In particular, the consumer in deciding on consumption of the preventive service will take future reductions (cost offsets) for the cost of treatment into account, along with the value of health benefits. That is, the consumer sees and satisfies condition, eqn [1]. However, if there already exists coverage of the treatment, and there is a positive cost offset, there should be insurance coverage of the preventive service that reflects the part of any cost offset for treatment that is covered by insurance. This is a second-best argument. In the absence of such an adjustment, the consumer ignores the cost offset term, and underconsumes the preventive service.

In the limiting case in which the expected cost offset ($\Delta\Pi C$) exceeds the price of the preventive service, and the other service is fully covered, insurance coverage of the preventive service should be 100% in the absence of insurance administrative and claims processing costs, regardless of the degree of price responsiveness (Glaser and McGuire, 2012). If there is a positive marginal administrative cost to insurance coverage, that consideration would reduce the ideal extent of coverage. If there would be positive use of preventive care in the absence of coverage, then coverage should be higher the greater the price responsiveness of the use of coverage to cost sharing (Held and Pauly, 1990). If price responsiveness is low, coverage per se may increase the aggregate expected insured expenses, and the higher premium is offset by these higher benefits. However, if there are administrative costs, those additional costs, when applied to paying benefits where use would have occurred in any case, are wasteful.

The Setting and the Behavioral Model

This simple case is well known. But the more interesting questions arise in one of the most frequently discussed (and topical) applications of behavioral economics: the idea that cost sharing in health insurance might be used to guide people to choose more efficient levels of consumption of effective medical care than they might otherwise select, which is commonly called 'value-based' cost sharing in the health insurance literature (Chernew *et al.*, 2007). The alternative model in the discussion is usually one in which cost sharing is uniform across all settings associated with a given level of spending (e.g. 20% coinsurance or a $1000 deductible for all covered medical expenses); it is alleged that value-based cost sharing will produce a better outcome than this.

But this status quo is not the best alternative system. The theory of optimal insurance (Pauly, 1968; Zeckhauser, 1970) envisions varying coinsurance as well, but for different reasons and in different ways than prescribed by value-based cost sharing. Therefore, the question arises whether value-based cost sharing that dominates some or all of these alternatives in terms of ex post net benefits would be preferred by consumers.

The benchmark framework in mind is this: competitive insurers in unsubsidized and unregulated markets are free to set cost sharing levels (as proportional coinsurance) at different levels for different services. Consumers choose among insurance plans based on their premiums and their cost sharing. Each insurance plan's premium must cover the costs of the benefits it pays out plus administrative expenses, and may yield positive economic profits if the market is not competitive. The first question is whether a plan that selects the level of cost sharing prescribed by the value-based approach will be preferred by consumers to plans offering other levels of benefits and associated premia. (The second question, whether insurers will offer that plan, will be considered later in the article.)

It has been shown (Pauly and Blavin, 2008) that, in the absence of cost offsets, a necessary condition for value-based

cost sharing to improve outcomes in competitive insurance markets is that the patient's marginal benefit or demand curve differs from the curve that represents true marginal benefits. If patients always consider correctly the value of effective medical care, they will use highly (marginally) effective care even if cost sharing is high, but will use only less marginally effective and inefficient care if cost sharing is low. To control this moral hazard, coinsurance will be chosen to make the second-best optimal tradeoff between such overuse and risk protection. Moreover, under full information but with variation across types of care in patient response to cost sharing, uniform cost sharing will not be optimal; rather, other things equal, optimal cost sharing will vary directly with patient demand responsiveness. No further consideration of 'value' is needed to specify the ideal level of cost sharing.

Figure 1 provides an illustration of this model. D_I represents the true marginal (expected) health benefit curve for some kind of care; *MC* is its cost and price net of cost offsets. (This service is both uncertain and has a positive marginal net cost; cost offsets from its use are not sufficiently large that consuming more of it reduces total benefits cost.) Because of uncertainty, it is assumed that consumers get benefit from insurance coverage of this service. There is a (second best) optimal level of coinsurance, indicated in the diagram as c_I^*, which consumers will also prefer to any other level of coinsurance. At this point the marginal welfare cost from lowering coinsurance will equal the marginal benefit from further risk reduction. At that point, the quantity will be second best optimal. Optimal coinsurance (other things, including risk characteristics and risk aversion, held constant) will be lower for less price responsive types of care and higher for more price responsive types of care. In all cases, the marginal benefit will be less than the marginal cost. At this optimal pattern of coinsurance, the value or marginal benefit from each type of care will equal the level of coinsurance per unit. At a given level of coinsurance, when informed consumers are in equilibrium, no one type of care will have higher marginal value than any other, so there is no need to further vary coinsurance with value. However, at the optimal level of coinsurance the marginal value of less price responsive care will be lower than that for more price responsive care because the lower coinsurance that leads to a lower value provides an offsetting benefit in terms of better risk protection.

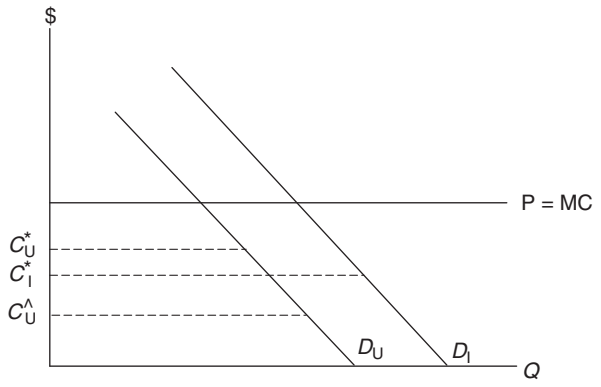


Figure 1 Optimal coinsurance and the demand for care.

Deviations from Optimality

Now suppose that the marginal benefit curves that patients are using are lower than the true curve. What then? Start with a simple comparison. Suppose that three plans are offered. One (informed plan) sets the coinsurance rate (as described above) at the optimal level given the consumer's risk aversion and given the marginal benefit curve that would be generated if patient demands were based on accurate estimates of the marginal benefit from different amounts of care. However, patients are assumed to underestimate the benefit from some important service, and so would consume less than the full optimal information amount if they were in the first plan. The inaccurate expected marginal benefit curve is indicated in Figure 1 as D_U. So an alternative value-based plan (Nudge Plan) is offered with lower cost sharing at $c_U^\wedge$. The purpose of the lower cost sharing is to offset the effect of benefit underestimation by increasing quantity demanded by using a lower user price. This is the optimal level of cost sharing, given that the marginal benefit curve is underestimated.

There is, however, a third alternative insurance plan, one that the consumer might prefer: Specify c_U^*, the coinsurance rate and premium that would be optimal given the actual (though underestimated) demand, and the lower rate of use and lower premium associated with that plan. The uninformed plan generally has a higher coinsurance rate (at c_U^*) than either $c_U^\wedge$ or c_I^*, with both a lower premium and lower expected medical costs than the Value-Based Nudge Plan. The reason why the coinsurance rate is generally higher than with the true marginal benefit curve is that, with lower demand at any level of coinsurance, there is less risk.

Figure 1 depicts each of the three plans under alternative assumptions that the marginal benefit curve is the informationally correct demand curve D_I or the actual (uninformed) demand curve D_U. Note that the welfare cost of moral hazard is smaller at all coinsurance levels along the D_U curve than it would be under the informed plan with the informed demand curve.

The Gain from 'Nudging'

This simple example shows that there can be gains from getting the consumer to choose the Nudge Plan. How does the size of the gain vary with the position of the uninformed demand curve? The answer depends in part on whether the informed plan or the uninformed plan is used as a benchmark. The case is simplest if welfare under the Nudge Plan is compared with what it would have been under the fully informed plan. Pauly and Blavin (2008) show that, over some of the range of possibly underestimated demand curves, welfare may actually be higher with underestimation and the Nudge Plan than with the informed plan. This is what they call the 'benefit of blissful ignorance.'

There is obviously a gain from permitting the marginal benefit curve to fall short of the true curve as long as it remains above that curve which hits the x-axis at the optimal level of use (ignoring income effects). That curve is D_U^* in Figure 2; should it prevail, coinsurance can be set at zero and yet use will be first best optimal (ignoring income effects). The

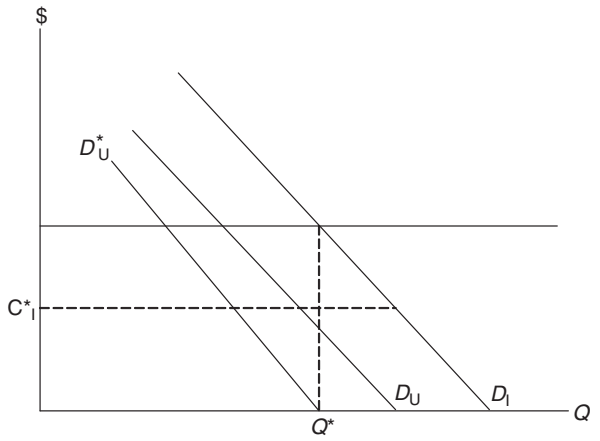


Figure 2 Lower bound on demand for care.

consumer can completely avoid the consequences of moral hazard, and have both full protection against risk and optimal use of medical care. To the left of D^*_U welfare begins to fall, but remains above that with the informed plan at c^*_I over some range.

Modeling Deviation

With this as background, a model can be made of the causes of deviations in the patient's marginal benefit curve from the true value and corrective strategies. Begin by thinking of what kind of medical service would be one for which consumers would demand insurance but underestimate true marginal value. Think of a service for which demand is stochastic today and which affects health tomorrow. Although some acute-care services yield immediate utility benefits (analgesics, suture of a bleeding wound), the bulk of medical services are of this 'two-period' character. Statins for people who have already had a heart attack, asthma medications, and cancer surgery are all things that a person might or might not need, depending on the onset of the chronic condition, but which then all generate disutility in one period in return for a benefit in the future. (There are some complexities associated with insurance coverage over multiple periods which will be ignored for the present.) The person decides on insurance coverage for such services at time t_0. It is assumed that there is such a service with a (gross) market price in period t of P_t and nonmarket costs (time, pain, bother) of C, all incurred at time t_1. If the person consumes the service, health is increased in period $t+1$ and future periods.

The first order condition for optimal choice (slightly re-written) is:

$$P_t + C_t + \sum_t \frac{\Delta C_{t+j}}{(1+r)^{t+j}} = \sum_t \frac{MH_{t+j}}{(1+r)^{t+j}} \qquad [2]$$

Here ΔC_{t+j} is the cost offset in the $J(j=1,2,\ldots,J)$ future periods the person will pay, MH_t is the value of the additional health at time t from the service measured in dollars, and r is the interest rate; the expression on the right gives the present discounted dollar value of additional health. It can be assumed that MH falls as the service is consumed at a higher rate. That is, the consumer compares the price with the discounted marginal benefits minus any nonmonetary cost from treatment.

There is only one way a consumer can estimate marginal benefit correctly, but there are (apparently) many incorrect ways to do it. Consider patient nonadherence to a physician's prescription to use some product or service, conditional on a diagnosis of some chronic condition. The things that can be, apparently, estimated or considered incorrectly are all in eqn [2]: the cost offset (because of insurance coverage), the value of the marginal health product or the service, the interest rate, and the nonmonetary cost of the service.

Insurance coverage distorting the consumer's value of the cost offset is one reason why patients may not use the care they should. Imperfect information is another likely reason offered, especially if patients have difficulty understanding the physician's explanation for some prescribed treatment. In addition, if people use nonexponential discounting, they may fail to consume services of high marginal (future) benefit, even if they correctly perceive that benefit, because they underestimate the value of that benefit. Quasi-hyperbolic discounting would be one way to model this imperfect discounting. Finally, the nonmonetary cost itself may be high, higher than is perceived by the physician who recommends the service and is then disappointed when patients are nonadherent to the recommendation. In this case it is the information for the potential 'nudger' that is incorrect, but it is a possible scenario for trying to persuade consumers to change.

The case for voluntary value-based cost sharing as a function of these four rationales for value-based cost sharing will now be explored.

The Best Case for Voluntary Value-Based Cost Sharing: Positive Insured Cost Offsets

The most obvious case why some level of coinsurance might be too high is if there are positive cost offsets (the current preventive service reduces future costs along with improved health) and, although those future services are covered by insurance, that is not taken into account in specifying the coinsurance for the preventive service. Then in calculating marginal benefit the consumer fails to take these reductions in cost (and in insurance payments to cover those costs) into account.

Most of the examples of the success of value-based cost sharing deal with this case. However, the conventional theory of optimal voluntary coinsurance in competitive markets is already supposed to have taken these effects into account because the cost of the preventive service is the net (of cost offsets, positive or negative). The idea is that the insurer will recognize these effects, and build them into the premium adjustment that matches any change in the level of coinsurance for the service. This happy state of affairs can be impeded if there is turnover among insureds – if the person potentially leaves the plan before the cost offsets occur. Other than this, however, the market solution to this case is well known; it is one in which the consumer does not require extra nudging beyond what would have been built into optimal insurance in the first place. The consumer notices that the plan

with 'nudging' coinsurance carries lower premiums and better benefits than any other, and chooses that plan.

A Good Case for Voluntary Value-Based Cost Sharing: The Consumer Looking to the Future Seeks to Control His Irrational Current Self

The strongest behaviorally motivated case for the value-based plan is to note that it is the plan that maximizes ex post utility, given the underestimated demand curve and the true marginal benefit curve. It is the former curve which describes behavior, but the latter which describes the actual outcome and its value. A split-personality or self-control model is a very common approach in the literature to this case, usually applied in cases where discounting is hyperbolic or inconsistent in some way. The approach of Della Vigna and Malmendier (2006) and Della Vigna (2009) that they used for health club annual memberships (as opposed to paying per use of the gym), modified to fit the health insurance case, will be followed.

The idea is the consumers realize that, although they should exercise, get their test, or take their medicine, because of lack of self-control they will not do so when they are facing the full price per unit, or even at the full information ideal level of cost sharing. They therefore prefer incentives that will be set at a level such that, given their expected attenuated future demand behavior, they do what they should. They therefore sheepishly prefer a plan with low enough cost sharing to get them to do what they ought to with any alternative, because it dominates all other alternatives in terms of ex post net benefit. In this case, the most common interpretation is that it is not that the consumer misperceives, it is that he or she misbehaves.

An alternative interpretation, based on the psychological work of Zauberman *et al.* (2009) is the consumer misperceives time. Exponential discounting of misperceived time can be equivalent to hyperbolic discounting of correctly perceived time. Either way, this case can be described as resulting from using a discount function that differs from the conventional exponential one – for example, by being hyperbolic.

Formally, imagine multiplying the discount rate $(1/(1+r))$ by a term B that reflects the underestimation of the value of future benefits (at time $t+1$ and later). This value is the one the consumers attach to future benefits at the time t when they might consume the costly and bothersome service. The usual model at this point imagines that the consumer considering precommitment at time t_0 with regard to behavior at time t_1 seeks to reproduce the behavior at time t_1 which would have occurred under exponential (nonmyopic) discounting (Rasmusen, 2008).

But in this case the paradoxical results on optimal underestimation means that the goal is not to get the consumer to fully correct the imperfect discounting. Doing that would lead to the D_{I} demand curve, although utility is higher if the demand curve is only at D^*_U. Put slightly differently, in deciding how to control one's 'bad' self in the two-selves model, one does not want to get that second self to do what the first self would have done under a fully correct perception of the discount rate. Instead, one wants to adjust cost sharing to produce a rate of use potentially greater than current (myopic) self-use and coverage but not as much as would be used by the nonmyopic self. In effect, the first self takes advantage of the impatience of the second self as a way of controlling moral hazard on services with preventive benefits. Far from wanting to correct a later shortsighted behavior, one would want it to remain shortsighted, just not as much.

There is an additional issue here of some potential importance. It may well be that the decision to precommit in period p_0 changes the person's demand curve in period p_1 when period p_1 arrives. That is, recognizing the arrangements for precommitment, the person may be less resistant to proper discounting; there may be feedback from the decision to precommit to behavior in some event to the behavior that would occur in that event. This might be especially likely to happen if Zauberman's model of discounting holds: observing the precommitment device changes how one thinks about time closer to its true value. If this happens – if the discounted marginal benefit curve in period p_1 is moved closer to the true curve (as perceived in period p_0, or period p_2 or later), then the structure of the precommitment device – the lower user price – will need to be changed to one that has a higher user price. But if the demand curve is shifted up enough, the expected utility under precommitment may actually end up lower than at the initial no-precommitment point. In this case no voluntary nudging will occur.

More generally, what is the ideal level of cost sharing in such self-control cases? It depends on the position of the uninformed marginal benefit curve relative to the true curve. At one extreme, if the reduction in demand is so great that the quantity demanded at a zero user price (full coverage) is less than or equal to the quantity at which true marginal benefit equals marginal cost, then optimal cost sharing is zero regardless of risk or risk aversion. (Negative prices are ruled out in favor of the corner solution.) If the quantity at a user price of zero is greater than the quantity at which marginal benefit equals marginal cost, then the optimal extent of coinsurance is calculated, as in Pauly and Blavin, (2008), by comparing the true marginal welfare cost with the marginal risk premium given the level of use and loss distribution that prevails under the underestimated demand curve, taking into account the person's risk aversion or 'risk premium.' It will be optimal to have some positive coinsurance for the person who lacks self-control as long as the distortion is not too large. (This question will be regarded when the alternative form of nudging is considered involving providing information on what the marginal benefit actually is.)

The consumers will voluntarily choose the cost sharing option that precommits them to lower payment per unit in return for payment of a lump sum payment (premium or membership). This provides two kinds of gain. To the extent that the use of the service is stochastic (Della Vigna (2009) describes even health club visits as stochastic), there is a risk premium gain to a risk averse person from paying some of the expected cost in advance. And then there is the gain in expected utility from precommitment.

How binding is the precommitment? Once the people are facing the possibility of paying for and using the preventive service, they attach less value to using it and, in view of low likely use, would prefer an insurance with higher cost sharing and lower premiums. There is no absolute barrier to changing

insurance coverage at any point in time; usually contracts for employment related coverage are for a year but individual insurance can be changed at any time. However, it is likely that no single specific coverage would motivate a change – our consumer in the throes of myopia would just prefer insurance with less coverage of preventive care in general. One could model the decision to renege as based on a comparison of the transactions costs of changing behavior versus the difference in expected utility between the precommitted coverage and the myopically optimal coverage. Of course, if the medical event occurred close to the time when the person is deciding to make an annual election, things could be different.

Better Information as the Solution

Another reason for benefit underestimation is imperfect information about benefit. One strategy is to provide information. But when would it be socially efficient to provide information? If marginal benefit is underestimated by a sufficiently large amount, more information may help; if not, more information may do harm, even if it is costless, as discussed earlier.

Imagine varying the D_{U} curve by changing information, and plot various levels of coinsurance.

As illustrated in Figure 3, if the nudger can control both information and the coinsurance rate, it would provide no corrective information but set coinsurance along the *BC* line. In contrast, if it can only control information but consumers choose coinsurance along *AC*, it would choose information to move the marginal benefit curve to D', where use at the consumer-chosen coinsurance rate C'_U is just at the optimal level.

But in the context of voluntary choice of insurance, it may be hard to be explicit about keeping people ignorant. Thus both characterization of settings in which leaving underestimation untouched is a good strategy and asking whether people will voluntarily and explicitly choose to leave things that way, is required.

Here is the source of a serious dilemma. The simplest explanation is to assume that the consumer focuses only on the final outcome (in terms of comparison of health improvement, premiums, and expected out of pocket costs), recognizes that this outcome is superior to that with the uninformed plan, and therefore prefers the nudge plan with a health outcome which (given its cost) is better ex post. The dilemma arises if the consumers reflect on the source of this improvement. Suppose they realize that it comes from the fact that the initial perceived marginal benefit curve was incorrect, and was less than the true curve. But if the consumers become aware of this fact and respond by increasing their demand curve, use and outcomes under the Nudge plan's coinsurance rate will not replicate what is anticipated under the informed plan. Instead, faced with a relatively low coinsurance rate, use will be higher. Health outcomes will be even better than under the informed plan, but (in the short run) the insurer will lose money, and, after seeing an increased premium to cover this higher use, the consumer will judge the higher premium and higher expected out of pocket cost to overshadow the health improvement. If the demand curve shifts all the way to the true demand curve, the ideal coinsurance rate will also rise to the level that is optimal.

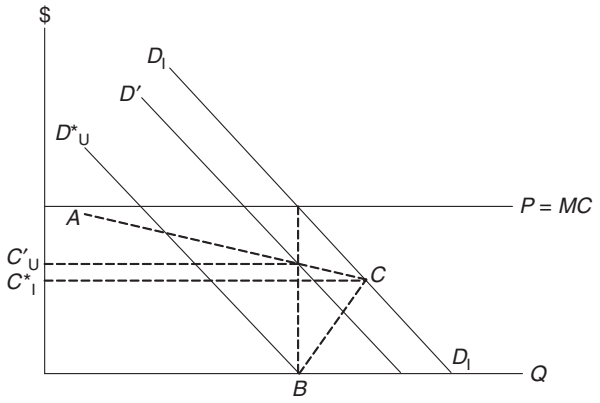


Figure 3 Tradeoff between demander information and coinsurance.

For information to work properly to achieve a better outcome, the consumer must purchase insurance and care based on the uninformed demand curve. But to be motivated to buy the plan, the consumers may need to know (and be convinced) that the health outcome they will achieve at the level of use they target under the plan's coinsurance rate will be much higher than they would expect. This is the heart of the dilemma: they will prefer the nudge plan only if they think their health outcome will be better than what they think it will be as embodied in their (uninformed) demand curve. Convincing them of this will arguably shift the actual demand curve.

Putting the pieces together, it is noted that the strategy of improving information only improves ex post welfare if the marginal benefit curve is below D^*_U. Even then, the uninformed demander will not be willing to pay the higher premium for better coverage as long as they remain ignorant. Providing information can shift the demand curve for care and coverage, so if information is relatively cheap and effective it may pay for a firm to charge a little more after paying to shift demand. Even here, however, the level of coverage will be on the *AC*-locus, not on the *BC*-locus. Merely offering the optimal level of coverage on any underestimated demand curve will not be persuasive. To get the consumers to prefer coverage on the *BC*-locus one would have to fully inform them, but then demand would shift to a level with higher coinsurance and higher moral hazard. Less information means less moral hazard but less correction in coverage. It does not seem possible to reach the first best outcome in a voluntary way.

The Cost of Bother

Both imperfect self-control and imperfect information are reasons why demand does not reflect marginal health benefit. But the informal literature on adherence also suggests another reason: The consumer forgets, or it is too much bother – there are subjective costs. In effect, there are additional costs on top of any cost sharing. It is these subjective costs that shift the marginal benefit curve downward.

But note that what prompts them to take their medicine is the lower user price, implying that decisions are made

rationally by comparing perceived benefits to 'short run' costs including time and bother. In the ignorant case it is the perceived benefit estimate that is wrong; in the 'bother' case there are some additional (uninsured) costs. If these costs are expected to be real (that is, if people know from past experience that they must work hard to remember, and do not just overestimate the actual effort in remembering), then people will expect to incur those costs if the user price is lowered enough to result in the desired behavior. Here again, but for a different reason, ex post welfare will be lower under the Value-Based Nudge Plan, and people may refuse to be subject to the push.

Perception of Nonmonetary Cost

Another reason why people may fail to follow provider advice about services which affect future health is because they experience or expect to experience nonmonetary costs associated with those services. Those costs may represent physical side effects (nausea, impotence, dizziness) or they may represent the effort needed to remember to take the medicines or treatment on the prescribed basis or even the bother of filling the prescription.

In the first best world these costs will be taken into account in determining the net marginal benefits, but in the setting in which physicians write prescriptions they may have a difficult time knowing what these costs are. If they prescribe based on, say, the average patient's net benefit from a drug, those who have above-average nonmonetary costs may rationally choose not to comply; there may be rational nonadherence. Lower the user price, and there will be more adherence as those who rationally failed to adhere when they were exposed to the full monetary cost of the treatment and their nonmonetary costs now find positive net benefit when the nonmonetary cost falls. But when confronted with a higher premium to pay for this change in future behavior, this group will correctly judge its net benefit to be negative.

Other Threats: Heterogeneity

What will the market look like if some consumers underestimate marginal benefits but others do not? If some consumers understand marginal benefits correctly but insurers cannot tell who is who, the well informed patients will find it advantageous to themselves to be pooled with other underestimating consumers who use less care at given coinsurance levels. The breakeven premiums will then reflect an average of the use of both classes of consumers, just as in conventional adverse selection models, which will make low coinsurance value-based policies even less attractive.

The Mixed Case

Now consider the most complex but probably the most realistic case: one in which patients underestimate the true marginal benefit curve (imperfect information) but also overdiscount those benefits (imperfect discounting). If the underestimation can be kept secret, the consumer might be willing to agree to a plan that lowers the user price enough to offset the imperfect discounting.

The key issue here as before is how far the demand curve is shifted to the left by these two influences. If it is still to the right of D_U^*, the two-self model will tolerate some reduction in cost sharing without having to turn to information to move the curve. The change will be the utility maximizing coverage based on a correctly discounted but underestimated marginal benefit curve.

Focus Group Evidence

A series of focus groups using subjects from the state of Michigan were asked about various aspects of alteration of insurance coverage to vary cost sharing with measures of clinical (net) value using a set of scenarios (Swinburn *et al.*, 2012). There was no scenario based on quantitative values for changes in use of care, health outcome, and total medical spending or premiums but there are some qualitative results of interest for the models that have been discussed.

One scenario proposed lowering copayments to zero for a medical condition (diabetes) thought to be characterized by underuse of recommended care and poor health outcomes. The scenario envisioned cost offsets in employment based insurance (patients with lower cost barriers are more likely to stick with the care they need, which would make them 'more likely to use fewer healthcare dollars'), but the total premium will initially increase though the employer 'expects to make it up with healthier employees' and will eventually have less costly health insurance.

Participants generally supported the intervention as described but with several caveats that are important for our analysis. First, they would support lowering copayments for diabetics 'only if the program saved money.' Although this response is somewhat vague, a reasonable interpretation is that they would not favor the program if premium costs they had to pay were increased even if health was improved for those workers with diabetes. There was also an equity consideration that offset efficiency gains: discounts should be available (a majority thought) only to those low income people who could not afford the prior cost sharing – even if lower cost sharing might change behavior of wealthier participants in a health-improving way.

The other interesting finding was also couched in terms of equity. It was felt to be unfair to provide lower cost sharing benefits to people with conditions under which they failed to follow physician advice. This was both rewarding irresponsible behavior and failing to reward people with conditions where adherence was high, or who had no chronic conditions. These observations could also be interpreted as referring to risk selection, benefitting high risk patients who behave incorrectly at the expense of those who manage their care properly or are low risk to begin with.

Overall, respondents felt that this new design should be used some of the time, but only in certain circumstances.

Another report (Midwest Business Group on Health and Buck Consultants, 2012) obtained similar results from another set of focus group participants. There was skepticism about the

ability of insurers to identify these cases, and a feeling that those who were compliant needed the help of lower cost sharing more than those who were less compliant.

Solutions

Solutions to these problems depend in large part on the cause. In the case of people with understanding of self-control problems, there should be a demand for nudging even without any intervention. Here providing accurate and persuasive information on the actual benefit ex post will be helpful not only to get the demand as right as it can be but also to motivate the demand for insurance.

For people who underestimate marginal benefit because of information imperfections, the strategy of providing information on actual benefit may backfire, as noted, because it will be associated with a greater rate of use in inefficient situations. There is a partial solution that may work in some cases. Suppose that the perceived marginal benefit curve is to the left of D'. Then it is possible that the utility under full information with cost sharing at the optimal level is higher than the utility with no action. In such a case providing information about the true curve and then doing the best that can be done may be preferred to the original state.

However, in this case the optimal and demander-chosen levels of coinsurance coincide; there is no need for value-based adjustments. In the more general case, it seems difficult to get the person to prefer the insurance with value-based cost sharing *a priori*.

See also: Moral Hazard. Value-Based Insurance Design

References

Bernheim, B. D. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics* **124**, 51–104.

Chernew, M. E., Rosen, A. B. and Fendrick, A. M. (2007). Value-based insurance design. *Health Affairs* **26**, w195–w203.

Della Vigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature* **47**, 315–372.

Della Vigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review* **96**, 694–719.

Fendrick, A. M. and Chernew, M. E. (2009). Value based insurance design: Maintaining a focus on health in an era of cost containment. *American Journal of Managed Care* **15**, 338–343.

Fendrick, A. M., Martin, J. J. and Weiss, A. E. (2012). Value-based insurance design: More health at any price. *Health Services Research* **47**, 404–413.

Glaser, J. and McGuire, T. G. (2012). A welfare measure of 'offset effects' in health insurance. *Journal of Public Economics* **96**, 520–523.

Held, P. J. and Pauly, M. V. (1990). Benign moral hazard and the cost-effectiveness analysis of insurance coverage. *Journal of Health Economics* **9**, 447–461.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus, and Giroux.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics* **116**, 1149–1187.

Midwest Business Group on Health and Buck Consultants (2012). Communicating value-based benefits: Employee research project results. Center for Value-Based Insurance Design, University of Michigan. Available at: http://www.sph.umich.edu/vbidcenter/publications/pdfs/CommunicatingVBBenefits-Apr12.pdf (accessed 29.06.12).

Pauly, M. V. (1968). The economics of moral hazard. *American Economic Review* **58**, 531–537.

Pauly, M. V. and Blavin, F. E. (2008). Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics* **27**, 1407–1417.

Rasmusen, E. B. (2008). Some common confusions about hyperbolic discounting. *Working Paper No. 2008–11.* Bloomington, IN: Kelley School of Business, Department of Business Economics and Public Policy, Indiana University.

Swinburn, T., Ginsburg, M., Benzik, M. E. and Clark, R. (2012). *Probing the public's view on V-BID.* Ann Arbor, MI: Center for Value-Based Insurance Design, University of Michigan. Available at: http://chcd.org/docs/vbid_report_5.12.pdf (accessed 29.06.12).

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Zauberman, G., Kim, B. K., Malkoc, S. A. and Bettman, J. R. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. *Journal of Marketing Research* **46**, 543–556.

Zeckhauser, R. J. (1970). Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**, 10–26.

Dentistry, Economics of

TN Wanchek and TJ Rephann, Charlottesville, VA, USA

Introduction

Dentistry is the field of medicine that is concerned about diseases of the teeth and other tissues and bone structures in the oral cavity. It is different to a degree from other medical services in its product attributes, market characteristics, and the level of government involvement. Although dental disease is not completely predictable, it is less random than other diseases and disorders, some of which can have potentially immediate catastrophic effects. In addition, preventative services make up a much larger portion of the care provided than for other health care services. When treatment of dental caries (tooth decay) is needed, patients are often presented with the option of restoration or removal. This type of choice may not exist for many other medical problems. These characteristics of dentistry allow consumers more flexibility in when and what they purchase. In addition, consumers are generally better informed about dental services. They can often observe and identify their disorder and have more contacts and familiarity with the limited number of dental diseases, diagnostic tools, and procedures than they would with conditions in other fields of medicine. Although poor dental health can affect one's general health in many ways over time and impede workforce performance and childhood development, dental diseases are not communicable unlike some diseases. Dentist services are generally provided by small, independently owned providers in a situation approximating pure or monopolistic competition in contrast to physicians who are more likely to be organized in larger consortia of providers such as hospitals and health groups with significant local market control. These features suggest that dental care may function more like a standard product market than other health care services where market failures are more pronounced.

Oral health has improved markedly in high-income countries over the last several decades. In contrast, many low-income countries have experienced deterioration in oral health in recent years. The improvements in high-income countries can be attributed to a variety of factors, including increasing incomes, expanded access to dental insurance, improved technology, dietary changes, and fluoridation. Many developing countries are seeing an increase in the prevalence of dental caries, largely due to an increase in consumption of sugars and inadequate exposure to fluorides.

In the US, dental service prices have increased at a much faster clip than other goods and services and even slightly faster than other medical services. However, expenditures on dental care are still a relatively small share (less than 5%) of total health care costs in the US. Unlike the large expenditures going toward medical care, the public sector in the US funds only approximately 6% of the costs. Most of the funds are for low-income children's programs (i.e., Medicaid and CHIP) and military and veterans care. Private insurance accounts for about half of spending with out-of-pocket funding the residual.

There are also notable differences in the physician and dental workforces. Dentists make up a relatively large percentage of the total health care provider labor force, with an estimated 181 725 active dentists in the US in 2010 compared to 784 199 active physicians who work on all other parts of the body combined. However, although approximately 80% of physicians are specialists and 20% are general practitioners, the reverse is true for dentists. Moreover, dentists are increasingly more likely to rely on auxiliaries to assist with dental procedures and to provide preventative care. New laws and legislation have been introduced to expand the range of services provided by auxiliaries even further. In contrast, other countries, such as New Zealand and the UK, rely more extensively on the use of mid-level oral health care providers.

Although problems in dentistry have featured less prominently in discussions about health care reform, the field is presented with its own set of challenges. Industrialized countries such as Japan and many members of the European Union offer public dental insurance. In the US and elsewhere, a relatively large percentage of the population is uninsured, resulting in serious inequities in access to care. In the US, disadvantaged individuals, minorities, and rural residents are much less likely to exhibit good oral health. In addition, dental labor markets may not work as efficiently as they could if they were less impeded by licensure/regulatory requirements that do little to enhance patient welfare.

The following sections examine these areas in further detail in order to provide a more comprehensive picture of the economics of dentistry in the present day. The first section looks at the chief focus of dentistry: improving oral health. It reviews determinants of oral health, the economic and general health consequences of poor oral health and trends in oral health outcomes. The second section examines dental demand and its determinants, including availability of insurance, time and out-of-pocket costs, public programs, oral health conditions, and other factors. The third section reviews important issues with respect to dental service supply including the supply and distribution of dentists. The fourth section considers the expanding role of other dental care providers in the US and elsewhere.

Oral Health

Determinants of Oral Health

There are many factors that ultimately determine an individual's oral health, including oral hygiene habits and behaviors, dietary choices, tobacco use, genetics, use of dental services, income, tastes and preferences, and age. One way to conceptualize individual choice about oral health outcomes is to use Grossman's well-known model of health capital in which individuals choose between spending time producing health and purchasing medical services. Health depreciates

with age, whereas education increases one's efficiency at producing health at all ages. In the dental context, people demand oral health. Oral health can be produced with various productive inputs. Individual behavior, such as brushing and flossing regularly and consuming less sugar, constitutes one type of input. Other inputs can be purchased such as checkup exams, cleanings, filling and caps for carious teeth, or extraction and replacement with bridges and dentures. People can also consume other goods, such as fluoridated water or fluoride supplements that improve oral health. How much of each input a person purchases in the market depends on a variety of factors including the price or opportunity cost of the services, the present quality of teeth, tastes and preferences for good teeth, age, etc. Applying Grossman's model to oral health, people demand less dental care as the cost of care and the time needed to produce oral health increases, suggesting people do make trade-offs between good teeth, consumption of other goods, and time.

People vary in their tastes and preferences for good oral health outcomes. Studies have found lower perceived need for care in rural areas and among individuals with a low socioeconomic status, which may be due to the social environment and expectations for good teeth. Family environment, particularly among children, is an important factor in health outcomes. In the US, whether a parent visited a dentist is strongly correlated with whether the child also had a dental visit. Similar results are observed in China. A survey of adolescents (11, 13, and 15 years old) from eight Chinese provincial capitals found that there is a strong relationship between oral health behaviors and the socioeconomic status of parents, school performance, and peer relationships. Looking at Medicaid programs, even when states increased children's Medicaid provider compensation to levels comparable to private insurance, utilization rates do not rise to the level of those with private insurance. The lower utilization rates suggest that there are significant nonfinancial barriers among low-income populations in seeking dental care, which could be interpreted as a lower preference for good oral health outcomes, increased costs of gaining access to dental services or a shortage of providers.

Age also plays a role in the demand for oral health care services. In the US, the elderly tend to have low utilization rates. In 1999, 53.5% of adults 65 and older reported that they had visited a dentist, the lowest rate of any adult age group. Although costs are a factor, even when services are available for free or at a reduced cost or when insurance is available, utilization only increases slightly. Low income and less-educated elders often have lower expectations of good oral health in their old age. As a result, they may be more accepting of pain as a normal part of aging rather than an indicator of the need for oral health care.

Consequences of Poor Oral Health

Economic consequences

Oral disease has negative economic consequences for both individuals and society. Oral disease increases consumers' direct spending on care and also creates indirect expenditures through lost worker productivity. These expenditures could be reduced with a greater investment in preventative care including better oral hygiene habits, decreased prevalence of families consuming unflouridated water, and greater use of dental sealants and fluoride varnish.

Adults also suffer reduced hours of work and earnings when burdened with oral disease. Much of the loss in hours of work appears to accrue to lower income individuals and is often a result of delaying treatment until symptoms are severe. Time lost from work tends to be correlated with previous time lost, low income, being nonwhite, and having poorer oral health. Interestingly, preventative visits account for the most episodes of lost time, but the fewest hours of lost work, suggesting that delaying treatment resulted in greater treatment need. Not only is there a loss in productivity due to the time needed to receive treatment, but poor oral health also appears to affect earnings more generally. The implementation of community water fluoridation during childhood increases earnings for women by 4%, but does not have an effect on men's earnings. These findings are consistent with a differing effect of physical appearance on earnings of women and men.

Among children, oral disease is correlated with greater absenteeism and poorer academic performance. For example, children with oral health pain are three times more likely to miss school due to pain and that missing school due to pain results in poorer school performance. However, the absence for routine oral care is not correlated with poor school performance.

Medical consequences

Traditionally, oral health was viewed in terms of esthetics or localized pain and was compartmentalized from overall health. Recent research, however, has found numerous links from oral health to overall health and well-being, including a correlation with general health, nutrition, digestion, speech, social mobility, employability, self-image and esteem, school absences, quality of life, and well-being. Both bacteria and inflammation resulting from oral disease appears to have a negative association with other chronic diseases such as cardiovascular disease, stroke, adverse pregnancy outcomes, respiratory infections, diabetes, and osteoporosis.

Oral Health Over Time

Over the past few decades, oral health has improved dramatically for the average individual in high-income countries. Adults have fewer dental caries, the prevalence of dental sealants has increased, and the elderly are less likely to have edentulism (i.e., the loss of some or all teeth) and periodontitis. Over the past few decades, the prevalence of cavities in US children has declined, as has the mean number of missing teeth and percentage of edentulous adults. Among the reasons for this general trend are increased utilization of dental care caused by expanded dental insurance coverage and higher incomes, improved quality of dental care, better oral hygiene practices, widespread adoption of fluoridation in public water supplies and fluoride in dental hygiene products, and greater prevalence of sugar substitutes.

Worldwide, trends in oral health are more mixed. International comparisons of oral health typically rely on the

Decayed, Missing and Filled (DMFT) Index. In general, high-income countries have high, but decreasing rates of dental caries. Lower income countries tend to have low levels of dental caries, but the prevalence of caries is increasing. In recent years, there have been an increase in the DMFT index for 12-year olds in the World Health Organization Regions of Africa, Eastern Mediterranean, and Southeast Asia, but a decrease in the Americas, Europe, and the Western Pacific (see Table 1). The result is that the difference in caries experienced by high- and low-income countries over the past two decades has narrowed. The consequence of low levels of oral health care can also be observed in the likelihood of caries being treated. In low-income countries, almost all caries remain untreated, in middle-income countries the proportion of the DMFT index that is filled is only 20%, and in high-income countries the rate is 50%. Within all countries sociobehavioral risk factors play a significant role in oral health outcomes. The increasing consumption of sugar, particularly in areas with inadequate fluoride, and high use of tobacco, is a major risk factor.

Table 1 Regional oral health trends among 12-year olds (DMFT Index)

	2004	*2011*
Africa	1.15	1.19
America	2.76	2.35
Eastern Mediterranean	1.58	1.63
Europe	2.57	1.95
Southeast Asia	1.12	1.87
Western Pacific	1.48	1.39
Global	1.61	1.67

Source: Reproduced from Oral Health Database, Malmö University. Available at: http://www.mah.se/CAPP/Country-Oral-Health-Profiles/According-to-Alphabetical/Global-DMFT-for-12-year-olds-2011/ (accessed 21.02.13).

In the US, utilization of dental services, defined as the percent of adults with a dental visit in the past year, increased dramatically from a little more than 30% in 1950 to more than 65% in 2009 (see Figure 1). Real per capita expenditures have more than doubled over the same time period from $116 to $312 per person. As a result of general improvement in oral health, demand for dental services has shifted toward preventative, diagnostic, and cosmetic care and away from restorative work.

Despite this general trend, there are still segments of the US population that have continued to suffer from generally poorer oral health, such as low-income, minority, and rural populations. Adults 20–64 years of age who are below or near the poverty level (less than 200% of the Federal Poverty Level) are more than twice as likely to have untreated tooth decay than the nonpoor (see Figure 2). Moreover, black and Mexican-American adults are twice as likely to have untreated tooth decay as whites. Similar disparities are found among children. The rate of untreated dental disease among low-income children is significantly higher than that of high-income children. Among 14-year-old white children, the use of dental sealants, a preventive service, is almost four times that among African-American children.

Rural US populations often have poorer oral health than their urban counterparts. Among the reasons for the disparity are lower rates of dental insurance and higher rates of poverty. There are also differences in culture and environment, which may affect the perceived need for dental care. Lower levels of water fluoridation due to reliance on wells and small water supplies may also play a role. Beyond these factors, rural populations also must contend with a lower per capita supply of dentists and longer distances to providers. In 2008, there were

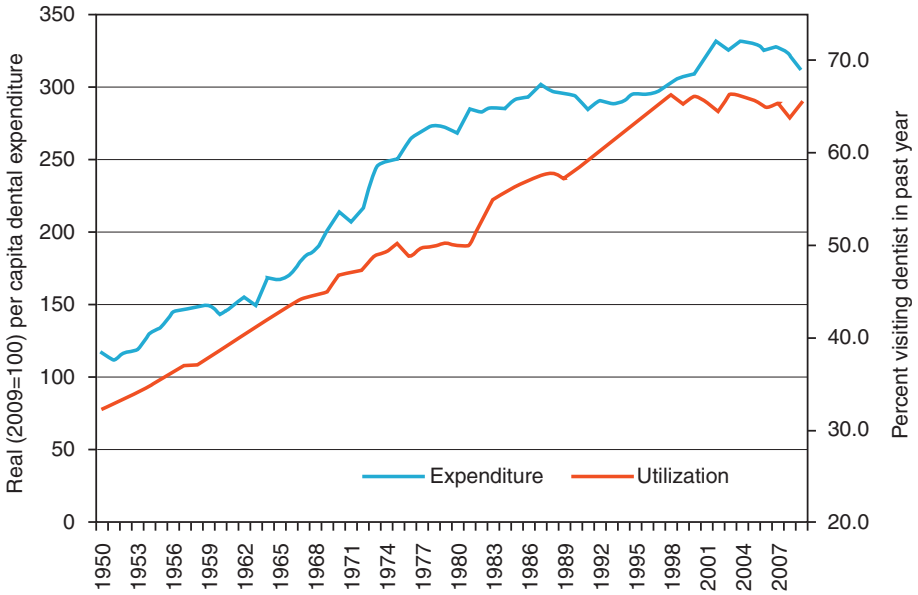


Figure 1 Dental utilization and per capita spending in the US, 1950–2009. Reproduced from U.S. Department of Commerce, Bureau of Economic Analysis (2011). Personal consumption expenditures for dental services, 1950–2009. Available at: http://www.bea.gov/national/nipaweb/index.asp (accessed 25.04.11), and Centers for Disease Control and Prevention, National Center for Health Statistics (2011) National Health Interview Surveys. Available at: http://www.cdc.gov/nchs/nhis.htm (accessed 22.04.11).

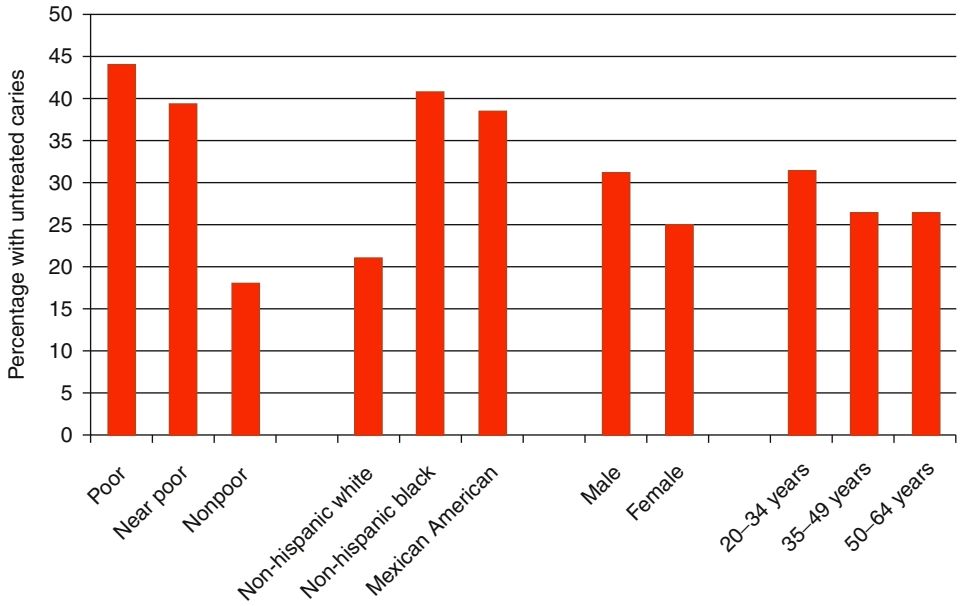


Figure 2 Prevalence of untreated tooth decay in permanent teeth for adults 20–64 years of age in the US, 1999–2004. Reproduced from Dye B. A., Tan, S., Smith, V., et al. (2007). Trends in oral health status: United States, 1988–1994 and 1999–2004. *Vital and Health Statistics* **11**(248), 1–92.

22 nonspecialist dentists per 100 000 population in rural areas in the US and 30 per 100 000 in urban areas. Additionally, a higher proportion of rural dentists were more than 55 years old.

Determinants of Dental Demand

Private Insurance and Income

Countries vary in their use of public or private dental insurance. By reducing the out-of-pocket cost of care, dental insurance can be an important component in the decision to seek dental care. Having private dental coverage significantly increased the proportion of individuals visiting a dentist. In the US, approximately 54% of the population has private dental insurance and 12% has public insurance, leaving 34% without coverage at all. Among those with private dental coverage, 56.9% had a dental visit, whereas only 32% with public coverage had a dental visit and 27% with no dental coverage had a dental visit in 2004. Among people with a dental visit, having insurance is associated with more visits per year and higher dental expenditures. However, some positive correlation between dental insurance and utilization would be expected due to adverse selection. Individuals who expect to need dental care are more likely to buy coverage. Thus, those with insurance would be expected to have higher rates of utilization.

The extent of dental insurance used in other industrialized countries varies. For example, in Norway dental care is provided by private practitioners without public or private insurance, whereas Sweden offers dental services to all adults either free of charge or with a large subsidy. In low-income countries, dental insurance is rare. Oral health services are often provided at urban hospitals where the focus is on pain relief and emergency care rather than prevention or restoration.

Income is also an important component of dental utilization. Of those who are poor in the US only 26.5% had at least one dental visit, whereas the rate was 57.9% among high-income individuals in 2004. However, higher income families are much more likely to have dental coverage. Nonetheless, even after controlling for dental coverage, lower income individuals without coverage are less likely to report a dental visit.

Out-of-Pocket Monetary Cost

Not only is having dental insurance important, but also the generosity of coverage matters. Unlike medical insurance, dental care routinely requires a substantial out-of-pocket payment. In the 10 largest US states, for example, 49.1% of dental expenditures are paid out of pocket, relative to 16.2% for all health care expenditure. Furthermore, utilization of dental services increases significantly as cost-sharing declines. Enrollees in free plans have 34% more visits and 46% higher dental expenses than enrollees in the 95% coinsurance plan. The mix of dental services may also be sensitive to the degree of cost sharing where prosthodontics, endodontics, and periodontics are more responsive to changes in coinsurance. In general, insurance has had a pronounced effect on the use of more expensive dental care, almost doubling the likelihood that a user will obtain bridge work and increasing the probability of a crown by 38%. Dental insurance, however, has had little or no effect on the use of X-rays and dental cleanings.

Evidence from the RAND Health Insurance Experiment, conducted between 1974 and 1982 in the US, found that dental services were significantly more responsive to cost sharing than other out-patient health services during the first

year, but not during subsequent years. The high response during the first year was due to a transitory surge, with individuals taking care of a backlog of problems when low-cost or free care became available. This response was significantly higher than that for other outpatient health services and was observed primarily among low-income groups.

Public Dental Insurance

Although private dental insurance is clearly linked to greater dental utilization, the same trend does not exist with public dental insurance in the US where dental insurance primarily targets low-income children. The most common form of public dental insurance is Medicaid, which typically covers children through age of 20 years, although some limited dental coverage is often available for adults. Medicare for seniors does not include dental coverage. Most US states have found that both enrollment and utilization are both low for Medicaid dental insurance. Nationally, among the children without dental insurance, approximately 3 million are likely eligible for public insurance but had not enrolled. Among those enrolled, often only 20–30% of children actually receive dental care in a given year.

There are several reasons for the low utilization rates. A major hindrance to utilization is that reimbursement rates for dentists serving Medicaid recipients is significantly below usual and customary dental fees in most states, reducing the number of dentists willing to accept Medicaid patients. Dentists also cite administrative difficulties (prior authorization and eligibility verification) and an excessive number of broken appointments as reasons for not accepting Medicaid patients. In fact, Medicaid utilization rates are typically not related to the absolute number of dentists in a county, but rather to the number of dentists accepting Medicaid patients. This suggests that simply increasing the number of providers may not be sufficient to increase use of dental services in underserved areas.

Some states have developed innovative Medicaid programs that have dramatically increased utilization rates. For example, in 2000, Michigan implemented a Medicaid program, Healthy Kids Dental, where in select counties a private insurance carrier, Delta Dental, administered the program and reimbursed dentists at the private rate. The results of the program were to increase utilization by 31.4% overall and 39% among continuously enrolled children. Furthermore, the program resulted in a substantial increase in dental participation and a decline in the distance between providers and the children receiving care.

Time Costs

Beyond the direct monetary costs of dental care, there are also indirect costs to seeking care such as the time spent traveling to care and waiting on service. The empirical evidence on the importance of these costs is inconclusive and measuring the effects is complicated by the fact that individuals often bundle their purchases of dental services with other goods and services, and that provider prices may vary in response to expected wait times. Furthermore, wages, which are the opportunity cost of visiting a provider during working hours, tend to be lower in rural areas. As a result, the cost of the extra travel time is at least partially offset by the lower opportunity cost of time.

Other Variables

Money is not the only determinant of the demand for dental services. Dental anxiety may curtail demand for some individuals. Educational achievement probably affects awareness of the benefits of dental care and may make it possible to lower the costs of obtaining dental care. As one would expect, an immediate need for care as measured by presence of tooth pain or gum bleeding has been found to be associated with a greater likelihood of seeking care. Less obviously, a very low state of dental health may actually lower an individual's need for care. Although poorer quality dentition on average might indicate greater need for care, lost teeth no longer need preventive care and costly restoration procedures over an individual's lifetime. This is one reason why studies investigating the effect of community water fluoridation on dental demand have been inconclusive. Although fluoridation is effective in reducing decay, it results in the retention of more teeth over a lifetime, which could increase the need for care during a person's life.

Determinants of Dental Service Supply

Roughly speaking, the supply of dental services can usefully be broken down into three parts: the supply of dental professionals, the hours those professionals choose to work, and the mix of services offered by dentists, hygienists, and auxiliaries. In the short run, supply of all trained dental professionals and the mix of services offered by each type of professional is relatively fixed. It takes time to gain the required education and begin practicing, and the service mix is largely determined by state licensing regulations. The third factor, hours worked, can respond relatively rapidly to changes in wages.

Dentist Profession

The chief dental care provider is a dentist. Typically, industrialized countries require a dental degree from a university to become a licensed dentist. In the US, entry into the profession requires a graduate degree consisting of four years of training leading to a Doctor of Dental Surgery degree or the equivalent Doctor of Dental Medicine degree. The first two years include basic medical and dental science and the second two years focus on clinical training. In many other countries, a dental degree is provided as a bachelor's degree program.

Graduates of accredited dentistry programs in most countries must also obtain a license to practice dentistry. In the US and Canada, graduates must pass a national licensure board exam and meet other state or province licensure requirements in order to practice. European countries, alternatively, permit free movement within the European Economic Area once a dentist is licensed to practice. However, other restrictions, such

as the ability to treat patients participating in Germany's sickness funds, limit the mobility of dentists.

Dental schools and many other institutions (generally, universities with medical schools, or large hospitals) offer advanced education programs of one to six years duration that train dentists to provide better quality clinical care or specialty care such as endodontics, periodontics, orthodontics, prosthodontics, and oral/maxillofacial surgery.

Supply of Dentists

Supply over time

Higher wages may induce fairly rapid changes in the supply of dentists' services as some dentists postpone retirement or work more hours. However, higher wages could also induce some dentists to work fewer hours, choosing to substitute leisure for work, a phenomenon referred to by economists as a backward-bending labor supply curve because increasing wages have the unexpected effect of reducing the amount of work people are willing to provide. Among dentists in the US, this choice to work less as wages rise may in fact be occurring at the margin. Dentists work, on average, far fewer hours than physicians. Therefore, higher wages could actually reduce the amount of services available. In the long run, expanding the number of licensed dentists will require an expansion of the number of spaces available in dental schools either through expansion of existing schools or the building of new ones.

It is interesting to note that, if dentists are indeed on the backward-bending portion of their supply curve, then increasing the supply of dentists has a larger effect on the supply of dental services than it does on the supply of dentists. Insofar as the added dental graduates drive dentist wages down, then, on average, already licensed dentists will expand their hours. In this way, each new graduate increases the supply of dental services by more than their own contribution of hours to the labor force.

In estimating the supply of dentists needed in the coming years, changes in productivity should also be taken into account. Since 1960, there have been significant fluctuations in productivity, ranging from an increase of 3.95% annually from 1960 to 1974, 0.13% annually from 1974 to 1991, and 1.05% growth annually from 1991 to 1998. The first increase was due to the use of high-speed drills and more auxiliary labor, whereas the 1991–98 increase was likely due to general economic expansion and the further increase in auxiliary employment.

The composition of the dentist workforce can also influence the supply of services. Studies have found differences in hours worked between male and female dentists, as well as differences by the age of the dentist. Older dentists, particularly males, worked fewer hours. Having children reduced the hours work among female dentists. Men and women are equally productive on a per hour basis, but women work part-time twice as often, at least up through age 45.

Internationally, the dentist-to-population ratio varies significantly. Low-income countries tend to have very low dentist to population ratios. In Africa, for example, the dentist-to-population ratio is 1:150 000 compared to 1:2000 in most industrialized countries. Furthermore, most dentists are located in big cities, resulting in even lower dentist-to-population ratios in rural areas.

However, simply increasing the number of dentists may not solve the problem. Between 1985 and 1998, the number of dentists in Syria increased from 2000 to 11 000, resulting in a ratio of 1:1500 dentists per population. The Care Index (F/DMFT∗100%) of the child population remained unchanged and adults only increased from 17% to 33%. Similarly, in the Philippines the dentist-to-population ratio is similar to high-income countries at 1:5000, but the Care Index of children remains very low. The likely explanation is that the majority of the population cannot afford restorative work even when dentists are available.

Licensure and regulation

Occupational licensure has been shown to be an important source of variation among US states in the supply of dentists, other dental professionals, and dental services. Each state sets its own licensure requirements for both new dentists and for those moving into the state. Licensed dentists who wish to move to a new state must obtain a license in that state. This process can be facilitated if states have reciprocity agreements in which state dental licensing agencies agree to recognize the validity of each other's license, or if they have licensure by credential in which states will grant licenses to practicing dentists who have met certain criteria, such as being in continuous practice for a specified period of time.

There are both benefits and costs to occupational licensure. On the one hand, licensure is intended to reduce uncertainty for consumers by ensuring a minimum level of competency or through greater standardization in care. On the other hand, licensing may increase cost and reduce supply by limiting entry into a market. It could also potentially reduce quality by screening out the most qualified individuals. Individuals with a high opportunity cost of time may opt not to enter a profession because of the high cost of obtaining a license.

Licensure generally does not have a significant direct effect on the quality of oral health outcomes, but can influence prices and the supply of dentists. For example, dental records from US Air Force enlistees reveals that stricter regulation has no effect on overall quality of outcomes. Restrictive licensing does, however, raise prices for consumers and earnings for dentists. A state that increased from low or medium to high restrictiveness could expect an 11% increase in the price of dental services. State-mandated restrictions on the number of branches of a dental practice and on the use of dental hygienists also results in higher prices.

Distribution of Dentists within US States

Where dentists settle within the US depends in large part on the size of the state's population and the state's per capita income, both of which are correlated with the per capita number of dental providers. Within the health care sector overall, there is virtually no relationship between the state of degree production and employment. Rather, the production of advanced degrees tends to be concentrated in large, densely

populated states, and providers disperse across the country after degree completion. Often, however, providers do return to their home state after degree completion.

A separate issue from the total number of dentists in a state is how they are distributed within the state. The distribution of providers within states and the decision to locate to rural and/or underserved areas has been studied more thoroughly in the context of physicians than dentists, although many of the findings can be applied to the dental profession. For medical students, having a rural upbringing and specialty preference for rural practice mattered. For medical schools, a commitment to rural curriculum and rotations were the most significant factors in encouraging graduates to locate in rural areas. Similar results were found when UCLA/Drew Medical Education Program students participated in medical rotations in South Los Angeles, an impoverished urban area. After 10 years, 53% of graduates were located in an impoverished or rural area, compared with 26% of other UCLA graduates, even after controlling for race and ethnicity.

Applying these results to dental education, a recent national demonstration project involving 15 schools established goals of increasing senior students' time providing care to underserved patients, educating students to treat underserved populations and expanding enrollment of underrepresented minorities. Results reveal that the quantity of time spent in community settings increased from 10 to 50 days, the participating schools developed courses in cultural competency and public health, and underrepresented minority enrollment increased. However, support from a government sponsored loan repayment program was the most significant predictor of plans to go into public service. Alternatively, increasing educational debt was the most significant barrier to public service plans.

Dental Auxiliaries and Other Providers of Oral Health Services

Types of Oral Health Providers

In addition to dentists, there are a variety of dental auxiliaries and other health professionals that provide oral health services. They include dental hygienists, dental assistants, dental therapists, and dental laboratory technicians. Regulations, training requirements, and the specific functions performed by each auxiliary type varies from country to country, and not every country licenses each type of auxiliary. Table 2 summarizes the training, licensure, or certification typically required, and the functions typically performed by dental auxiliaries found in the US. Dental hygienists focus on preventative care. Dental assistants provide more direct aid, working alongside the dentist. Most states have enacted provisions to permit dental assistants to conduct more tasks after obtaining additional certification.

Dental therapists are less widespread in the US. As of 2011, only Minnesota and some remote parts of Alaska permitted dental therapists. The safety and effectiveness of dental therapists tend to be high. In Alaska, dental therapists exercise good judgment, provide appropriate care, and have highly satisfied patients.

In contrast, dental therapist is a well-established profession in a number of countries. In New Zealand, where dental therapy began in 1921, dental therapists focus on children. The result is more than 60% of children from 2 to 4 years old utilize public oral health services, with an average cost of US $99 the per child. Currently, there are at least 54 countries that use dental therapists, most often staffing school-based programs.

Beyond these standard auxiliaries, individuals may receive oral health care from other providers. Primary care physicians, for example, can be involved early in oral health through a number of possible interventions including screening, counseling, referral to dentists, application of fluoride varnish, and the provision of supplemental fluoride. However, many pediatricians are not actively providing oral health services. More than 90% of pediatricians said that they should examine teeth for caries and educate families, but only 54% did so for more than one-half of their patients between the ages of zero and three. Only 4% of pediatricians regularly applied fluoride varnish. The most common barrier is lack of training.

Regulation of Dental Hygienists

The experience of the dental hygiene profession in the US illustrates the potentially negative effects of restrictive licensure practices for dental auxiliaries. In the US, state dental boards are typically responsible for regulating the dental hygiene profession, making dental hygiene the only licensed profession regulated by another profession. States vary significantly in both their entry requirements, including what is required to obtain a licensure by credentials, and in their

Table 2 Types of dental auxiliaries

Auxiliary	*Typical education/ training in US*	*US Licensure/ certification*	*Typical functions*
Dental hygienists	2–4 years	License	Preventive oral health services including oral prophylaxis and dental hygiene education services
Dental assistants	On-the-job or 1-year program	No (certification optional in most states)	Prepare equipment, update patients' records, work along side dentists during procedures, remove sutures, apply topical anesthetics or cavity-preventive agents, remove excess cement during filling processes
Dental therapist	4–6 years	License	Function of Dental Hygienists plus extractions and simple fillings
Dental laboratory technician	On-the-job or 2 year accredited program	No (certification optional in most states)	Creates dentures, bridges, crowns, and orthodontic appliances by following a dentist's written instructions

scope of practice restrictions, which range from allowing only basic teeth cleaning and polishing services to conducting more complex or potentially hazardous procedures such as administering anesthesia and conducting restorative functions. States also regulate the level of supervision required by dentists, ranging from direct supervision to complete autonomy.

The consequence of restrictive scope of practice regulations is to increase the demand for dentists, while underutilizing hygienists. The US Federal Trade Commission (FTC) estimated that the price effects of state-imposed restrictions on the number of dental auxiliaries that dentists are permitted to employ or the functions hygienists can perform resulted in a 7–11% increase in prices, which cost consumers approximately $700 million in 1982. In 2003, the FTC issued a complaint against the South Carolina Dental Board for prohibiting hygienists from providing teeth cleaning services to Medicaid children. The case was eventually settled. The restrictions dentists have placed on hygienists suggest that dentists view expanded dental hygiene services as a substitute for dental services. What the evidence suggests is that hygienists may also serve as a complement to dentists allowing dentists to specialize in more complex procedures.

The consequence of regulations restricting the use of hygienists and other mid-level providers is often to eliminate the lower cost, lower quality segment of the market. The elimination of the lower cost market segment is troubling when it serves to prevent lower income individuals from purchasing important services and simply forgo treatment that would improve health outcomes. Researchers have found a correlation between restrictive dental hygiene regulations and dental hygienist salaries, dental office visits, hygienist employment levels and, ultimately, access to oral health care.

Conclusion

Oral health has generally been improving in industrialized countries, along with increased utilization and per capita spending on dental care. However, the US experience illustrates that such improvements can occur while significant disparities persist. Tooth decay continues to be a major problem among its low-income, rural, and minority populations. The trend goes the opposite direction for low-income countries where the prevalence of dental caries is increasing from previously low levels because of the adoption of developed nation diets without the corresponding dental care infrastructure.

There are both economic and medical consequences of poor oral health, yet the solutions for improving oral health outcomes are far from clear. Insurance and income are both strong predictors of demand for dental services, but as the US illustrates low reimbursement rates and administrative difficulties can render public insurance for low-income individuals less effective at raising utilization rates. The availability of oral health services is likely to increase in the coming decades. If the international pattern follows that of developed countries, the supply of dentists is likely to increase with rising incomes and greater need in developing countries in the coming decades. Dental auxiliaries and other oral health providers may also play an increasingly important role in the provision of dental services as oral health systems modernize. Without adequately funded and managed public programs to target disadvantaged populations and prudent consumer-friendly regulations, much of this increase in dental service will bypass many of those most in need.

See also: Access and Health Insurance. Health Care Demand, Empirical Determinants of. Health Labor Markets in Developing Countries. Health Status in the Developing World, Determinants of. Occupational Licensing in Health Care. Peer Effects in Health Behaviors. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment

Further Reading

Baelum, V., Van Palenstein Helderman, W., Hugoson, A. and Yee, R. (2007). A global perspective on changes in the burden of caries and periodontitis: implications for dentistry. *Journal of Oral Rehabilitation* **34**, 872–906.

Beazoglou, T., Heffley, D., Brown, L. J. and Bailit, H. (2002). The importance of productivity in estimating need for dentists. *Journal of the American Dental Association* **133**(10), 1399–1404.

Dye, B. A., Tan, S., Smith, V., et al. (2007). Trends in oral health status: United States, 1988–1994 and 1999–2004. *Vital and Health Statistics* **11**(248), 1–92.

Glied, S. and Matthew, N. (2010). The economic value of teeth. *Journal of Human Resources* **45**(2), 468–496.

Kleiner, M. M. (2006). Licensing occupations: ensuring quality or restricting competition? Upjohn Institute.

Kleiner, M. M. and Kudrle, R. T. (2000). Does regulation affect economic outcomes?: The case of dentistry. *The Journal of Law and Economics* **43**(2), 547–582.

Lewis, C. W., Boulter, S., Keels, M. A., et al. (2009). Oral health and pediatricians: Results of a national survey. *Academic Pediatrics* **9**(6), 457–561.

Liang, J. N. and Ogur, J. (1987). *Restrictions on dental auxiliaries*. Washington, DC: Federal Trade Commission.

Petersen, P. E., Bourgeois, D., Ogawa, H., Estupinan-Day, S. and Ndiaye, C. (2005). The global burden of oral diseases and risks to oral health. *Bulletin of the World Health Organization* **83**(9), 661–669.

Skillman, S. M., Doescher, M. P., Mouradian, W. E. and Diane, K. B. (2010). The challenge to delivering oral health services in rural America. *Journal of Public Health Dentistry* **70**(Suppl. s1), S49–S57.

U.S. Department of Health and Human Services (DHHS) (2000). *Oral health in America: A report of the Surgeon General*. Rockville. MD: U.S. Department of Health and Human Services, National Institute of Dental and Craniofacial Research, National Institutes of Health. Available at: http://www.surgeon general.gov/library/oralhealth/ (accessed 29.04.11).

Wanchek, T. (2010). Dental hygiene regulation and access to oral healthcare: assessing the variation across the U.S. states. *British Journal of Industrial Relations* **48**(4), 706–725.

Wing, P., Langelier, M., Continelli, T. and Battrell, A. (2005). A dental hygiene professional practice index (DHPPI) and access to oral health status and service use in the United States. *Journal of Dental Hygiene* **79**, 10–20.

World Health Organization (2001). *Global oral health data bank*. Geneva: World Health Organization.

Development Assistance in Health, Economics of

AK Acharya, OP Jindal Global University, Sonipat, India, and London School of Hygiene and Tropical Medicine, London, UK

Glossary

Agency relationship The relationship between an agent and a principal. Classically in health care, the role of a physician or other health professional in determining the patient's (or other client's) best interest and acting in a fashion consistent with it. The patient or client is the principal and the professional is the agent. More generally, the agent is anyone acting on behalf of a principal, usually because of asymmetry of information. In health care, other examples include health managers acting as agents for their principals such as owners of firms or ministers, regulators as agents for politically accountable ministers, ministers as agents for the electorate. In health care, the situation can become even more complicated by virtue of the facts, first, that the professional thereby has an important role in determining the demand for a service as well as its supply and, second, that doctors are expected (in many systems) to act not only for the 'patient' but also for 'society' in the form, say, of other patients or of an organization with wider societal responsibilities (like a managed health care organization), or taxpayers, or all potential patients. There can be much ambiguity, as in seeking to understand the agency relationships in overseas aid giving and management, and as in establishing the extent to which formal contracts can enhance efficiency.

Aid effectiveness A measure of the effectiveness of aid by examining the contribution of Overseas Development Assistance to the extent to which countries have achieved a reduction in poverty or an increase in growth.

Conditionalities Many countries stipulate that aid is given on particular conditions being met, for example, a package of macroeconomic policies is undertaken.

Development Assistance for Health Overseas Development Assistance that is specifically earmarked for use only within the health sector.

Developmental aid Aid given solely for the purpose of alleviation of poverty or for achieving a higher growth rate compared, say, with aid for military improvements or for foreign policy reasons.

Donor coordination A means of avoiding fragmented aid giving, thereby avoiding needless project duplication. One of the easiest ways to coordinate funding streams is to fund particular ministries rather than single specific projects.

Earmarking International development assistance (aid) or taxes within a jurisdiction stipulated for a particular purpose.

Fungibility A term used to describe the substitutability of one entity for another. For example, money is fungible, in that a ten dollar bill is equivalent to ten one dollar bills. In aid policy, the phenomenon of external funding intended for one purpose but ultimately used by a recipient government for another is another example.

Background

In 1990, development assistance for health (DAH) flowing from the Organization of Economic Cooperation and Development (OECD) countries amounted to only US$4 billon accounted in the index year of 2009. This figure had increased to US$19 billion by 2010, although the year 2009 saw a decline in DAH perhaps due to the economic downturn in the developed countries. Figure 1 shows the dramatic increase since 2000. Although much of it can be due to the commitment to combat human immunodeficiency virus (HIV)/acquired immune deficiency syndrome (AIDS) epidemics, the overall increase is substantial, and there has been recognition of other health care issues as well, as shown in Figure 2. Naturally, questions have risen as to the effectiveness of development assistance for health. Of course, the pathways through which one can examine whether aid has contributed to improved health are extremely difficult to discern. This is one of the issues addressed below. A substantial number of intermediary issues have been examined in the literature. Among the important issues concerning the pathways are what has been recognized as fungibility, coordination, and fragmentation. After a brief description on flow of development assistance on health (DAH), the authors ascertain the current thinking on aid effectiveness.

Trends in Development Assistance on Health

Of the US$127 billion distributed in overseas development assistance (ODA) in 2009 from OECD-DAC, approximately 16% (19 billion) was directed toward health; the corresponding figure for the sub-Saharan Africa (SSA) is 44% (US$12 billion of US$26.7 billion ODA). Thus, health issues play a prominent role in the total development assistance where poverty issues loom large. The prominence of recognition of HIV/AIDS as a global problem resulted in a proportion of DAH going to HIV/AIDS, rising from being approximately 10% of the total amount DAH in 2000 to nearly 40% by 2007 (see Figure 2). Perhaps, due to this crisis there have been proliferations of other actors such as private foundation, global health partnership, and NGOs toward ensuring greater DAH. Once these mechanisms have been taken into account, estimation of total DAH can rise by 20–30%.

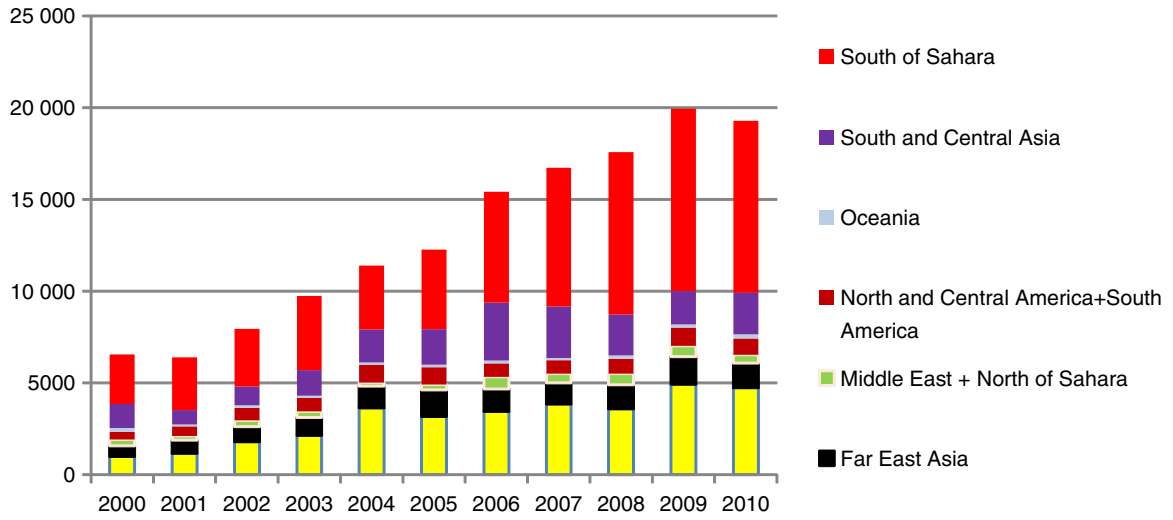


Figure 1 Total and regional patterns in DAH (in millions of 2009 US$). Reproduced with permission from OECD (2013). Available at: http://www.oecd.org/dac/stats/ (accessed 15.07.13).

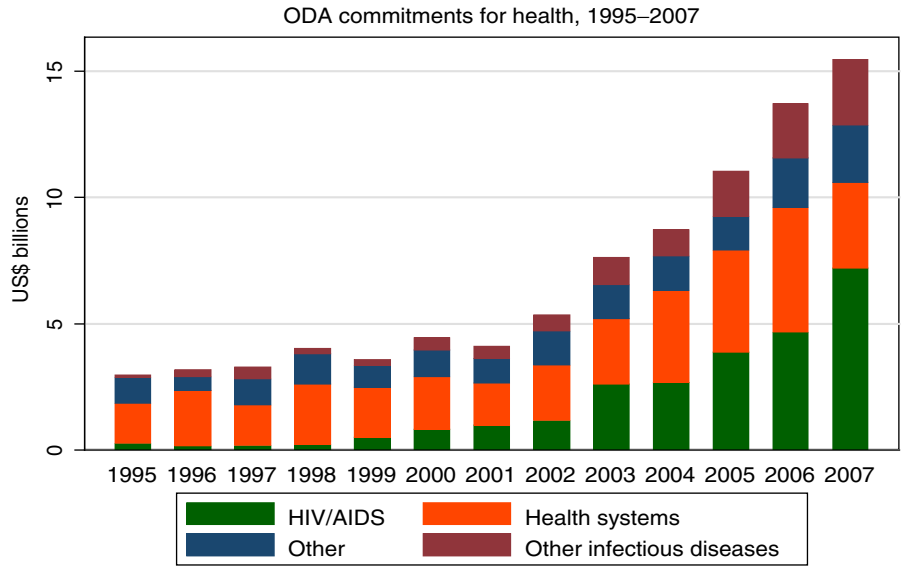


Figure 2 Trends in priorities for development assistance for health (2009 constant US$ billion commitment).

Given that for some countries DAH can amount to a large proportion of the public health budget, the modality of funding is an important issue. A useful literature that feeds into analysis of aid-effectiveness is the examination of the wide variety of funding modalities, which depend on the amount of earmarking for specific usage and the extent to which the myriad of actors rely on government systems for planning, disbursement, and monitoring of funds. The myriad of modalities include direct funding of projects, program aid, sector-wide approaches, and budget support, with projects having the most earmarking and budget support the least. A more recent form of aid giving has been through direct NGO funding. Discerning these channels from the existing data sets has been difficult. At least one study concludes that aid modalities independent of government may induce greater commitment of funding toward health improvement from the government. Although one can take account of a certain degree of endogeneity from the fact that the recipient country's governance structure influences the method through which aid is delivered, these results do not prove conclusive. Knack notes that the use of country infrastructure by the donors is related to (1) donor's share of aid provided to the recipient, (2) perception of corruption in the recipient country, and (3) the public support for aid in the donor country. The perception of corruption in aid has been tied to the issues regarding fungibility. Another concern has been that aid is provided through multiple transfer instances what one may label as aid events; and for a single country there are multiple donors as indicated already. There are two concerns underlying here: firstly that aid giving involves multiple episodes of transactions between the donors and the recipients; and secondly that there may

be too many aid givers. These two issues have been recognized as aid fragmentation, perhaps stemming from lack of coordination among donors.

As ODA can be specified to be used only for developmental use as opposed to military use, within the developmental budget, there is, of course, a mandate as to what can be funded, such as health and education. Various modes have been used to mandate that the health budget have to provide for both specific and general use. One mode of delivery is to offer through basket funding to the government's nonmilitary budget; this has little or no earmarking. A widely used method is called the sector-wide approach (popularly known as SWAP) which can be described as a coordination mechanism for donors working on the same sector. It is a form of budget support where funding is more targeted. Program-based approach has gained prominence; they are characterized by having a single comprehensive program and budget framework, donor coordination in budgeting procedures, management, procurement and reporting. In recent years, nearly exclusive to funding health problems in developing countries are the global health initiatives, which are vertical programs to tackle a single public health issue through a consortium of ODA and private funding.

Effectiveness of ODA is a much discussed recent topic in the economics literature, although initial literature dates back to Pack and Pack in 1983 when much of ODA may have been driven by geopolitical concerns; and there were essentially two donors. The geopolitical nature of aid giving induces Rajan and Subramanian, for example in their now well-known paper, to leave out Egypt.

Measuring Effectiveness of DAH

The unambiguous conclusion from the empirical literature on aid effectiveness as stated by Bourguignon and Sundberg is that it has 'yielded unclear and ambiguous results.' They also state that this should not be a surprise given the politics of aid with the heterogeneity of motives; and more importantly, the complex causality chain linking external aid to final outcomes. For ODA, one of the most important outcome measures of development has been growth rate. Ignoring the mechanism through which ODA may affect the growth rate, a set of oft-quoted studies using reduced form equations have estimated the impact on growth rate to get results that indicate that the relationship between aid and development outcomes is fragile and often ambiguous. The results are slightly more optimistic when some form of mechanism through which aid can affect growth rate has been taken into account. For example, Arndt and coauthors show a positive impact of ODA on growth through a structural model where life expectancy along with investment and education are intermediary factors through which aid affects growth.

The use of reduced form approaches have prevailed in showing the impact of DAH. DAH is, of course, intended for improving health in most cases. It is also part of the developmental aid as opposed to the military or the politically motivated assistance. Clements and coauthors have indicated that for the short run aid allocated to support budget and balance of payments commitments and infrastructure result in rising income. They speculate that aid promoting democracy, health, and education will have a long run impact on growth. Minoiu and Reddy have shown through a Gaussian Mixture Model that developmental aid contributes to growth, whereas the same cannot be said of nondevelopmental aid. Burnside and Dollar conclude that ODA equivalent to 1% of GDP in the recipient country reduces child mortality by 0.9%. Mishra and Newhouse have shown through a Generalized Method of Moments estimation for data from 1975 to 2004 that doubling per-capita health aid decreases infant mortality by 2% for the subsequent five-year period. Earlier, Peck and Peck had showed statistically insignificant results for infant mortality rate.

As mentioned, the mechanism through which aid can improve an outcome is too complex, and simply answering whether an outcome is achieved or not is not very helpful especially if the answer is that the outcome of interest does not seem to have a desirable relationship with ODA. One way to discern pathways through which ODA may or may not work is to ask questions as to whether elements of an economy that ODA funds make for sound policy-making. This may involve macroeconomic analyses or an impact evaluation of projects. It is not possible to evaluate all projects; it is certainly not possible to list all projects that may have used DAH wisely in this monograph. Given the nature of implementation of projects, it is also likely to be misled to list successful projects to be applied from places and times different from the original situation in which they were placed.

The general structure of aid giving is likely to have played a significant role in achieving the outcome which the authors finally arrive at. Many ODA was distributed through conditionality which may have resulted in binding policy makers around donor priorities to ensure policy compliance and implementation. The Paris Declaration on Aid Effectiveness adopted in March 2005, with 100 country signatories, recommended improving aid coordination, promoting donor alignment with country strategies, and cutting the 'compliance burden'. Examining the methods through which ODA have been delivered and the intermediary processes it may engender can help us to understand whether ODA would be effective or not. In doing so, it is not emphasized on private philanthropy which is already engaged in development project funding in a significant way, especially in aiming to improve health. Private philanthropy should not be considered as DAH as for some countries it is also domestic fund; and secondly the rules governing such funding are entirely different from that of ODA funding.

Factors Affecting the Effectiveness of DAH

Clearly a factor that would affect aid effectiveness is whether or not it goes to the right place; that is, do the poorest people of the world receive DAH. Secondly, given that some of the DAH is targeted toward particular activities, are these in some ways the right activities that should be funded? It is then turned to a more subtle point of aid architecture or process factors that might be affecting how well DAH is able to improve health: predictability, fragmentation, and

fungibility. Finally, some important elements that motivate the actions of the players involved in making aid more effective are noted.

Funding the poor

Given that the countries of SSA, the poorest region to which aid flows, have received a large share of foreign aid in terms of DAH, the aim of development assistance to generate human development perhaps is likely to have been met. Despite the published results of a cross-country analysis that found no correlation between countries' GDP per capita and the amount of DAH they received, although this is improving, in terms of per-capita DAH is indeed aimed toward poorer countries. Although in terms of total aid the amount of DAH or ODA is aimed at the country which has the largest number of poor – India, as it is a middle income country, it receives a very low-level of per-capita DAH. Distribution of DAH is fairly consistent with the motive of aiding the poor in the poorest country. Figure 3 shows the relationship between the cumulative proportion of poor (defined as living under US$1 a day) and the cumulative amount of DAH distributed for 56 countries, including India and China, but excluding countries with a population smaller than one million and those for which DAH made up for less than 1% of their total government budget. These countries were ranked by per-capita income, averaged over 1995–2006. For this sample of countries, the first 25 countries amounted to containing 26% of the total poor, whereas the amount of health ODA going to these countries amounted to 51.5% of the total amount of aid in our sample. Of these countries, 22 were in SSA. In some countries DAH does nearly make up for the entire public sector health budget and this may perhaps lead to aid dependency.

Funding Illnesses

As funding can be earmarked, it is important to know how it is earmarked. One way to measure the impact is to see how the burden of illness matches funding. A commonly used measure is disability-adjusted life years (DALYs) for burden. Nugent has shown that while US$0.78 per DALY is allocated toward combating noncommunicable diseases in 2007, US$23.9 per DALY was allocated to HIV/AIDS, Malaria, and tuberculosis. However, a bulk of the latter funding was targeted toward combating HIV/AIDS. Of the US$13.8 DAH that could be accounted for by Ravishankar and coauthors, US$4.9 was spent on HIV/AIDS, compared with US$0.6 billion spent on tuberculosis; the corresponding numbers for malaria and health system were US$0.7 billion and US$0.9 billion. More funding is allocated toward drugs than to human infrastructure.

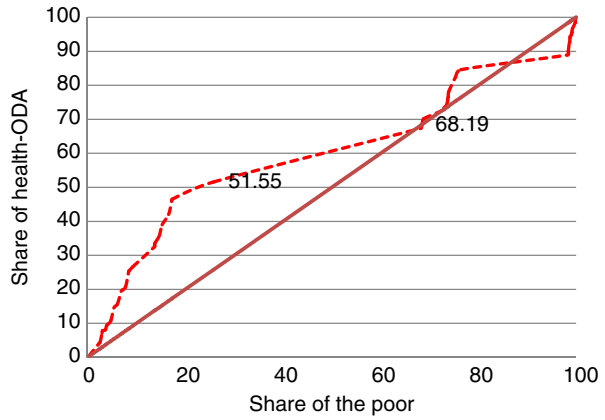


Figure 3 Cumulative distribution of health ODA in relation with the distribution of poor in 2006 (Martinez-Alvarez and Acharya, 2012).

Predictability

For donor countries, ODA is discretionary spending without the backing of any electoral constituency that needs to be placated through political seigniorage. Predictability in regard to ODA is defined by OECD as (1) long-term consistency and (2) disbursement of committed funds in a timely manner. A panel regression of data from 60 low-income countries from the time period of 1990–2005 found that annually there was a great deal of differences between disbursements and commitments, particularly in SSA and the time trend did not show an improvement. Some of this discrepancy can be attributed to a lack of stability in the recipient country. However, the larger reasons for the discrepancy may well be due to the unmet policy conditions by the recipients, donor administrative and political problems. A lack of consistency in funding availability hinders planning for the long term and may force adjustments and changes to original budget plans in the recipient country.

Fragmentation

An increased level of development funding has resulted in the proliferation in the number of donors as well the number of transactions that mobilize the funding processes. Frot and Santiso found Tanzania, a poor stable democracy, had 1601 aid projects in 2007, although the attraction of Tanzania may be due to its stronger institutions. Acharya and coauthors note that fragmentation causes direct transaction costs both to the total aid budget and the recipient country; further, it exacerbates skill shortage in the recipient country by diverting management attentiveness. Anderson shows through econometric techniques that fragmentation does impose administrative costs. Fragmentation may lead to duplication of projects and repetitive activities. Mueller and coauthors have observed that there is great many similar types of training for health workers in Malawi. As fragmentation can be due to the presence of increasing number of donors, Knack and Rahman have emphasized that bearing of responsibilities of outcome of developmental funding can be diluted. Individual country will be less able to claim credit for success; and the result may be that fear of free-ridership induces a lack of effort on the part of donor countries. Fragmentation may also limit economies of scale as project expansion may be limited by the donor's budget ceilings. Principally, the 2005 Paris declaration may have been aimed at fragmentation; 100 countries have recognized that improving aid coordination and promoting alignment with country strategies is a big step toward making ODA more effective.

Fungibility

Fungibility centers around the possibility that ODA becomes a substitute for developmental expenditure that recipient countries are willing to undertake rather than complement the government's developmental budget. It has also become synonymous with corruption. However, fungibility can be seen to be a rational response to sectoral earmarked funds. It also signals that the donor and recipients may have different priorities. One way that a finance ministry can see any type of ODA is to view it as extra revenue. Naturally, for a particular sector the recipient countries would increase the total expenditure in that sector; it may or may not maintain or increase its funding from its own revenue. However, among some policymakers there seems to be an expectation that any increase in DAH should not reduce any domestic expenditure. The response in the academic literature has been very different than policy makers. The academic literature sees fungibility as an extension of the literature on centralized allocation under a federalized system. Economists generally would be surprised by the fact that the local expenditure actually exceeds what would be predicted by the income effect of additional revenue allocated from the central government. Estimates of the extent of fungibility in the health sector for every dollar allocated through DAH on average to a country vary from a decrease in US$0.27–1.65 to a US$1.50 increase. These results depend on the methodologies used including how the dependent variable of total domestic expenditure is calculated. Some factors can be associated with increased fungibility; these include low-levels of GDP per capita of the recipient, fragmentation, and lack of predictability of DAH flow. From a fiscal point of view, the optimal response to lack of reliability of ODA flow is to smooth DAH by spreading it across different years, a practice advised by the IMF.

Motivations and Relations

The issue of fungibility highlights the fact that as economic agents donors and recipients are likely to have different motives. Donors may well be monolithic in their home political structures but by no means in their home country's attitude toward ODA; and the recipient make up for divergent types of governments ranging from those that are war torn to those that have experienced more or less stable democracy since independence. The donors are accountable to their government and domestic public opinion. The recipients stand in relation perhaps to fill the revenue gap for funds needed for developmental project.

As Knack and others have pointed out, the number of donors shapes donor incentives where development can be seen as a public good which is likely to result in donors eluding individual responsibility. The equivalence of Niskanen type of rational bureaucracy on the donor's part may well be toward spending of funds rather than achieving results where the links from funds to outcome may be tenuous. This has come to be known as 'money-moving syndrome.' The consistency between the donor government's motives and development also plays a role. One must also note that the governments of the recipient countries may have different developmental interest or may even have little interest beyond remaining in power. The relation between the donor and the recipient can be understood as something similar to the canonical principal–agent relationship. The donor is not able to judge or monitor the recipient's commitment to development the way that may have been agreed. The donor stands in relation as the principal who may wish for outcomes which can only be achieved by the recipient, the agent. Usual aid practices ignore this fact and unenforceable conditionalities have been usually implemented. Bourguignon and Sunderberg recommend that the aid recipient be free to choose development policies and to implement them and that aid should be "made dependent on observed or possibly foreseeable progress in development outcomes like poverty reduction, improved literacy rates, lower child mortality, etc., and on the observable general quality of policies."

Discussion

What can be highlighted here is that effectiveness of DAH measured in terms of outcome is inconclusive; but most likely DAH along with ODA has not resulted in very significant changes in health outcomes. The key factors affecting the impact of aid on the development that are emphasized here are allocation of resources, donor fragmentation, fungibility of funding, and issues related to making the recipient accountable. That the auhors are unable to gauge the performance of ODA or DAH clearly does not entail that assistance to poor countries should be stopped or even drastically curtailed. Further, political expediency is not likely to move toward such a situation. Thus, making aid effective is a priority for many countries.

Where the link between outcomes and DAH would always be statistically questionable, for ODA relation to be based on performance one would need to examine factors such as governance, country practices, and the outcome results that are observables or can be monitored. An examination of small-scale programs will be valuable toward determining a set of best practices. For successful scaling up of best practices, as has been noted by Medlin and coauthors, the important factors are country ownership, strong leadership and management, and realistic financing.

See also: Disability-Adjusted Life Years. Global Health Initiatives and Financing for Health

Further Reading

Acharya, A., Fuzzo de Lima, A. T. and Moore, M. (2006). Proliferation and fragmentation: Transaction costs and the value of aid. *Journal of Development Studies* **42**, 1–21.

Anderson, E. (2011) *Aid fragmentation and donor transaction costs.* Working Paper 31. UEA, UK: School of International Development.

Arndt, C., Jones S. and Tarp, F. (2011). *Aid effectiveness: Opening the black box.* UNU-WIDER Working Paper No. 2011/44. Helsinki: UNU-WIDER.

Banerjee, A. (2006). *Making aid work: How to fight global poverty effectively.* Working paper. Cambridge, MA, USA: MIT, Economics Department.

Bourguignon, F. and Sundberg, M. (2007). Aid effectiveness – Opening the black box. *American Economic Review* **97**, 316–321.
Burnside, C. and Dollar, D. (2000). Aid, policies, and growth. *American Economic Review* **90**(4), 847–868.
Celasun, O. and Walliser, J. (2008). Predictability of aid: Do fickle donors undermine aid effectiveness? *Economic Policy* **23**, 545–594.
Clemens, M., Radelet, S. and Bhavnani R. (2004). *Counting chickens when they hatch: The short term effect of aid on growth.* Center for Global Development Working Paper 44. Washington, DC: Center for Global Development.
Easerly, W., Levine, R. and Roodman, D. (2003). *New data, new doubts: Revisiting "Aid, Policies and Growth"*, vol. 26. Washington, DC: Centre for Global Development.
Farag, M., Nandakumar, A. K., Wallack, S. S., Gaumer, G. and Hodgkin, D. (2009). Does funding from donors displace government spending for health in developing countries? *Health Affairs (Millwood)* **28**, 1045–1055.
Foster, M. and Leavy, J. (2001). *The choice of financial aid instruments.* London: Overseas Development Institute.
Frot, E. and Santiso J. (2010). *Crushed aid: Fragmentation in sectoral aid.* OECD Development Centre Working Papers 284. Paris: Organisation for Economic Development and Co-operation.
Gottret, P. and Schieber, G. (2006). *Health Financing Revisited – A Practitioner's Guide.* Washington, DC: World Bank.
Institute of Health Metrics and Evaluation (2010). *Financing Global Health 2010: Development assistance and country spending in economic uncertainty.* Seattle, WA: IHME.
Juliet, N. O., Freddie, S. and Okuonzi, S. (2009). Can donor aid for health be effective in a poor country? Assessment of prerequisites for aid effectiveness in Uganda. *Pan African Medical Journal* **3**, 9. Available at: http://www.panafrican-med-journal.com/content/article/3/9/pdf/9.pdf
Knack, S. (2012). *When do donors trust recipient country system.* World Bank Working Paper No. 6019. Washington, DC.
Knack, S. and Rahman, A. (2007). Donor fragmentation and bureaucratic quality in aid recipients. *Journal of Development Economics* **83**, 176–197.
Lahiri, S. and Raimondos-Moller, P. (2004). Donor strategy under the fungibility of foreign aid. *Economics and Politics* **16**, 213–231.
Lu, C., Schneider, M. T., Gubbins, P., et al. (2010). Public financing of health in developing countries: A cross-national systematic analysis. *Lancet* **375**, 1375–1387.
Martinez Álvarez, M. and Acharya A. (2012). *Aid-effectiveness in the health sector*, Working Paper No. 2012/69. Helsinki: UN_WIDER.
McCoy, D., Chand, S. and Sridhar, D. (2009). Global health funding: How much, where it comes from and where it goes. *Health Policy and Planning* **24**, 407–417.
Medlin, C. A., Chowdhury, M., Jamison, D. T. and Measham, A. R. (2006). Improving the health of populations: Lessons of experience. In Jamison, D. T., Breman, A. R., Measham, A. R., et al. (eds.) Disease control priorities in developing countries, 2nd ed., Ch. 8., pp. 161–180. New York: Oxford University Press.
Minoiu, C. and Reddy, S. G. (2010). Development aid and economic growth: A positive long-run relation. *The Quarterly Review of Economics and Finance* **50**, 27–39.
Mishra, P. and Newhouse, D. (2007). *Health aid and infant mortality.* Washington, DC: International Monetary Fund.
Monkam, N. F. K. (2008). The money-moving syndrome and the effectiveness of foreign aid. PhD Thesis, Georgia State University.
Mueller, D. H., Lungu, D., Acharya, A. and Palmer, N. (2011). Constraints to implementing the essential health package in Malawi. *Public Library of Science One* **6**, e2071–e2075.
Nugent, R. A. (2010). Where have all the donors gone? Scarce donor funding for non-communicable diseases. Center for Global Development.
OECD (2008a). *2008 Survey on monitoring the Paris declaration. Effective aid by 2010? What will it take. Overview*, vol. 1. Paris and Washington, DC: Organization for Economic Co-operation and Development.
OECD (2008b). *2008 Survey on monitoring the Paris declaration: Making aid more effective by 2010. Better aid.* Paris: Organization for Economic Co-operation and Development.
OECD (2013). Available at: http://www.oecd.org/dac/stats/ (accessed 15.07.13).
Pack, H. and Pack, J. R. (1993). Foreign aid and the question of fungibility. *Review of Economics and Statistics* **75**, 258–265.
Rajan, R. and Subramanian, A. (2008). Aid and growth: What does the cross-country evidence really show? *The Review of Economics and Statistics* **90**, 643–665.
Ravishankar, N., Gubbins, P., Cooley, R. J., et al. (2009). Financing of global health: Tracking development assistance for health from 1990 to 2007. *Lancet* **373**, 2113–2124.
Stuckler, D., Basu, S. and McKee, M. (2011). International Monetary Fund and aid displacement. *International Journal of Health Services* **41**, 67–76.

Relevant Websites

http://www.cgdev.org/section/topics/aid_effectiveness
Center for Global Development.
http://www.wider.unu.edu/research/current-programme/en_GB/Foreign-Aid-2011/
UN-WIDER.

Diagnostic Imaging, Economic Issues in

BW Bresnahan and LP Garrison Jr., University of Washington, Seattle, WA, USA

Introduction

Diagnostic imaging uses noninvasive devices to visualize internal human anatomy and physiology. In higher-income, developed economies of the world there is enormous variation in the use and rate of growth of use of diagnostic imaging technology like computed tomography (CT). Even in high use jurisdictions like the US there is a large variation. Compared to other Organization for Economic Co-operation and Development (OECD) countries, the most recent per-capita use rates in the US for CT and magnetic resonance imaging (MRI) are more than twice the median OECD rate. There are also substantial geographic variations among the Medicare regions – with the highest per-capita rate of noninvasive imaging in the Atlanta region being 150% greater than Seattle, the lowest use region (Parker *et al.*, 2010). Rapid growth and extreme and variable utilization rates suggest that there may be a lot of inappropriate diagnostic imaging use in the US, and this is especially of concern given that it has relatively poor population-level health indicators among the OECD countries. Inappropriate use has health and economic consequences, which from an economic perspective can be framed as questions of static efficiency and dynamic efficiency. The former is about obtaining good value of money in current spending, and the latter is about eliciting the optimal rate of innovation given the substantial global fixed costs of research and development (R&D). And they are related in that current spending provides the funds needed to support R&D for the long term.

Some characteristics of diagnostic imaging as a medical product – such as strong economies of scale for high fixed cost equipment, informational asymmetry between providers and patients, and the potential for moral hazard – create obvious challenges to promoting its efficient use. For the majority of imaging applications, the marginal cost is relatively low, leading patients and their doctors to seek the information even if the marginal health benefit is low. However, marginal cost pricing would neither cover the fixed costs of the equipment nor provide sufficient funds for long-term dynamic efficiency. However, fee-for-service (FFS) average cost pricing may induce providers to increase volumes to cover the fixed cost, providing incentives for greater use of imaging rather than lesser use – whether the use is appropriate or not. Between 2000 and 2010, physician fees for diagnostic imaging in US Medicare population grew by more than 80%, targeting attention on diagnostic imaging (Medicare Payment Advisory Commission (MedPAC), 2012). Different countries have tried very different approaches to controlling use and costs – discussed elsewhere – the focus here will be on structural economic incentives in the US marketplace and related evidence – such as cost-effectiveness studies, appropriate use strategies, health spending trends, and the impact of payer policies. Finally, designing reimbursement policies that will support appropriate utilization and stimulate levels of innovation consistent with dynamic efficiency is discussed.

The Role of Diagnostic Imaging in Healthcare Delivery

Diagnostic imaging is clearly about reducing uncertainty when diagnosing health conditions. Over the past 120 years, the major 'modalities' of diagnostic imaging (**Table 1**) have become essential components of medical care. Early innovation in imaging was linked to the advent of radiographic imaging (X-ray) in Germany in the 1890s and the introduction thanks to French scientists Pierre Curie and Paul Langevin of ultrasound in the 1940s and 1950s. The so-called 'advanced imaging technologies' were created in the 1970s by Hounsfield in the UK and Cormack in the US with CT imaging or computed axial tomography, again in the 1990s by Bloch and Purcell in the US with MR or MRI, and during the late-1990s by Townsend and Nutt in the US with the nuclear medicine innovation of positron emission tomography (PET) and combined PET/CT imaging.

Imaging can be used in many ways in healthcare delivery. Imaging can inform healthcare decision-making to assist inpatient planning and management, or can be used with interventional procedures (which are not discussed here). Imaging can be used only once or multiple times during the process of making decisions about using specific medications, procedures, surgeries, or other treatments. The resulting information helps the managing physician to refine the diagnosis to support better overall clinical decision-making. This information can increase the likelihood that the patient will ultimately receive the appropriate stream of treatments in order to reduce morbidity and mortality. But the diagnostic test information can also increase the physician's and patient's confidence in the chosen course of clinical action. This can add valuable comfort and peace of mind for the patient. This benefit has been called the 'intrinsic value' or the 'value of knowing,' and can undoubtedly be important to patients and their providers.

Imaging testing relies on a complex mix of specialized labor and capital, information technology (IT) applications, and processes of communication in ordering and reporting. The high-cost equipment producing high-quality images may provide clinical utility through accurate information for treating providers if the scans are ordered, performed, and interpreted appropriately. From an economic perspective, appropriate use would generally be defined as use for which the long-term marginal social benefit exceeds the long-term marginal social cost. Clinically, the goal has been to use a 'correct test, correct indication, and correct timing for the correct patient.' Although it seems likely the majority of use would be found to be appropriate, identifying, and measuring

Table 1 Description of select modalities

Analog radiography, also referred to as conventional plain film X-ray, uses radiation beams that pass through the body and are absorbed in different amounts depending on the density or composition of the anatomic material subjected to the radiograph. Dense bones appear as white on X-rays, whereas organs, fat, or muscles appear as darker in the image.
Digital X-ray machines capture a computerized image rather than using traditional photographic film capture of images. Digital X-ray technology has become the standard for institutions updating their equipment, due to lower radiation emission and faster scanning capabilities.
Mammography machines are specialized low-dose X-ray machines designed specifically for imaging breast conditions. Mammograms are used during the detection and diagnosis of breast disease. Digital mammography, also referred to as full-field digital mammography, uses computerized technology to create the image, evaluate the image, and store the image, rather than printing mammograms on photographic film.
Sonography or ultrasound uses equipment with a transducer probe which emits high-frequency sound waves through the body, reflecting information signals that provide details related to anatomical abnormalities or health status related to pregnancy, thyroid conditions, organ damage, or other internal conditions.
CT imaging combines X-ray equipment with computers. The X-ray is enclosed in a rotating cylinder that sends signals to high-powered computers to produce cross-sectional images of organs or tissue in the body. CT scans generate multiple diagnostic pictures depicting 'slices' of specific portions of the body. CT scans, like conventional X-ray, have radiation exposure risk for patients, so inappropriate utilization is a concern.
Magnetic resonance imaging (MRI) uses magnets encapsulated in a cylinder, rotating around the patient to send strong magnetic fields and radio waves through the body to depict conditions and to detect abnormalities. MRI equipment does not emit radiation and thus does not have the same risk as CT or radiography. However, owing to the use of magnets, MRI use includes restrictions for patients with internal metallic items or specific devices. MRI is best able to depict blood vessels and soft tissues, but does not depict bone structure.
Hybrid imaging: Positron emission tomography (PET) equipment uses a gamma camera along with radioactive pharmaceuticals (tracers) to detect disease and molecular activity. PET imaging is most often aligned with CT (PET/CT) images to coordinate anatomical positioning and computers generate 3-dimensional images of the anatomy, organs, and tumor presence, with its size, spread, and severity in the body. This nuclear medicine modality is able to indicate with high-quality images how tissues and organs are functioning, including molecular function and activity associated with oncologic tumors.

the amount and consequences of inappropriate use is difficult but important. Nonetheless, as will be described below, many of the market and policy responses observed in the US in the past decade represent efforts to control or limit inappropriate use.

Overview of the Market for Diagnostic Imaging

The market for diagnostic imaging is about the demand for and supply of information. All of the various modalities provide the managing physician – i.e., the patient's agent – with additional information to reduce diagnostic uncertainty and therefore to enhance the probability of successful treatment. It is, thus, a derived demand from the patient's point of view, but it is also an imperfect good, subject to some testing and diagnostic inaccuracy. Also, imaging can be subject to considerable moral hazard if neither the patient nor physician face substantial direct monetary consequences. From the supply side, equipment with high fixed and high operating costs are involved, and also complementary services are provided by the radiologist or other interpreting physicians. Throughput in imaging interpretation can be very high (e.g., 7–8 scans read per hour), and the short-run and long-run marginal cost can be fairly low (e.g., the average US payer reimbursement amounts can be on the order of less than US$100 for an X-ray, US$200–300 for a CT scan, and US$400–500 for an MRI). None of these alone would be sufficient to warrant the purchase of catastrophic insurance protection by patients. Still, this equipment represents major investments for most health systems or providers, and many national health systems control their acquisition and deployment. However, in the US, a healthcare system where many insured patients have first-dollar insurance and tax-subsidies for insurance purchase at the margin, it is easy to understand the potential for moral hazard, when the out-of-pocket cost for patients (i.e., often approximately 20% of reimbursed amounts) is far below marginal social cost.

Supply of Equipment: Cost, Location, and Regulation

Imaging manufacturing has a relatively high fixed cost of entry, but the marketplace includes a mix of diversified large global firms, medium-sized innovative cross-industry firms, and smaller niche firms specializing in specific equipment with more focused applications for different conditions. The global market for advanced imaging equipment is substantial – on the order of US$5 billion per year. Manufacturers have an incentive to produce equipment at a quality level that their customers (hospitals, outpatient centers, and physician offices) find sustainable: i.e., given the reimbursement level, the customers can recover their investment and be in a position to upgrade equipment to be competitive in their local provider market. Given that advanced imaging modalities, such as CT or MRI machines, may have a useful life of 5–10 years, individual customers do not purchase new equipment each year, but manufacturers offer improved software upgrades. The larger manufacturers also provide lease/purchase arrangements, financing mechanisms, imaging software, bundled contracting for multiple purchases, and service contracts, effectively reducing the transparency of specific imaging equipment purchase arrangements, and presumably allowing some price discrimination among buyers.

Imaging providers are located in three main locations: hospital facilities, physician offices or clinics, and independent diagnostic testing facilities (IDTFs). Emergency departments are most often connected to hospitals and usually have dedicated imaging equipment available, but may share resources

with the hospital-based inpatient providers. The principal manufacturers are located in Germany (Siemens), the UK (Philips), and the US (General Electric).

Imaging devices are costly and require major investments on the part of health systems. Less advanced imaging equipment, such as ultrasound or analog X-ray machines can range from US$25 000 to more than US$100 000, with the most advanced digital radiography equipment costing several hundred thousand dollars. CT equipment costs may be closer to US$1 000 000, whereas new MRIs and PET/CT scanners can cost US$2 000 000 or more. Service contracts with vendors usually add approximately 8–10% of the purchase price. Countries with national health systems must make major, central financing decisions about purchasing and allocating imaging equipment. In the US, imaging equipment purchases are made by private and public institutions with limited federal guidance or restrictions. The US system is more decentralized, with for-profit and nonprofit hospitals acquiring equipment independently, along with health maintenance organizations (HMOs) and outpatient facilities, such as IDTF. Other more centralized systems use a more publicly reported, transparent planning approach to imaging equipment acquisition.

The regulatory hurdles for new diagnostic imaging devices differ substantially from those of pharmaceuticals and are generally less burdensome. In the US, the Food and Drug Administration oversees safety and efficacy standards for both medical devices and drugs. Devices are categorized as either class I, II, or III, which align with specific premarket authorization notification requirements, or the 510(k) process, as well as with demonstrating that good manufacturing practice compliance standards are met. The class III devices are generally considered higher risk and thus have higher evidentiary standards for manufacturers to meet – more similar to innovative drugs. For lower-risk devices, such as 'next generation' diagnostic imaging equipment that is similar to older models and has relatively marginal modifications, diagnostic equipment suppliers can add innovative features or updates with a limited regulatory evidence requirement: clinical trials are not required.

Demand for Imaging: Moral Hazard and Asymmetric Information among Providers and Patients

At the point of care, managing physicians and their insured patients often have a strong incentive to get more information regardless of social marginal cost. The traditional principal-agent relationship applies to imaging, with the patient being the principal consumer and the physician agents providing technical expertise (e.g., radiologist and ordering provider). The size of the market and the complexity of selecting which test is appropriate and choosing how to interpret images have led to further technical subspecialization among imaging professionals, for example, cardiologists also read scans. Patients often have little knowledge of how imaging works technically and/or which modality, such as CT, MRI, or ultrasound, would be most appropriate for a particular condition. This provides an opportunity for radiologists to gain rents for providing their technical expertise for test appropriateness and test interpretation. Physician preference may increase further imaging in cases where initial testing is not definitive, as indicated in a radiology report or through a sequential testing strategy by ordering providers – though payers place constraints on this, as described below.

Patient demand can also lead to higher rates of medical imaging. Patients are becoming more informed and more active in making their healthcare decisions. Although patients rely on their providers as agents with technical expertise, they are more frequently engaging their providers with information obtained from their networks, the internet, or other sources. Given the influence of third-party insurance payments, because patients do not incur the full cost of imaging services, they may not have a strong copayment disincentive to not be tested. Shifting more financial responsibility to patients in the form of higher copayments for higher-cost imaging, such as for MRI, is being used in some US health systems.

Some observers see physician-induced demand as a factor in this market, particularly when providers have some ownership of the equipment that they are referring patients for imaging. This is, however, a more general phenomenon that links the increased use of healthcare services to physicians having a financial incentive to provide care, whether in the form of more office visits, conducting testing, or performing invasive procedures. In the US, this is generally called self-referral and has been a focus of attention by researchers and the government. **Figure 1** shows the change in specialty practice revenue estimated to come from diagnostic imaging, comparing 2000 with 2006. Imaging by cardiologists and vascular surgeons had the largest increase. Expenditures for imaging associated with self-referral are discussed further below.

Finally, the highly litigious US medical practice has been cited as a contributing factor to providers requesting more imaging, other things equal. Health systems, hospitals, outpatient clinics, and emergency departments may perform more imaging than is clinically or economically appropriate due to concerns of lawsuits and fears of misdiagnosing or under diagnosing a condition without imaging. Although it is difficult to prove that not conducting an imaging test is a wrongful act, providers and systems are at some risk and may choose to image more patients to protect against legal consequences. Tort reform has been suggested by medical professionals as needed to reduce defensive imaging. Smith-Bindman *et al.* (2011) assessed diagnostic imaging tests of the head in an emergency department setting in 10 US states with varying medical malpractice laws. They found that states with more reforms restricting monetary payments from lawsuits against providers or reforms limiting legal fees had a reduced usage of neurologic imaging.

Fee-for-Service Payment and Incentives

Market incentive structures are a key issue when assessing the behavior of providers and the use of diagnostic imaging. In the US, providers are largely reimbursed under the FFS system, creating a limited incentive to reduce the number of imaging tests being conducted, or to put strong mechanisms in place to increase the proportion of appropriate imaging

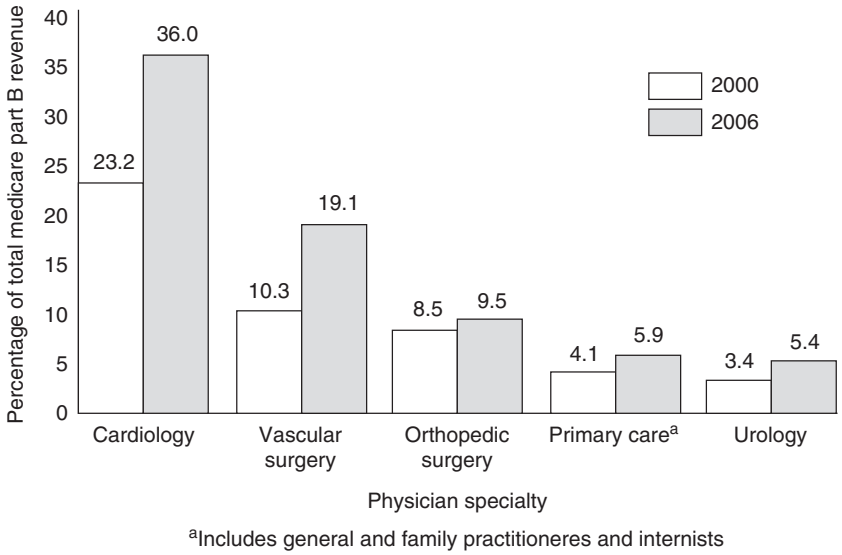


Figure 1 Share of total Medicare part B revenues derived from in-office Imaging services by physician specialty, 2000 and 2006. Reproduced from United States Government Accountability Office (US GAO) (2008a). Rapid spending growth and shift to physician offices indicate need for CMS to consider additional management practices. *Report to Congressional Requestors, Medicare Part B Imaging Services*. Washington, DC: US GAO. GAO-08-452.

performed and to reduce more inappropriate scans. In general, the FFS model applies most directly to outpatient imaging, where a technical (facility) fee is charged, along with a professional fee charged that is linked to work-related relative value units, which serves as a proxy for intensity of provider services used and physician time allocated for a particular service. In addition, for imaging in outpatient, emergency department, and inpatient settings, insured patients most often will be responsible for a copayment associated with imaging services, whereas self-pay patients (uninsured) will be expected to pay the full associated charges with limited ability to negotiate reduced payments. The US system used for current Medicare and non-Medicare reimbursement for diagnostic imaging services is essentially a side product of a system designed primarily to reimburse physician services: the Resource-Based Relative Value Scale. Payment, assigned thorough 9600 Current Procedural Terminology codes, is divided into three parts – physician work, practice expenses, and professional liability – and is adjusted geographically. Approximately 600 of the codes apply to diagnostic imaging, although a small portion of these comprise the majority of use and cost to payers.

Medical charges for imaging services are also not closely tied to actual value delivered, rather tend to be based on expected average cost. Reimbursement amounts for Centers for Medicare and Medicaid Services (CMS's) are often determined through the use of resource use surveys completed by provider facilities, which provide estimates for the amount of time and intensity of resources used to provide a particular service, whether using an older piece of imaging equipment or a newer model. Rates for specific modalities and anatomical regions are modified regularly by CMS as well as by other payers, and can increase or decrease from year to year. It is not immediately obvious whether reimbursement for a particular imaging test or imaging modality would be profitable for a particular facility because it would likely vary by patient subgroup.

Incentives for inpatient imaging in the US may discourage imaging because hospitalizations are reimbursed using assigned diagnosis-related groups (DRGs) based on hospital discharge diagnoses and other factors. The DRG system reimburses hospital providers using a standard 'lump sum' payment per DRG for the hospital stay, with variation based on complication-related modifier codes. In general, in-hospital providers or individual practitioners requesting imaging tests may not have a strong incentive to reduce inpatient imaging for their own patients. The hospital administration, of course, has an incentive to control costs, but subject to professional norms and legal risks: they may seek to promote the appropriate use of imaging.

Modifying payment models is a centerpiece of US health reform and health system experiments. As an alternative to FFS in the US, many HMOs use salary-based models for compensating providers, which is likely to reduce direct financial incentives associated with ordering and conducting imaging studies at these institutions. Rather than paying non-HMO physicians on a salary basis, the US is experimenting with episode-based payments or bundled payments for outpatients, requiring coordination among multiple physicians or groups of providers. These models are expected to be a standard of reimbursement in the near future, at least for specific procedures and patient types. Simple examples of bundling imaging payments are payer policies that combine reimbursement for performing multiple imaging procedures at a lower total amount for imaging procedures that are often conducted together (e.g., 75% of time used simultaneously), rather than reimbursing for each imaging test separately. However, these strategies still do not reimburse based on value delivered: they are cost-based approaches based on utilization metrics (Figure 2).

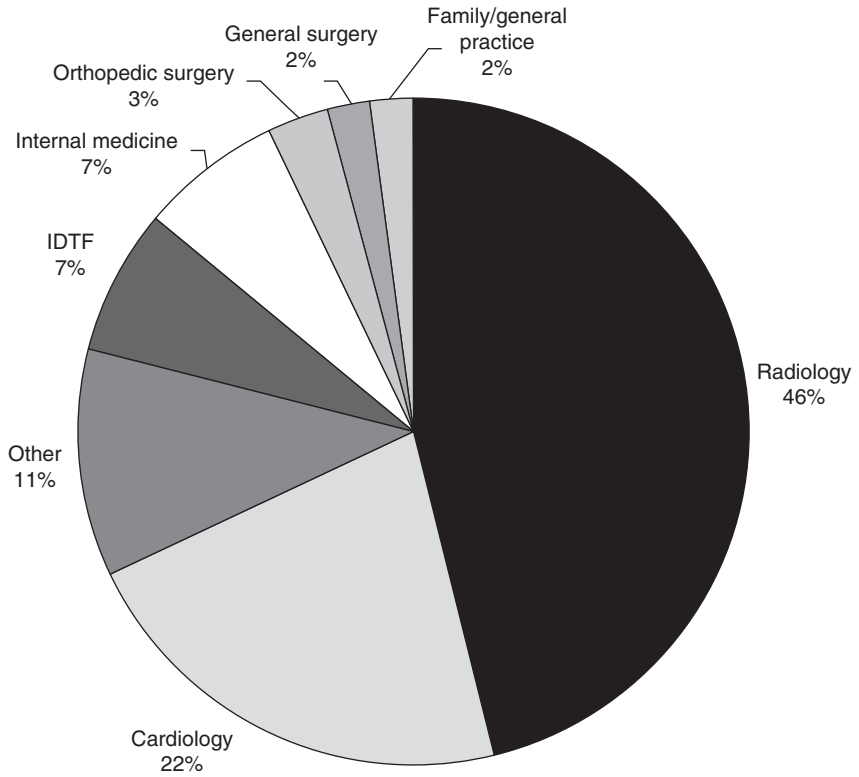


Figure 2 Radiologists received nearly half of physician fee-schedule payments for imaging services, 2009. Reproduced from Medicare Payment Advisory Commission (MedPAC) (2012). A databook: Health care spending and the Medicare program.

Economic and Comparative Evaluations of Imaging Appropriateness

This section begins with a discussion of issues to consider in thinking about an economic framework for analyzing imaging appropriateness. Then, two general groups of evaluations are discussed, in turn: (1) economic evaluations, including cost-effectiveness analyses, and (2) comparative effectiveness research (CER).

An Economic Framework for Evaluating Imaging Appropriateness

The assessment of imaging appropriateness faces several challenges common to devices in general. Drummond *et al.* (2009) assessed differences and similarities for medical devices compared to pharmaceuticals. They identified six primary differences: (1) many devices are diagnostic and do not provide final outcomes, (2) randomized controlled trial (RCT) data are more limited for devices compared to pharmaceuticals, partly due to rapid innovation, (3) device efficacy is in part determined by the device user, (4) there may be more extensive health system impacts for devices, such as training or infrastructure needs, (5) generally less comparative evidence is available for devices, and (6) dynamic pricing flexibility may be greater in device markets due to continuous new product introductions.

Most of these differences also apply to diagnostic imaging specifically. The intermediate nature of diagnostic imaging as informing the referring physician decision-making produces even more challenges for evaluation. Diagnostic scans occur in the middle of the value stream of healthcare delivery, which implies that measuring direct impact on patient outcomes depends on the choice of subsequent interventions. Imaging is more distal to patient outcomes compared to pharmaceutical or surgical interventions. Partly owing to less restrictive regulatory requirements, imaging manufacturers are able to bring new models of scanners to the market with relative frequence, most often without evidence from RCTs. Clinical trials in imaging are often smaller in size and have restrictions in generalizability due to new innovations being introduced during the study. Patients are not likely to enroll in RCTs where they may not be assigned to an arm with the diagnostic test, or the newest test available. All of these challenges contribute to lower levels of evidence for diagnostic imaging modalities.

Fryback and Thornbury (1991) provided an oft-cited hierarchical model for assessing levels of 'efficacy' in relation to diagnostic tests, with higher levels associated more closely with outcomes and economic metrics. The initial assessment for a test should focus technical efficacy (image quality) in Level 1. Once a test has shown good technical performance, Level 2 assesses diagnostic accuracy, sensitivity, and specificity associated in specific patient groups and the interpretation of their scans. Level 3 evaluates the impact on diagnostic thinking or modifications to the diagnostic plan of the referring provider. If diagnostic testing introduces a change in treatment planning, Level 4 assesses the impact on the patient management plan by the ordering provider. Level 5 efficacy measures

the impact on patient outcome, including survival, morbidity, and quality-of-life-metrics. If costs and resource constraints are considered, Level 6 would assess societal efficacy of resources used by evaluating costs and benefits or cost-effectiveness from a societal perspective.

A central feature of imaging is the strong capital and labor interaction that influences whether a test is effective for its intended use. The equipment should produce an accurate image and the radiologist should be able to provide a correct interpretation. However, more advanced and complex imaging tests may have associated imaging artifacts or incidental findings that require a determination on whether it is appropriate to conduct follow-up testing, related to Fryback and Thornbury's Level 3, where diagnostic testing findings can lead to more diagnostic testing. Likewise, lesser trained or lesser experienced radiologists, cardiologists, or other specialists conducting imaging may not interpret scans correctly, or may not be able to provide meaningful summaries of imaging findings, thus providing limited guidance to referring physicians. More advanced imaging equipment and more complicated medical scans will require greater levels of skilled providers to produce the imaging result and interpretation, thus potentially more controls on appropriate use. Newer molecular imaging studies or advanced cardiovascular studies require attenuation correction for image quality and reference points to allow more precision in specific anatomical positioning, and skilled technologists, radiologists, IT personnel, and physics teams are needed to complete accreditation requirements and monitor equipment.

To evaluate the clinical impact and resource use implications of imaging, a model is usually needed to simulate the use of the testing strategy, as well as downstream impacts from either false positives or false negatives. For example, false positives may lead to unnecessary biopsies being conducted for verification. Likewise, imaging may occur for exploring indeterminate results or incidental findings. False negatives from imaging may also have serious health consequences: patients may not receive needed treatments which could lead to disease progression with adverse clinical and/or economic impacts. Therefore, understanding the likelihood of sensitive and specific results from a diagnostic testing strategy and the follow-up implications is essential in estimating the comprehensive impact of diagnostic test use.

The identification of and follow up of incidental findings observed on imaging studies, meaning those not related to the primary condition of interest for obtaining the diagnostic imaging test, can lead to high resource use and high anxiety on behalf of patients. Establishing the costs and health consequences of incidental findings is a critical issue for imaging practice. These variable outcomes not only add complexity to conducting economic evaluations, but also are affected by health system and provider practice variations. Concern about legal liability may influence providers to follow up these findings more aggressively.

Economic Evaluations Including Cost-Effectiveness Analysis

Economic evaluations of diagnostic imaging address a range of different issues and involve a variety of assessment approaches. These analyses study: (1) diagnostic test utilization patterns, referral patterns and imaging, (2) impacts on use after introduction of decision support systems, (3) trends for specific conditions, and (4) the cost-effectiveness of particular diagnostic strategies in defined subpopulations. The data sources include observational studies, retrospective reviews, prospective studies, and secondary database analyses. As noted, imaging studies may result in findings that require further testing to more conclusively determine if patients have a condition or not. Clinical studies may not capture these comprehensive sets of events that occur due to diagnostic imaging, and retrospective analyses are suboptimal due to a limited availability of health status data or clinical information. For a variety of reasons, medical record data and resource use data are often not connected electronically.

Cost-effectiveness analysis (CEA) and cost-utility analysis (CUA) are important tools to assess the potential appropriateness of imaging interventions. Their growth has paralleled the growth in imaging spending and payer requirements to demonstrate value for expenditures. Cost-utility studies generally estimate cost per quality-adjusted life-years (QALYs) gained (a combined time and quality-of-time metric). In 2008, Otero and colleagues evaluated 20 years of cost-effectiveness studies for radiology (1985–2005), providing an assessment of 111 published CUAs. During this period, there was an increase from a few CUAs each year to approximately 10 per year. Nearly 80% of the CUAs they identified pertained to diagnostic radiology. They summarized studies by modality and disease/condition. Ultrasound and angiography were the most frequently studied imaging tests, followed by MRI and CT. The five most frequently assessed disease areas were peripheral vascular disease, cerebrovascular disease, ischemic heart disease, musculoskeletal and rheumatologic disease, and lung cancer. Importantly, approximately 80% of studies used secondary data from the literature to estimate quality of life 'utility scores' for the QALY estimation rather than primary data collection. This highlights the need for more comprehensive prospective studies to assess the economic impact of imaging on patient outcomes.

Most economic evaluations of diagnostic imaging have estimated the marginal effects of imaging interventions on particular types of patients by comparison with alternative testing strategies. The incremental costs and consequences associated with using health resources for one condition or type of medical test can be compared with those costs and outcomes from using other tests, and potentially compared among conditions. To date, the number of well-designed imaging evaluation studies is still very limited. The clinical scientific imaging literature has predominantly focused on diagnostic accuracy characteristics and comparisons. Recently, more incremental cost-effectiveness studies of imaging are being conducted and published. But these studies face considerable challenges in sorting out the heterogeneity associated with estimating cross-population or cross-indication effects associated with implementing diagnostic testing guidelines.

Economic assessments can be conducted at a health system level as well as for a typical patient with a health condition. Consider a policy that tries to encourage adherence to findings from a diagnostic test that indicates a low likelihood that a

surgery would improve a patient's morbidity or mortality status. This could result in fewer surgeries of that type being performed, thereby reducing the aggregate number of surgeries expected to have suboptimal outcomes and lower cost. Likewise, a diagnostic test that leads to additional testing, treatments, or procedures may result in other patients not receiving specific procedures or care, particularly in systems with a fixed health budget. Therefore, a comprehensive assessment of the economic impact on the health system of using a diagnostic test should include these direct and indirect effects. Practically speaking, however, few, if any, economic assessments of diagnostic imaging interventions have taken a comprehensive societal perspective.

Establishing a Comparative Framework for Appropriateness

In recent years, there have been calls for more CER in imaging. CER incorporates multiple stakeholder perspectives (e.g., patients, providers, payers, and systems) and attempts to identify those medical products or programs that provide substantial benefits to patients and those that do not. In addition, effectiveness data for patient subgroups are often lacking. In 2009, an Institute of Medicine report provided recommendations on the top 100 national priorities in the US for comparative effectiveness research, citing advanced imaging (CT, MRI, PET, and PET/CT) in oncology as a top-tier priority for additional comparative studies. A total of nine topic areas relevant to imaging approaches were included in the top 100 priorities. Most of these nine areas included recommendations to compare multiple imaging modalities used in specific indications.

Gazelle *et al.* (2011) suggested a framework – aligned with the Fryback–Thornbury hierarchy – for thinking about CER in diagnostic imaging. They suggest that designing practical evidence requirements for imaging technologies should consider the size of the population at risk for a condition, the likely clinical impact of imaging, and the overall cost impact of diagnostic testing, including the cost of the test, subsequent costs of treatments and testing, and the impact on payers' budgets. In a market-oriented system such as the US, differences in access to healthcare can affect health outcomes among ethnic, racial, or income groups. The CER initiatives and national priorities identified racial and ethnic disparities as a primary area of US healthcare requiring more comparative evaluations.

Imaging Utilization Management Strategies and Appropriateness Tools

The goal of improving the appropriateness of diagnostic imaging has been addressed in multiple ways by the various stakeholders, including payers, providers, professional societies, and policy makers. Six tools aim to limit overuse and promote efficiency are briefly described in this section: (1) professional appropriateness criteria, (2) radiology benefits management (RBM), (3) clinical decision support (CDS), (4) coverage with evidence development (CED) by CMS, (5) Congressionally mandated, across-the-board reductions in payment amounts, and (6) quality improvement (QI) metrics.

First, imaging appropriateness criteria have been developed by the American College of Radiology. Duszak and Berlin (2012) provide an overview of their rationale and a historical perspective of utilization management. Owing to the overall scarcity of comparative imaging evidence and long-term outcomes studies, these types of criteria often rely on expert opinion, supported by the medical literature when available: they are most often not based on large randomized studies or strict evidence-based clinical guidelines. However, they serve as a guide to using imaging more appropriately, which can reduce testing that does not provide high marginal clinical benefit but does impose cost on the system.

Second, RBM gained momentum in the 1990s as a mechanism to control use and costs, similar to prior authorization requirements for prescribing expensive biotechnology medications. Insurance companies hire RBM brokers to manage their imaging-related benefits, such as in requiring preauthorization for MRI scans or other expensive tests. Although providers argue these systems are restrictive and remove patient-provider preferences from decision-making, assessments have shown a reduction in imaging expenditures related to a 'gatekeeper effect.'

Third, CDS systems are generally less restrictive than RBMs and are more real-time use oriented at the point of ordering, but require a computerized ordering system. Providers enter patient-level clinical and demographic information, including diagnosis codes, and then request an imaging test. The tools provide an appropriateness score based on an embedded algorithm. Individual imaging managers or health systems can decide how restrictive to make the algorithm and whether to allow all orders to be processed or to disallow imaging tests based on specific appropriateness ranges.

The introduction of RBMs and CDS systems to control imaging requisitions highlights a key point related to imaging use: radiologists do not typically order scans. They traditionally have not served as effective self-regulating providers who perform only appropriate imaging tests based on requisitions; hence, ordering-point controls have arisen.

Fourth, CMS has attempted to control or better understand imaging use though CED. In 2005, a CED approach was used to require enrollment in a registry as a mechanism to restrict coverage and reimbursement for PET and PET/CT scans in oncology indications while more evidence was gathered about this new modality. CED can effectively slow the diffusion of new products and the CED cohort can be linked to utilization claims to evaluate how innovations impact overall utilization.

Fifth, in a broad, national-level effort to control costs, the US Congress passed the Deficit Reduction Act (DRA) of 2005, which included mechanisms to slow the growth of medical spending on Medicare and Medicaid. The law imposed reductions in reimbursement rates for imaging services, as well as allowing states to modify conditions associated with Medicaid programs. At the state level, in particular, access to care for lower-income individuals and families was affected due to provisions allowing states to modify eligibility or documentation requirements, with a goal of saving billions of dollars in the Medicaid program.

Finally, QI metrics are being used by providers, payers, and health technology assessment organizations to directly and indirectly incentivize appropriate care. For example, emergency department throughput reporting is required by CMS, direct cost comparisons for academic medical centers have been added to the University Hospital Consortium, and reimbursement restrictions are being implemented for cardiovascular patients having hospital readmissions within 30 days of discharge. Imaging in inpatients impacts overall efficiency of workflow and net reimbursement for providers, leading hospitals to evaluate imaging use and direct costs relative to other hospitals.

Trends in US Imaging Spending: Growth and Controls

In terms of the growth in imaging spending, the period since 2000 can be divided into two intervals: the period before the DRA of 2005 and the period since then. In the first period, imaging growth outpaced all other medical expenditures, leading to the initiatives described above. In the period 2006–11, imaging growth slowed, and in later years even declined in evaluations of specific payers. Nonetheless, attention to physician ordering patterns and geographic variability continues to be a target for standardization as well as efficiency and appropriateness assessment.

A 2008 report by the US Government Accountability Office (US GAO, 2008a) evaluated use and expenditures for different imaging modalities, including MRI, CT, nuclear medicine, ultrasound, X-ray, and other procedures between 2000 and 2006. They analyzed trends in the number of tests per Medicare beneficiary and the estimated payments for technical and professional fees for imaging per beneficiary. Overall, there was a steady increase in imaging use and payments to physicians in per beneficiary spending. The report also highlighted substantial state-level variability in imaging outpatient expenditures per beneficiary, ranging from less than US$100 per beneficiary to greater than US$400, with Florida and Nevada having the highest levels of per beneficiary spending. The GAO assessment noted a shift in the proportion of physician services paid through physician offices and IDTFs, rather than through institutional outpatient settings of hospitals.

Overall Medicare Part B spending on imaging during this period increased from under US$7 billion to more than US$14 billion, including imaging in hospitals, provider offices, and IDTFs (Figure 3). The overall size of the imaging spending pie doubled, although the allocation of spending shifted more heavily toward nonhospital imaging (Figure 3). These expenditure trends were influenced by higher rates of increases in advanced imaging, such as for CT, MRI, and nuclear medicine (Figure 4).

Levin *et al.* (2011) assessed Medicare trends in noninvasive diagnostic imaging from 1998 to 2008 and reported steady increases in overall imaging utilization rates during this period. Their assessment of advanced imaging showed that CT rates per 1000 Medicare beneficiaries continued to increase, but MRI and nuclear medicine testing started to level off from 2005 to 2008. A MedPAC assessment of essentially the same time period indicated that the number of head CT increased from 112 per 1000 Medicare beneficiaries in 2000 to 205 per 1000 in 2010. In the same period, all other CTs increased from 258 per 1000 to 548 per 1000 beneficiaries. Overall MRI rates essentially doubled from 2000 to 2010, with MRI of the brain increasing from 45 per 1000 beneficiaries to 79 per 1000, and all other MRIs increased from 64 per 1000 to 141 per 1000

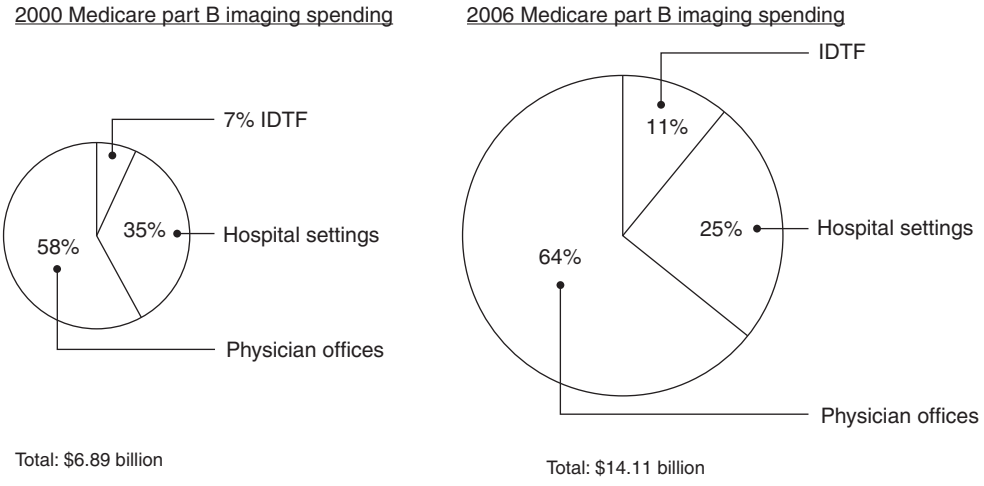


Figure 3 Medicare part B spending on imaging by setting, 2000 and 2006. Hospital settings include inpatient and outpatient departments and emergency rooms. The IDTF category also includes imaging services provided in other outpatient facilities such as mammography screening centers and independent physiological laboratories that are paid under the physician fee schedule. Expenditures include fees for physician interpretation of imaging services in hospital settings, and fees for interpretation and provision of services in physician offices and IDTFs. When the imaging examination is performed in an institutional setting, such as a hospital or skilled nursing facility, the physician can bill Medicare only for interpreting the examination, while payment for performing the examination is covered under a different Medicare payment system. Reproduced from United States Government Accountability Office (US GAO) (2008a). Rapid spending growth and shift to physician offices indicate need for CMS to consider additional management practices. *Report to Congressional Requestors, Medicare Part B Imaging Services.* Washington, DC: US GAO. GAO-08-452.

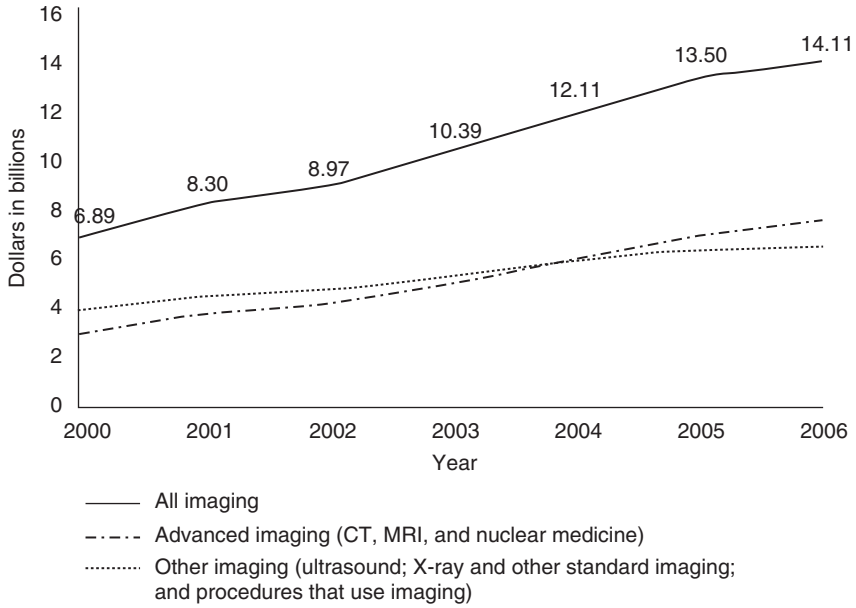


Figure 4 Total Medicare expenditures for imaging services paid under the physician fee schedule, 2000–06. Reproduced from United States Government Accountability Office (US GAO) (2008a). Rapid spending growth and shift to physician offices indicate need for CMS to consider additional management practices. *Report to Congressional Requestors, Medicare Part B Imaging Services*. Washington, DC: US GAO. GAO-08-452.

beneficiaries (Medicare Payment Advisory Commission (MedPAC), 2012).

A recent 2012 GAO report evaluated the role of self-referral of imaging services to assess Medicare trends in imaging utilization. As noted earlier, self-referral implies that an ordering provider has an ownership interest in a facility they direct patients to have imaging testing. The report presents the number of CTs and MRIs between 2004 and 2010 for self-referred imaging and nonself-referred imaging. Between 2006 and 2010, nonself-referred MR imaging was associated with a decline in office-based or IDTFs. Self-referred MRIs continued to increase during this same period. For CT services, nonself-referred imaging increased at a slowing rate between 2004 and 2009, and declined from 2009 to 2010. Self-referred CTs, however, although lower in magnitude, gradually continued increasing throughout the period 2004–10.

Levin *et al.* (2010) found a decreased effect on outpatient imaging rates in a multistate pre-RBM and post-RBM analysis of a large private insurer introducing this control mechanism. Rates of CTs, MRIs, and PET scans were reduced subsequent to the introduction of increased management of ordering by approximately 10–20% measured as the number of imaging studies per 1000 members. Blackmore *et al.* (2012) summarized the impact of several imaging utilization programs tested in statewide initiatives, hospitals, and individual health plans. An observed reduction in growth rates for CT scans in Massachusetts General Hospital (MGH) of approximately 10% was greater than the impact on MRI growth, which was negligible. In Minnesota, the Institute for Clinical Systems Improvement (ICSI) coordinated a CDS initiative with five large provider groups, and reported a restriction in imaging growth to nearly zero. In MGH and ICSI, these tools were also used as an education tool for providers consistently ordering less appropriate scans, with an intention of changing behavior. A Virginia Mason CDS tool implemented in Seattle, WA was able to reduce overall imaging rates in target conditions.

A 2008 GAO report estimated the impact of the DRA on imaging expenditures in the Medicare population (US GAO, 2008b). The outpatient prospective payment system cap resulted in fee reductions for approximately 25% of overall imaging tests, with a greater relative reduction for advanced imaging tests. Following DRA outpatient reimbursement rate reductions, overall imaging expenditures per beneficiary decreased by approximately 10% between 2006 and 2007, although the number of total imaging tests per FFS Medicare beneficiary continued to increase.

The overall financial impact of QI initiatives focused on efficiency metrics have likely put downward pressure on costs and potentially have reduced inappropriate imaging, although no comprehensive published studies are available. Several policies have been recently introduced, so there is limited data on the effects of QI programs on expenditures. As the impacts of broader US health reform and QI are studied in the coming years, the impact of cost-focused QI programs will be better understood.

Recently, Lee and Levy (2012) analyzed multiple samples of insured populations and found that annual rates of CT and MRI growth, although increasing fairly rapidly from 2000 to 2006, started to decline after 2006. In some instances, CT utilization per 1000 beneficiaries showed absolute declines in use in 2008 and 2009. Their evaluation of a combined set of 47 health plans indicated a doubling of CT and MRI rates per 1000 plan members between 2002 and 2009, but MRI growth was close to zero from 2006 to 2009. They describe that findings were contemporaneous with general policy trends of increased prior authorization, CDS, RBM, and general

economic challenges. Nevertheless, more attention toward appropriateness and utilization strategies – including more attention to CT radiation dose exposure–seemed to contribute to slowing the growth of imaging.

It would, of course, be very difficult to estimate accurately the resulting health impacts on society or to separate the role of each of these influences. It is also difficult to claim causal effects due to any specific payer or government policy related to imaging, but taken as a whole, the attention placed on reducing imaging expenditures since 2005 was associated with at least a leveling off of growth rates and a bending of the cost curve for imaging. However, the US still uses the highest amount of advanced imaging of nearly all countries, spends more on healthcare in general, and does not have adequate structural incentives to encourage substantial reductions in diagnostic imaging. Although imaging is not the highest category of US medical expenditures, more alignment with appropriateness at the point of imaging ordering should improve the static efficiency of use.

Innovation and Dynamic Efficiency

Given the complexities of providing diagnostic imaging in the US healthcare system, it is unclear how much use is inappropriate (i.e., inefficient in a static sense), and it also unclear how much underuse there is for those with access problems. Furthermore, no national estimates are available of the social cost of either overuse or underuse.

In such a second-best world, it is also unclear how close the system is to achieving dynamic efficiency, i.e., eliciting the optimal amount of innovation from a longer-term perspective. Given the lack of hard evidence and estimates, reasoning about incentives may be the best option. In this vein, it has been argued that the lack of value-based reimbursement in radiology is likely to inhibit innovation (Garrison *et al.*, 2011). In theory, fixed payments per scans of different types lead manufacturers to provide a quality level of imaging that is only financially sustainable within that payment limit. Furthermore, it is not clear whether the amount the reimbursement system pays for imaging results is being divided between the capital equipment owners and scan readers (usually the radiologist) in a manner that supports optimal capital innovation. The science behind imaging is a global public good, and the equipment is sold worldwide, including the sale of lower quality or refurbished equipment in developing country settings. Such differential pricing can provide greater support for research and development and counter the incentives in the US that might hinder the rate of innovation.

Conclusion

Advanced imaging modalities have revolutionized medical practice by improving clinical diagnostic ability to meet the goal of having reliable, condition-specific test results to support better decision making. The potential benefits associated with an improved ability to accurately diagnose medical conditions using advanced imaging should be weighed against the resource costs for payers and society in order to assess the appropriateness and efficiency of its use. However, specific features of diagnostic imaging provide unique challenges for economic evaluations. Also, policy attention to imaging use has increased in an effort to target the most rapidly increasing components of medical imaging. Rates of spending growth have slowed since 2007 due presumably to several payer and policy initiatives, but overall imaging spending remains high. Nonetheless, there is clearly a dearth of economic research on either the actual cost-effectiveness of specific imaging applications or on the impact of current reimbursement rules and other market incentives on health system performance. At best, most cost-effectiveness analyses show only the potential value of appropriate imaging in specific applications. Continuing high and variable utilization rates suggest significant overuse in the US. Across-the-board cuts and other utilization controls have curbed spending growth, but the extent of inefficiency – both static and dynamic – remains unclear.

See also: Adoption of New Technologies, Using Economic Evaluation. Analysing Heterogeneity to Support Decision Making. Biopharmaceutical and Medical Equipment Industries, Economics of. Budget-Impact Analysis. Cross-National Evidence on Use of Radiology. Economic Evaluation, Uncertainty in. Information Analysis, Value of. Medical Decision Making and Demand. Observational Studies in Economic Evaluation. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Primary Care, Gatekeeping, and Incentives. Problem Structuring for Health Economic Model Development. Research and Development Costs and Productivity in Biopharmaceuticals. Statistical Issues in Economic Evaluations. Value of Information Methods to Prioritize Research

References

Blackmore, C. C. and Mecklenburg, R. S. (2012). Taking charge of imaging: Implementing a utilization program. *Applied Radiology* 18–23.

Drummond, M. F., Griffin, A. and Terricone, R. (2009). Economic evaluation for drugs and devices – Same or different? *Value in Health* **12**(4), 402–404.

Duszak, R. and Berlin, J. W. (2012). Utilization management in radiology, part 1: Rationale, history, and current status. *Journal of the American College of Radiology* **9**, 694–699.

Fryback, D. G. and Thornbury, J. R. (1991). The efficacy of diagnostic imaging. *Medical Decision Making* **11**, 88–94.

Garrison, L. P., Bresnahan, B. W., Higashi, M. K., Hollingworth, W. and Jarvik, G. J. (2011). Innovation in diagnostic imaging services: Assessing the potential for value-based reimbursement. *Academic Radiology* **18**(9), 1109–1114.

Gazelle, G. S., Kessler, L., Lee, D. L., et al. (2011). A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. *Radiology* **261**(3), 692–698.

Lee, D. W. and Levy, F. (2012). The sharp slowdown in growth of medical imaging: An early analysis suggests combination of policies was the cause. *Health Affairs* **31**(8), 1876–1884.

Levin, D. C., Bree, R. L., Rao, V. M. and Johnson, J. (2010). A prior authorization program of a radiology benefits management company and how it has affected utilization of advanced diagnostic imaging. *Journal of the American College of Radiology* **7**, 33–38.

Levin, D. C., Rao, V. M., Parker, L., Frangos, A. J. and Sunshine, J. H. (2011). Bending the curve: The recent marked slowdown in growth of noninvasive diagnostic imaging. *American Journal of Roentgenology* **196**, W25–W29.

Medicare Payment Advisory Commission (MedPAC) (2012). *A databook: Health care spending and the Medicare program*. Washington, DC: MedPAC.

Parker, L., Levin, D. C., Frangos, A. and Rao, V. M. (2010). Geographic variation in the utilization of noninvasive diagnostic imaging: National Medicare data, 1998–2007. *American Journal of Roentgenology* **194**, 1034–1039.

Smith-Bindman, R., McCulloch, C. E., Ding, A., Quale, C. and Chu, P. W. (2011). Diagnostic imaging rates for head injury in the ED and states' medical malpractice tort reforms. *American Journal of Emergency Medicine* **29**, 656–664.

United States Government Accountability Office (US GAO) (2008a). Rapid spending growth and shift to physician offices indicate need for CMS to consider additional management practices. *Report to Congressional Requestors, Medicare Part B Imaging Services*. Washington, DC: US GAO. GAO-08-452.

United States Government Accountability Office (US GAO) (2008b). Trends in fees, utilization, and expenditures for imaging services before and after implementation of the deficit reduction act of 2005. Report to Congressional Requestors, Medicare. Washington, DC: US GAO. GAO-08-1102R.

Further Reading

Bresnahan, B. W. (2010). Economic evaluation in radiology: Reviewing the literature and examples in oncology. *Academic Radiology* **17**, 1090–1095.

Duszak, R. (2012). Medical imaging: Is the growth boom over? *The Neiman Report, No. 1*. Reston, VA: Harvey L. Neiman Health Policy Institute.

Gazelle, G. S., McMahon, P. M., Siebert, U. and Beinfeld, M. T. (2005). Cost-effectiveness analysis in the assessment of diagnostic imaging technologies. *Radiology* **235**(2), 361–370.

Hollingworth, W. (2005). Radiology cost and outcomes studies: standard practice and emerging methods. *American Journal of Roentgenology* **185**, 833–839.

Institute of Medicine (2009). Initial national priorities for comparative effectiveness research. Report Brief. June. Available at: http://www.iom.edu/CMS/3809/63608/71025.aspx (accessed 21.03.13).

Levin, D. C. and Rao, V. M. (2008). Turf wars in radiology: Updated evidence on the relationship between self-referral and the overutilization of imaging. *Journal of the American College of Radiology* **5**, 806–810.

Massachusetts Medical Society (2008). Investigation of defensive imaging in Massachusetts. Available at: http://www.massmed.org/AM/Template.cfm Section=Research_Reports_and_Studies2&TEMPLATE=/CM/ContentDisplay.cfm&CONTENTID=27797 (accessed 14.03.13).

Miller, R. A., Sampson, N. R. and Flynn, J. M. (2012). The prevalence of defensive orthopaedic imaging: A prospective practice audit in Pennsylvania. *Journal of Bone and Joint Surgery* **94**(3), e18, doi:10.2106/JBJS.K.00646.

Otero, H. J., Rybicki, F. J., Greenberg, D. and Neumann, P. J. (2008). Twenty years of cost-effectiveness analysis in medical imaging: Are we improving? *Radiology* **249**(3), 917–925.

Pandharipande, P. V. and Gazelle, G. S. (2009). Comparative effectiveness research: What it means for radiology. *Radiology* **253**, 600–605.

Ramsey, S. (2010). Comparative assessment for medications and devices: Apples and oranges? *Value in Health* **13**(supplement 1), S12–S14.

United States Government Accountability Office (US GAO) (2012). Higher use of advanced imaging services by providers who self-refer costing Medicare millions. *Report to Congressional Requestors, Medicare*. Washington, DC: US GAO. GAO-12-966.

Disability-Adjusted Life Years

JA Salomon, Harvard School of Public Health, Boston, MA, USA

Introduction

The disability-adjusted life year (DALY) is a summary measure of population health that accounts for both mortality and nonfatal health consequences. DALYs were first developed for the primary purpose of quantifying the global burden of disease (GBD). In this context, the DALY was designed as the unit of analysis for measuring the relative magnitude of losses of healthy life associated with different causes of disease and injury. In addition to measurement of the burden of disease, another intended use for DALYs was as a metric for health benefits in the denominator of cost-effectiveness ratios. This article introduces the conceptual and computational basis for DALYs and discusses key issues relating to value choices underlying DALYs, with a brief discussion of how DALYs relate to quality-adjusted life-years (QALYs).

Basic Concepts

A DALY is equivalent to one lost year of healthy life. DALYs accumulate when individuals die prematurely or when they live with the health consequences of diseases, injuries, or risk factors. For a particular cause of disease or injury, DALYs are computed as the sum of (1) 'years of life lost' (YLLs), which capture premature mortality and (2) health losses in 'years lived with disability' (YLDs), which capture lost healthy life due to living in states worse than perfect health. The following sections elaborate on these two components.

Years of Life Lost

YLLs are measures of health losses due to premature mortality. Calculation of YLLs requires some quantification of how long people 'should' live, that is, a normative target lifespan by which the length of life lost due to each death at a certain age may be evaluated. For example, the normative target might imply that a person who is aged 60 years should expect to live another 10 years, that is, until the age of 70 years. In that case, 100 deaths among people at the age of 60 years translates to $100 \times 10 = 1000$ YLLs. There are various possible ways to define a normative target lifespan, as described by Christopher Murray in his 1996 essay entitled 'Rethinking DALYs.' Some norms imply a target lifespan that is constant across ages at death, whereas others imply a target lifespan that shifts depending on the age that has been attained. The latter are typically based on life tables that give expectations of life at different ages. In the GBD study, a global standard life table has been used based on an egalitarian argument for valuing a death at a particular age as the same loss irrespective of where the person lived. Another choice that is made is whether the same life table is used for males and females. In the GBD for the year 1990 and revisions through 2008, two different life tables were used, with the standard for females based on a life expectancy at birth that was 2.5 years greater than the life expectancy at birth in the standard for males. The argument for the different standards was based on a plausible biological difference in longevity. In the revision of the GBD for the year 2010 (hereafter 'GBD 2010'), a new standard was used, based on a synthetic life table constructed from the lowest currently observed mortality rates at each age. The other change in the GBD 2010 was that a single standard life table was defined for both males and females.

Years Lived with Disability

YLDs may be understood conceptually as partial losses of healthy years due to living in health states that are worse than optimal health, weighted for the severity of the states. For example, 10 years lived in a health state that constitutes a 50% reduction in health, i.e., a state that resides halfway between death and perfect health, would imply a total of $10 \times 0.5 = 5$ YLDs. Construction of YLDs requires a defined measurement construct for health losses, a way to quantify these losses, and an approach to attribute losses to years of life lived with a particular condition.

Cases and sequelae

The GBD maps losses of health due to disease and injury through the concepts of cases and sequelae. For cases of a given disease or injury in the population, the experience of health until remission or death will include an array of different health states. For the sake of parsimony, burden of disease calculations require that this multitude of health states be approximated by a small number of discrete entities characterized under the umbrella term of sequela. The sequela is the unit of analysis for epidemiological estimates and YLD calculations. In the GBD, health states are defined by levels of functioning within a set of health domains, for example, mobility, pain, vision, or cognition. These health states are not defined in reference to general well-being or 'quality of life' (both broader constructs). Nor do the health states refer to aspects of participation in society, although different levels of functioning in domains of health may clearly affect – and be affected by – these other aspects.

Incidence and prevalence

YLDs may be computed based on either an incidence or a prevalence perspective. In an incidence perspective, the YLDs associated with a particular sequela are computed in terms of the number of incident cases of the sequela, times the average duration of time spent in the sequela, times a disability weight reflecting the magnitude of health loss experienced for each unit of time lived with the sequela. (Disability weights are discussed further in the next section.) For example, if there were 100 new cases of blindness in a population, and each case of blindness had an average duration of 20 years and an average disability weight of 0.25, then the YLDs due to blindness computed from

 doi:10.1016/B978-0-12-375678-7.00511-3

an incidence perspective would be $100 \times 20 \times 0.25 = 500$. From a prevalence perspective, the calculation is simply the prevalence of a sequela at a defined point in time (e.g., the midpoint in the year of interest), multiplied by the disability weight. For example, if in a population there were 1000 people living with asthma this year, and asthma had a disability weight of 0.10, then the YLDs due to asthma computed from a prevalence perspective would be $1000 \times 0.10 = 100$.

Disability weights

Disability weights provide the bridge between information on mortality and information on nonfatal outcomes in DALYs. These weights represent cardinal measures of health decrements on a scale ranging from 0 (signifying conditions that are equivalent to ideal health) to 1 (signifying conditions that are equivalent to being dead). Thus, for example, if a year lived with deafness has a disability weight of 0.25, this implies that 4 years lived in deafness would be an equivalent health loss to dying 1 year 'too early' in reference to some defined target for longevity (i.e., $4 \times 0.25 = 1$).

Disability weights are needed for every sequela that is included in the study. For most sequelae, a single disability weight is applied to time spent in that sequela under the simplifying assumption of an approximately constant, homogeneous health experience for those living with the sequela over its specified, average duration (taking an incidence perspective). Within this framework it is important to recognize that an individual may have more than one disabling sequela at the same time. The disability weight refers to the average health loss for individuals with a particular condition in the absence of other comorbidities. Without adjustment for comorbidities, the implicit assumption is that multiple sequelae in the same person combine additively, which may not accurately describe the real effects of comorbidity on functional health. Some researchers have suggested various alternative approaches to account for the presence of multiple sequelae, other than assuming additivity.

Assignment of disability weights to the range of sequelae in the first iteration of the GBD 1990, undertaken during the early 1990s, was based on first defining six different disability classes, and then mapping from each sequela into the class or classes that applied to incident cases of that sequela. The six disability classes were defined in reference to limitations in activities of daily living such as eating and personal hygiene; instrumental activities of daily living such as meal preparation; and four other domains (procreation, occupation, education, and recreation). Weights were assigned to the different classes by a panel of public health experts using a rating scale approach. Once the weights attached to each of the six classes were determined (by averaging the values from the expert panel), the disability weight for a particular sequela was estimated by (1) specifying a distribution of incident cases across the different classes – reflecting either the proportion of time an average incident case would spend in different disability classes, or the proportion of incident cases that would be characterized by different severity levels and (2) computing the average weight across this distribution.

For the revision of the GBD 1990 that was completed in 1996, a new approach to estimating disability weights was devised based on two variants of the person trade-off (PTO) method. The revision of the approach was inspired by some specific criticisms of the original approach: (1) that the disability classes were appropriate only for adults (because, e.g., children were naturally dependent on adults for some of the referenced activities); (2) that no formal, replicable protocol was available to guide those aspiring to undertake a national burden of disease exercise; (3) that the class with the lowest level of disability was valued at 0.096, which produced a scale that was too blunt to capture very mild conditions; and (4) that the valuation task itself did not allow the expert panelists to reflect on the policy implications of their values.

New disability weights in the 1996 exercise were elicited from a panel of health professionals following an explicit protocol. In the protocol, a series of 22 indicator conditions were evaluated through an intensive group exercise involving two variants of the PTO and incorporating a deliberative process to encourage reflection on the values that emerged during the exercise. The first type of PTO question asked participants to trade off life extension in a population of healthy individuals versus life extension in a population of individuals having a particular condition. The second type of PTO question asked participants to trade off life extension for healthy individuals versus health improvements in individuals with the reference condition. Participants were required to resolve inconsistencies in the numerical weights implied by the two alternative framings of the PTO. The final consistent values implied by the reconciled PTO responses, averaged across participants, defined the disability weights for the 22 indicator conditions, which were then clustered into seven different classes of severity. As each class contained several of the indicator conditions, these indicators thereby supplied an intuitive and easy-to-convey operational definition of the severity of each class (see **Box 1**).

To generate disability weights for the remainder of the disabling sequelae in the study, participants were asked to estimate distributions across the seven classes for each sequela. In this second stage, the indicator conditions in each class were used as 'pegs' on the scale from perfect health to conditions equivalent to being dead to guide estimation of the distribution across the seven classes of disability. As described above for the first iteration of GBD 1990, the distributions across classes were intended to reflect either the proportion of time a typical case for a given sequela would spend in each class or the percentage of cases that would be categorized in each of the different classes. Distributions across disability classes were estimated separately for treated and untreated cases where relevant, and weights could also vary by age group. The box below presents a few examples of disability weights for common causes.

Various critiques have challenged aspects of the 1996 disability weight measurement exercise. For instance, several critics have questioned the use of healthcare professionals as respondents and suggested that there might be cross-cultural variation in disability weights that should be evaluated. In 1999, Trude Arnesen and Erik Nord argued that there was a serious ethical problem with the first variant of the PTO question used and a logical problem with the requirement that there should be numerical consistency between responses to the two different variants, given that these addressed two different issues. These critiques notwithstanding, the disability weights used for updates of the GBD undertaken through

Box 1 Disability weights in the Global Burden of Disease study, 1996 revision

Based on a deliberative protocol built around the PTO method, disability weights for 22 indicator conditions were estimated, and the conditions were then grouped into seven different classes reflecting a spectrum of severity levels:

- Class 1, with weights ranging from 0.00 to 0.02, included vitiligo on face; and weight-for-height 2 SDs or more below the reference median
- Class 2, with weights ranging from 0.02 to 0.12, included watery diarrhea, severe sore throat, and severe anemia
- Class 3, with weights ranging from 0.12 to 0.24, included radius fracture in a stiff cast; infertility; erectile dysfunction; rheumatoid arthritis; and agina
- Class 4, with weights ranging from 0.24 to 0.36, included below-the-knee amputation; and deafness
- Class 5, with weights ranging from 0.36 to 0.50, included rectovaginal fistula; mild mental retardation; and Down syndrome
- Class 6, with weights ranging from 0.50 to 0.70, included unipolar major depression; blindness; and paraplegia
- Class 7, with weights ranging from 0.70 to 1.00, included active psychosis; dementia; severe migraine; and quadriplegia

Weights for all the full range of sequelae in the study were estimated by defining the distribution of incident cases across these seven classes, using the indicator conditions in each class as illustrative benchmarks. Examples of the resulting weights include:

- Episodes of otitis media: 0.02
- Cases of asthma: 0.10 (untreated); 0.06 (treated)
- Episodes of malaria: 0.21 (ages 0–4 years); 0.17 (ages 15 years and older)
- Rheumatoid arthritis cases: 0.23 (untreated); 0.17 (treated)
- Episodes of meningitis: 0.62
- Terminal cancer: 0.81

Source: Reproduced with permission from Murray, C. J. L. (1996), Rethinking DALYs. In Murray, C. J. L. and Lopez, A. D. (eds.) *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, pp 1–98. Boston: Harvard School of Public Health.

2008 were still largely based on the GBD disability weights as measured in the 1996 exercise. For certain conditions, where weights were not available from the original GBD Study, provisional weights were used from the Dutch Disability Weights study or from the Australian Burden of Disease study. The Dutch Disability Weights study used a similar protocol to the GBD 1996 revision, with the addition of health state distributions for sequelae described in terms of a variant of the EQ-5D classification system.

More recently, prompted by a more general research agenda on developing internationally comparable summary measures of population health at the World Health Organization, the use of the PTO as the basis for disability weights in DALYs has been reconsidered. The most recent thinking on DALYs reflects an effort to more precisely delineate the concept embodied in the nonfatal component of the measure, which has led to the explicit definition of disability weights as measures of overall levels of health associated with health states rather than as measures of the utility associated with these states or the contribution of health to overall welfare. Although some have argued that the burden of disease must be quantified in terms of overall welfare loss because health and well-being are not separable, others have challenged this view, and this debate goes on. In the GBD 2010, a large empirical exercise to measure disability weights has been conducted using household surveys in five countries and an open-access internet survey. This study uses a much simpler method for eliciting weights, based on simple paired comparisons of sequelae described with brief labels. A number of the new weights are lower than the previous ones, including weights on sensory impairments, infertility, and intellectual disability. Other weights are higher in the new study, including weights for some states relating to epilepsy, illicit drug use disorders, and low-back pain. Another significant finding in the new study is that responses to comparisons of health states are remarkably consistent across the diverse sampled populations, which contradicts the prevailing hypothesis that assessments of disability must vary widely across cultures.

Other Value Choices Relevant to Both Years of Life Lost and Years Lived with Disability

Discounting

Many of the arguments around discounting invoked in the context of QALY measures have also been rehearsed in the discussion of DALYs as population health measures. Until recently, the use of an annual discount rate of 3% has been the default standard in the construction of the DALY, as in the recommended base case analysis for cost-effectiveness studies; in both cases it is typically advised that alternatives should be considered in sensitivity analyses. For the GBD 2010, a simpler variant of DALYs has been adopted for the base case, with no discounting.

Age Weights

In addition to discounting, some have argued for assigning unequal weights to life years lived at different ages, and the standard DALY prior to the GBD 2010 included weights that give the highest values to years lived in young adulthood. A range of arguments have been considered in relation to age weighting, with reference to empirical findings on weights that people attach to years over the life course and to important ethical considerations. The developers of the DALY measure previously argued for unequal age weighting based on the social roles played at different ages, but age weights remain controversial. For the GBD 2010, the base case DALYs are not differentially weighted by age.

Applications

DALYs have been used for both quantifying the burden of disease and as the unit of effectiveness in the denominator of cost-effectiveness ratios for economic evaluation of health interventions and programs. The major debut of the DALY in the World Bank's *World Development Report 1993* introduced applications of the measure toward both ends. Various revisions of the GBD have continued to use DALYs as the main unit of account for assessing the relative magnitude of health

losses associated with various diseases, injuries, and risk factors, with the latest revision (GBD 2010) introducing some changes in the specific value choices reflected in the construction of DALYs for base-case analyses, as described above. For use in cost-effectiveness analyses, guidelines from the World Health Organization on conducting 'generalized cost-effectiveness analyses' – with a particular focus on health policies in developing countries – have included an explicit recommendation to use DALYs as the measure of benefit in these analyses.

DALYs and QALYs

DALYs are closely related in concept to QALYs. Both are metrics that take healthy time as the unit of account. Both attach weights to the continuum of health outcomes residing between optimal health and death. An important distinction is in the intended uses of the two metrics. As noted above, DALYs are used both as summary health measures for purposes of descriptive epidemiology, i.e., as units for measuring burden of disease, and as measures of the health benefits of interventions, for example, in cost-effectiveness analyses. QALYs are used primarily for the latter purpose, but there have been assessments of a related measure called 'quality-adjusted life expectancy' as a measure of the overall average level of health in a population. The construction of summary measures of population health has much in common with the construction of measures of the benefits from health interventions, so the distinction is unimportant when considering many of the features of the measures.

Christopher Murray and Arnab Acharya, in their 1997 essay on DALYs, characterized the relationship between DALYs and QALYs as follows: "DALYs can be considered as a variant of QALYs which have been standardized for comparative use." There are certain key distinctions worth noting.

As DALYs are negative measures that reflect health losses, the scale used to quantify nonfatal health outcomes in DALYs is inverted compared with the scale used in QALYs; that is, numbers near 0 represent relatively good health levels (or small losses) in DALYs, whereas numbers near 1 represent relatively poor health levels (or large losses). The inverted scale means that interventions that improve health result in DALYs averted, whereas QALYs are gained.

Disability weights in DALYs, which are the health state valuations analogous to the 'quality' adjustments in QALYs, are intended to reflect the degree to which health is reduced by the presence of different conditions, whereas at least one interpretation of the weights in QALYs is based on the individual utility derived from different states. The current interpretation of weights in DALYs reflects some evolution over time, as discussed above. Another distinction relating to disability weights is that in the GBD disability weights are assigned to health states that are attached explicitly to the sequelae of specific diseases and injuries, whereas in many applications of QALYs health states are described in terms of concrete symptoms and functional losses, without reference to specific conditions.

The standard formulation of DALYs used in revisions through 2008 has weighted healthy life lived at different ages according to a variable function that peaks at young adult ages, whereas QALYs do not typically incorporate unequal age weights. As noted above, the GBD 2010 revision has moved to using DALYs without age weights.

For measuring the burden of disease, YLLs due to premature mortality at different ages are computed with reference to a standard life table. For purposes of cost-effectiveness, this distinction is largely inconsequential, because the standard life expectancy largely nets out when benefits of interventions are computed as the change in DALYs. As a simplified example, imagine an intervention that defers one death from the age of 50 years to the age of 70 years, and suppose that the normative target lifespan used as the yardstick for DALYs is 80 years (irrespective of one's current age). Then the number of DALYs averted through intervention is a change from $80-50=30$ to $80-70=10$, for a net of 20 DALYs averted, which is the same as the number of QALYs gained through the intervention. (Note that in the actual standard life table that is used, as in most life tables, the target lifespan, equal to the number of years of remaining life expectancy at age x plus x, rises slightly with advancing adult ages rather than remaining constant as per the simple example here. This will produce a slight discrepancy between DALYs averted and QALYs gained, but this difference is usually negligible.)

See also: Multiattribute Utility Instruments: Condition-Specific Versions. Quality-Adjusted Life-Years. Time Preference and Discounting. Valuing Health States, Techniques for

Further Reading

Anand, S. and Hanson, K. (1997). Disability-adjusted life years: A critical review. *Journal of Health Economics* **16**(6), 685–702.

Arnesen, T. and Nord, E. (1999). The value of DALY life problems with ethics and validity of disability adjusted life years. *British Medical Journal* **319**(7222), 1423–1425.

Hausman, D. M. (2012). Health, well-being, and measuring the burden of disease. *Population Health Metrics* **10**(1), article 13.

Murray, C. J. L. (1994). Quantifying the burden of disease: The technical basis for disability-adjusted life years. *Bulletin of the World Health Organization* **72**(3), 429–445.

Murray, C. J. L. (1996). Rethinking DALYs. In Murray, C. J. L. and Lopez, A. D. (eds.) *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, pp 1–98. Boston: Harvard School of Public Health.

Murray, C. J. L. and Acharya, A. K. (1997). Understanding DALYs. *Journal of Health Economics* **16**(6), 703–730.

Murray, C. J. L., Ezzati, M., Flaxman, A. D., et al. (2012). Comprehensive systematic analysis of global epidemiology: definitions, methods, simplification of DALYs, and comparative results from the global burden of disease 2010 study. *Lancet* **380**(9859), pp 2055–2058.

Nord, E., Menzel, P. and Richardson, J. (2006). Multi-method approach to valuing health states: Problems with meaning. *Health Economics* **15**(2), 215–218.

Salomon, J. A. and Murray, C. J. L. (2004). A multi-method approach to measuring health-state valuations. *Health Economics* **13**(3), 281–290.

Salomon, J. A., Vos, T., Hogan, D., et al. (2012). Common values in assessing health outcomes from disease and injury: Disability weights measurement study for the Global Burden of Disease Study 2010. *Lancet* **380**(9859), pp 2129–2143.

Dominance and the Measurement of Inequality

D Madden, University College, Dublin, Ireland

Glossary

Bootstrap A technique for obtaining the sampling distribution of a statistic via resampling with replacement from the original sample.

Cardinalization A situation where a transformation is applied to ordered, categorical data whereby the transformed data can be regarded as cardinal.

Decomposability In the context of a measure of inequality, this is the property whereby if a population can be exclusively and exhaustively separated into a finite number of groups, then overall inequality of the population will exactly equal the sum of within group and between group inequality.

Dominance Dominance, as it is used in this article, refers to a situation whereby health in one population is regarded as superior to health in another population for a wide range of evaluation functions.

Entropy measures of inequality A family of measures of inequality deriving from the degree of order or disorder in a system. Complete equality corresponds to maximum disorder, so the gap between the actual order and maximum disorder is an index of inequality. One of the measures has the property of decomposability.

Gini coefficient A commonly used summary measure of inequality which ranges from zero to one. A value of zero indicates complete equality, whereas a value of one indicates that all health is concentrated in one person.

Jackknife A technique for obtaining the sampling distribution of a statistic via resampling from the original sample but with observations successively deleted.

Kolmogorov–Smirnov test A nonparametric test for the equality of continuous, one-dimensional probability distributions.

Likert scale In a situation where responses to questions come in the form of ordered categories with numbers assigned to them such as '1, 2, 3...' but where no cardinal interpretation can be assigned to the numbers, the Likert scale is obtained by summing the values of the responses. Thus, for example, given 10 questions, where the subject responds with '2' in each case, the Likert score would be 20.

Lorenz curve The Lorenz curve graphs the cumulative proportion of the population (in increasing order of health) on the horizontal axis against the cumulative proportion of total health on the vertical axis. If health is distributed exactly equal, then the Lorenz curve is a 45° line.

Mean/median-preserving spread The situation whereby the degree of variance or spread in a population is increased, whereas the mean/median remains unchanged.

Scale independence The property whereby a summary statistic, such as an inequality measure, is independent of the underlying scale of whatever is being measured. For example, an inequality measure is scale independent if inequality in weight is independent of whether weight is measured in pounds or kilograms.

Stochastic dominance Given an outcome such as health or income, stochastic dominance refers to a situation whereby the probability distribution of the outcome in population A is always ranked higher than the population in population B for all evaluation functions where more is better than less. Higher order definitions of stochastic dominance refer to situations where restrictions such as concavity are imposed upon the evaluation function.

Introduction

This article covers a number of measurement issues which arise in Health Economics. The first of these arises when economists wish to make comparisons between populations on the basis of some measure of health, h, where h_i refers to the value of the health measure for individual i. Such comparisons may be between different populations at the same point in time, or between the same population at different points in time, or indeed a combination of the two. In some cases it may be desirable to compare some measure of central tendency, such as the mean, μ_h.

In some cases however, there may also be of concern about how this health measure is distributed throughout the population. This may arise, for example, because the underlying individual utility function is increasing and concave in the health measure, h_i (presuming for the sake of exposition that a higher value of the health measure increases utility) or it may arise because the ethical views of society are such that society has a degree of 'inequality aversion' with respect to the distribution of this health measure. In the latter instance the inequality aversion of society will be reflected in the way in which individual utility functions are aggregated into some measure of social welfare. In both cases social welfare (defined as some aggregate of individual welfare) will be sensitive to both the level and distribution of h.

In either case, comparison of the health measure will be influenced by the precise utility and/or social welfare function employed, because this will determine the relative importance attached to the average value of the health measure and its distribution. This can be problematic, because the ranking of any two populations may well be sensitive to the choice of specific utility/welfare function. This is where the issue of dominance becomes relevant. Intuitively, a dominance result is obtained if it can be demonstrated that the distribution of health in one population, P will always be ranked better

 doi:10.1016/B978-0-12-375678-7.00725-2

(in terms of conferring higher welfare) that the distribution of health in population Q, for all welfare functions which obey certain broadly agreed upon properties. Dominance results are powerful in that they permit fairly unambiguous comparisons to be made between populations, where the term 'fairly unambiguous' is used in the sense that the ranking between the populations would hold for a wide range of welfare functions. Below, more formal, specific, definitions of dominance are given but for the moment the aforementioned explanation will suffice.

Where dominance is not found, then analysts must rely upon comparisons of some measure of central tendency, usually the mean or the median. If distribution is also an issue they must rely upon the specific utility/welfare function or, if the focus is solely upon distribution then specific inequality measures must be used, in either case running the risk that the ranking of populations may be sensitive to the choice of function/measure. In the case of health however, there may be a further complication. Some health measures are cardinal (e.g., life expectancy) and thus lend themselves to comparison via the mean and also via well-known inequality measures such as the Gini coefficient or coefficient of variation. In many cases however, the health measure is not cardinal but instead is ordinal and categorical, for example, self-assessed health (SAH). In such cases, analysts have essentially two choices: They can either transform their data from ordinal to cardinal, and then proceed using the cardinal approach referred to earlier. Alternatively, they can employ measures which are specifically designed to deal with ordinal data, bearing in mind, however, that there are relatively fewer such measures to choose from than in the case of cardinal data. The case of data which is measured in intervals lies somewhere in-between. Analysts have the choice to convert interval data into cardinal data by assuming that all observations within an interval take the range of, say, the median of that range. Of course, this may be an overly strong assumption to make and ignores any within interval variation, though it may be acceptable if the intervals are comparatively narrow. An alternative would be to convert interval data into cardinal data using the interval regression approach described later.

In this article, the application of dominance methods and the measurement of inequality in health economics, for the case of both cardinal and ordinal data, are reviewed. First, the case where the health measure is cardinal is considered. In the discussion which follows, it can be noted that what could be termed 'pure' health inequality, i.e., inequality in health without reference to an individual's socioeconomic resources will be discussed. This distinguishes this review from the extensive literature on inequality in health outcomes with respect to income or other measures of resources. The article concludes with a brief discussion on statistical inference.

Dominance and Health Inequality with Cardinal Data

In analyzing issues of dominance and inequality in the case where health is measured cardinally, the results and methods employed in the case of income inequality are available for use. It is probably easiest to deal with the case of inequality first. In what follows it is assumed that comparisons between two populations are made with respect to a measure of health h_i where it is assumed that higher values represent better health.

The primary dominance concept in the analysis of inequality is Lorenz dominance. This involves comparison of the Lorenz curve for h_i for the two populations. The Lorenz curve orders individuals in increasing order of h_i and then plots, against the cumulative proportion of the population so ordered, the cumulative proportion of total health going to each proportion of the population. The graph corresponding to the 45° line represents complete equality – everyone has the same health. The closer the graph is to the 45° line, the more equal are the distributions. Thus if one distribution lies above (nearer to the 45° line) for all values of p then that distribution is said to Lorenz dominate and would be ranked as more equal by all inequality measures obeying certain basic properties. These properties are anonymity (permutations of health among the population do not matter for overall inequality), population (the measure of inequality is independent of the size of the population), relativity (absolute levels of health do not matter for inequality measures), and transfer (inequality must fall if there is a transfer of a unit of health from a more to a less healthy person).

Where Lorenz dominance is found, the issue of inequality is essentially resolved. However, it is frequently the case that dominance is not found, in which instance specific inequality measures must be used. There is a wide range of such measures. Among the most frequently used are the Gini coefficient, the coefficient of variation, and the entropy family of measures. The Gini coefficient is closely related to the Lorenz curve and can be calculated as the ratio of the area between the Lorenz curve and 45° line of perfect equality to the area of the triangle below the 45° line. A more formal expression for the Gini coefficient is

$$G = \frac{1}{2N^2\mu_h}\sum_{j=1}^{N}\sum_{k=1}^{N}|h_j - h_k|,$$

i.e., the sum of all the differences between all pairs of health normalized by dividing by the squared population, where N is the total population and μ_h is the mean of the population health.

The coefficient of variation can be obtained from the expression

$$C = \frac{1}{\mu_h}\sqrt{\sum_{i=1}^{n}(h_i - \mu_h)^2},$$

i.e., the standard deviation of health divided by mean health.

The entropy family of inequality indices are given by

$$GE(a) = \frac{1}{a(a-1)}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{h_i}{\mu_h}\right)^a - 1\right]$$

where notation is as before and the parameter a reflects the weight attached to inequality at different parts of the distribution. More negative values of a reflect a higher weight on the lower part of the distribution, whereas higher positive values reflect a greater weight on the upper part of the distribution.

A further additional property which may be desirable in an inequality measure is that of decomposability. Suppose the

population can be clearly partitioned into groups, for example, by region, then the overall inequality index can be decomposed into inequalities within regions and inequalities between the regions. The only commonly used inequality index which can be exactly decomposed (in the sense that the sum of the within group inequalities and between group inequalities exactly add up to overall inequality with no residual) is the Theil index (one of the entropy family above, with $a=1$).

Lorenz dominance is concerned with comparing health in two populations purely on the basis of inequality, without any reference to the average level of health. From a social welfare perspective, greater inequality of health may be a trade off for a higher average level. To take an extreme example, suppose in population Q, there is complete equality of health, whereas in population P, there is a high degree of inequality, yet the least healthy person in P has higher health than the average level in Q. Many would regard P as having superior health to Q even though Q Lorenz dominates P.

In these instances, stochastic dominance results can be applied. The degree of stochastic dominance will depend upon whether the data are cardinal or ordinal and also on the nature of the underlying utility function. For example, with first-order stochastic dominance, suppose that the cumulative distributions of health in populations P and Q are given by $F_p(h)$ and $F_Q(h)$, respectively. Then distribution P dominates distribution Q if for any value of h, $F_Q(h) \geq F_P(h)$, i.e., for any value of health, h, the fraction of population with health lower than h is less in P than in Q. Alternatively, suppose there is a monotone nondecreasing function of h, $u(h)$, then P dominates Q if $\int u(h)dF_P \geq \int u(h)dF_Q$ for all values of h. In this case, $u(h)$ can be regarded as a utility function which is monotonically increasing in health.

Thus if it is simply assumed that individual utility is increasing in health, then dominance for population P over population Q holds if the cumulative distribution of health for population P first-order stochastically dominates that for population Q.

Assuming that individual utility functions are not only increasing, but are also concave in the measure of health, then provided the health measure is cardinal, dominance may also be observed if the cumulative distribution of population P second-order stochastically dominates that of population Q. Thus, $\int u(h)dF_P \geq \int u(h)dF_Q$ and now $u(h)$ is monotone increasing and concave. In terms of the comparison of cumulative distribution functions, what is now important is the area under the distribution functions. Thus P will second-order stochastically dominate Q if $D_Q(h) = \int F_Q(h)dh \geq \int F_P(h)dh = D_P(h)$. Note that comparison of the areas under the distribution function implies that h, the argument of the distribution function, can be summed in a meaningful way. This implies that second- and higher order stochastic dominance is only meaningful if h is cardinal and cannot be applied if h is ordinal. It is also worth noting that in this case second-order stochastic dominance is equivalent to what is known as Generalized Lorenz dominance, where the Generalized Lorenz curve is simply the original Lorenz curve scaled up by the average level of health.

There is one further branch of dominance theory which is of relevance for comparison of some specific health measures between populations. In some cases, it would be of concern if the value of a specific health measure lies above (or below) a critical threshold, although at the same time, it may not be of concern should the value of the health measure be below (above) that threshold. This has clear parallels with the study of poverty and dominance results from the poverty literature can be applied in these cases. One obvious area within health economics where such techniques could be applied is obesity, with its focus on individuals whose body mass index (BMI) lies above a critical threshold. This approach is particularly useful when there may not be complete agreement over where the critical threshold should be drawn. A further example of an application of this technique in health economics is with regard to mental stress (Madden, 2009). Here mental stress is measured via a Likert scale derived from answers to the General Health Questionnaire (GHQ) and once again the threshold value of the scale which indicates mental stress is open to question. Stochastic dominance techniques are used here to show that regardless of where the threshold is drawn, there was a fall in mental stress in Ireland over the 1994–2000 period.

The analysis of mental stress in Ireland referred to earlier essentially interpreted the Likert scale derived from the GHQ as a cardinal measure of mental health. Strictly speaking this is not true as the underlying data used to construct the scale are of an ordinal categorical nature. Much health data, including the frequently encountered SAH measures, are of this nature and the application of dominance techniques and the calculation of inequality in these instances raise particular questions, to which we now turn.

Dominance and Inequality with Ordinal Data

Whereas there are some health measures which are cardinal, they tend to concentrate only on specific dimensions of health, for example, BMI. More general cardinal health measures are comparatively difficult to come across. Measures such as the SF-36 or Euroqol are available only for a limited range of countries. Probably the most frequently employed measure of general health is SAH. Individuals answer a question of the form: In general, how good would you say your health is? The possible answers are: very bad, bad, fair, good, and very good (the exact wording can differ from survey to survey but it is generally of the aforementioned type). Whereas this measure appears to be a good indicator of overall health it is not cardinal, and with only five categories, it is not suited to the application of the standard inequality indices referred to earlier.

The breakthrough in analyzing inequality with such data came from Allinson and Foster (2004). They showed how standard measures of the spread of a distribution which use the mean as a reference point, such as the Gini, are inappropriate when dealing with categorical data. This is because the inequality ordering will not be independent of the (arbitrarily chosen) scale applied to the different categories. In this instance a more appropriate reference point is the median category and the cumulative proportions of the population in each category is the foundation of their analysis of inequality with categorical data. This is because, whereas changes in the

scale used will affect the width of the steps of the cumulative distribution, the height of the cumulative distribution is invariant to the choice of scale, thus providing the crucial property of scale independence.

Allison and Foster (2004) thus develop a partial ordering based on a median-preserving spread of the distribution (analogous to the partial ordering based on a mean preserving spread provided by say a Lorenz comparison). Thus, suppose a measure of SAH with n different categories which can be clearly ordered $1, \ldots, n$. Let m denote the median category and let P and Q denote two cumulative distributions of SAH with P_i and Q_i indicating the *cumulative* proportion of the population in category i, in each distribution, where $i = 1, \ldots, n$. For the case where both P and Q have identical median states m then P has less inequality than Q if for all categories $j < m$, $P_j \leq Q_j$ and for all $j \geq m$, $P_j \geq Q_j$. What this is effectively saying is that distribution Q could be obtained from distribution P via a sequence of median-preserving spreads.

Allison and Foster also deal with dominance when the focus is on the level of the health measure. In this case distribution P will dominate distribution Q if the cumulative frequency at each point on the ordinal scale (as we go from lower to higher) is always higher in Q than in P. This is equivalent to the first-order stochastic dominance condition referred to earlier. For a recent example of an application of this approach to a comparison of SAH between different social classes, see Dias (2009). As pointed out earlier, it is important to note that when data are ordinal then second-order stochastic dominance is not defined, because it requires that the health measure h can be summed in a meaningful way.

Of course, the Allison–Foster measure shares with Lorenz dominance the property that it only provides a partial ordering and there may be instances when the aforementioned conditions do not hold and it is not possible to rank different distributions of categorical data. Abul Naga and Yalcin (2008) address this issue and build upon the Allison–Foster approach in presenting a parametric family of inequality indices for qualitative data. Like its cardinal data counterparts such as the Gini coefficient or coefficient of variation, it will always provide a ranking, but it lacks the generality of the dominance approach. Subsequent to the Abul Naga–Yalcin paper, Lazar and Silber (2011) have provided an alternative index for ordinal data building upon work in the area of ordinal segregation. The Abul Naga–Yalcin work has also been extended to provide an index which can be used to make comparisons when the two distributions in question do not have the same median category.

The aforementioned contributions show that real progress has been made toward measuring inequality in the case of ordinal data. However, at this stage, in the literature there is still only a limited number of indices specifically designed for ordinal data, so, unlike the case with cardinal data, the analyst has less opportunity to check the sensitivity of results to alternative indices. In this instance, there is another approach which can be taken. It is possible to transform ordinal data to cardinal data, and then apply the cardinal indices referred to earlier.

Much of the literature in this area developed in the context of measuring health inequality related to socioeconomic resources and a very useful summary is available in van Doorslaer and Jones (2003). Their favored approach is to use interval regression to obtain a mapping from the empirical distribution function of what is regarded as a valid index of health (such as the McMaster Health Utility Index (HUI)) to SAH. By mapping from the cumulative frequencies of SAH categories into an index of health such as the McMaster HUI it is possible to obtain upper and lower limits of the intervals for the SAH categories. These can then be used in an interval regression to obtain a predicted value of the index for all individuals. Comparisons which they carry out for measures of SAH in Canada suggest that this approach to cardinalization outperforms other approaches and it also appears to be the case that the values of the health index obtained are not very sensitive to the cutoff points chosen. Hence it may be regarded as acceptable to use cutoff points from the Canadian HUI to calculate a cardinal index of health for other countries.

A key question then is, how do the results obtained from such an approach compare with those from an index specifically designed to deal with ordinal data? Madden (2010) carried out such an exercise, calculating ordinal inequality indices using the Abul Naga–Yalcin approach and also cardinal indices using generalized entropy measures and applying them to Irish data for the years 2003–06. In terms of the ranking of the different years there was very little correlation between the ordinal and cardinal indices. This is a specific result obtained with a specific dataset but it underlines that the choice between the application of an ordinal index versus transforming data into cardinal format and then using a cardinal index may not be trivial.

Statistical Inference

Sections 'Dominance and Health Inequality with Cardinal Data' and 'Dominance and Inequality with Ordinal Data' outlined approaches for the testing of dominance and measuring inequality, using both cardinal and ordinal data. Should dominance be found then of course, it is necessary to check if such a finding is statistically significant. Similarly, it may be used for the calculation of the standard errors associated with any particular index of inequality calculated.

Dealing first with dominance in the case of cardinal data, in the case of inequality alone this issue boils down to checking for statistically significant differences between the ordinates of the Lorenz curves. Suppose that L_i is the ith Lorenz ordinate ($i = 1, 2, ..k$), where the kth ordinate is equal to one. Then, as shown in Beach and Davidson (1983), given estimated Lorenz ordinates from two populations P and Q with sample sizes N_P and N_Q respectively, there are $k-1$ pairwise tests of sample Lorenz ordinates:

$$T_i = \frac{\hat{L}_i^P - \hat{L}_i^Q}{\sqrt{\frac{\hat{V}_i^P}{N_P} + \frac{\hat{V}_i^Q}{N_Q}}}, \quad i = 1, 2, ..k-1 \tag{1}$$

In large samples, T_i is asymptotically normally distributed. Bishop et al. (1991) suggest the following criteria when testing for Lorenz dominance: If there is at least one positive significant difference and no negative significant differences between Lorenz ordinates then dominance holds. Two distributions are

ranked as equivalent if there are no significant differences, whereas the curves cross if the difference in at least one set of ordinates is positive and significant although at least one other set is negative and significant.

In the case of first- and second-order stochastic dominance for cardinal data, Kolmogorov–Smirnov tests can be applied. Such tests can also be applied to ordinal data for first-order stochastic dominance.

If Lorenz dominance is not found then individual inequality indices must be calculated and the appropriate standard error obtained. Obtaining analytic expressions for standard errors in the case of many inequality indices is far from easy as the expressions may be highly nonlinear and whereas asymptotic results may exist, robust, small-sample results are more difficult to obtain. Given this problem, the bootstrap approach may be preferable, as evidence suggests that bootstrap tests perform reasonably well in these situations (see Biewen, 2002).

In the case of the ordinal inequality index developed by Abul Naga and Yalcin to date there has been no progress in terms of statistical inference for this index. However, the Lazar–Silber paper provides jackknifed standard errors and it is also worth observing that in the related literature of the measurement of polarization for ordinal data expressions for the calculation of confidence intervals for such measures have been produced.

Conclusion

This article has summarized some of the main results with respect to dominance and inequality in the case of health data. It was seen that a crucial distinction must be made between cardinal and ordinal health measures. In general the literature for cardinal health measures is more developed, in terms of dominance, indices, and statistical inference. There have also been developments in the analysis of dominance in more than one dimension. The area of multidimensional dominance raises important issues for the measurement of population health which are currently being vigorously debated in the poverty literature. One of the principal issues to be resolved is whether aggregation of different dimensions of health should take place before dominance or inequality analysis is applied (e.g., if single-dimensioned dominance/inequality analysis were to be applied to an aggregate cardinal health measure such as the SF-36), or whether alternatively an explicitly multidimensional approach is adopted whereby analysis is applied to separate dimensions of health and aggregation which takes place at the level of the inequality index itself.

For the case of ordinal health measures, which are arguably more widely employed, dominance results are generally less applicable, there are fewer inequality indices and statistical inference is less well developed. In this area, future developments are perhaps most likely to involve further contributions along the lines of Lazar and Silber (2011) with the development of a wider menu of inequality indices. It is also to be expected that further progress will be made in the area of statistical inference.

See also: Efficiency and Equity in Health: Philosophical Considerations. Equality of Opportunity in Health. Health Econometrics: Overview. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach. Unfair Health Inequality

References

Abul Naga, R. H. and Yalcin, T. (2008). Inequality measurement for ordered response health data. *Journal of Health Economics* **27**, 1614–1625.
Allison, R. A. and Foster, J. (2004). Measuring health inequality using qualitative data. *Journal of Health Economics* **23**, 505–524.
Beach, C. and Davidson, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies* **50**, 723–735.
Biewen, M. (2002). Bootstrap inference for inequality, poverty and mobility measurement. *Journal of Econometrics* **108**, 317–342.
Bishop, J. A., Formby, J. and Smith, W. J. (1991). Lorenz dominance and welfare: Changes in the US. distribution of income, 1967–1986. *Review of Economics and Statistics* **73**, 134–139.
Dias, P. R. (2009). Inequality of opportunity in health: Evidence from a UK Cohort Study. *Health Economics* **18**, 1057–1074.
Lazar, A. and Silber, J. (2011). On the cardinal measurement of health inequality when only ordinal information is available on individual health status. *Health Economics* **22**, 106–113.
Madden, D. (2009). Mental stress in Ireland, 1994–2000: A stochastic dominance approach. *Health Economics* **18**, 1202–1217.
Madden, D. (2010). Ordinal and cardinal measures of health inequality: An empirical comparison. *Health Economics* **19**, 243–250.

Further Reading

Abul Naga, R. H. and Yalcin, T. (2010). Median independent inequality orderings. *University of Aberdeen Business School Working Paper Series*, vol. 03, pp 1–25, Aberdeen: University of Aberdeen Business School.
Apouey, B. (2007). Measuring health polarisation using self-assessed data. *Health Economics* **16**, 875–894.
Atkinson, A. (1987). On the measurement of poverty. *Econometrica* **55**, 749–764.
Kakwani, N., Wagstaff, A. and Van Doorlsaer, E. (1997). Socioeconomic inequalities in health: Measurement, computation and statistical inference. *Journal of Econometrics* **77**, 87–103.
Madden, D. (2012). A profile of obesity in Ireland, 2002–2007. *Journal of the Royal Statistical Society A* **175**, 893–914.
Van Doorslaer, E. and Jones, A. (2003). Inequality in self-reported health: Validation of a new approach to measurement. *Journal of Health Economics* **22**, 61–87.

Dynamic Models: Econometric Considerations of Time

D Gilleskie, University of North Carolina, Chapel Hill, NC, USA

Introduction

Empirical health economists are likely to encounter questions regarding health and health behaviors that involve dynamics. What does one mean by dynamics? Put simply, a dynamic model of economic behavior captures the element of 'time.' In contrast, a static model leaves out time. More specifically, intertemporal dependence is made explicit in dynamic models. This article discusses some of the econometric methods used to estimate dynamic empirical models.

To motivate these methods, the article begins with three examples of individual behaviors studied by health economists that exhibit meaningful relationships across time. Building on these examples, the article presents the econometric methods that have been used by health economists to measure dynamic relationships or behaviors that are connected over time. Much of this work attempts to recover causal effects of variables on outcomes (as opposed to mere statistical correlations) in a dynamic empirical setting. The article concludes with a description of solution and estimation of optimization problems that recover the underlying 'primitive' (or structural) parameters that characterize how economists model dynamic decision making.

Examples of Dynamics in Health and Health-Related Behaviors

As demonstrated in each example in this section, behaviors that are dynamic involve an event, outcome, or action in the past that affects current decisions regarding that behavior or, related, an event, outcome, or action today that impacts future decision making. A distinction needs to be made as to whether the observed event, outcome, or action is endogenous or exogenous. The adjective endogenous implies that the agent (individual or firm) had a role in choosing or influencing the event, outcome, or action; exogenous implies that characteristics of the agent do not affect the event, outcome, or action. Most importantly, endogeneity suggests that unobservable characteristics of the agent likely affect adjacent (in time) behaviors. Another important characteristic of the examples that follow is that the observations of interest are specific to an individual (i) over time (t). We will assume that the econometrician has access to data with repeated observations on the same individuals. We need many observations in the 'individual' dimension, but may have only a small number of observations in the 'time' dimension.

Example 1: Health Production

It is difficult to think about dynamic models in health economics without thinking of Michael Grossman's seminal work on health capital and the demand for health (1972). Grossman was the first economist to formally model the optimal health and medical care consumption of individuals. He recognizes, and emphasizes, that health is inherently dynamic. Indeed, time is such an integral part of health-related decision making that Grossman framed it as an optimization problem being solved over a lifetime.

In this example, the evolution of an individual's health is a dynamic process. Grossman describes our health as a stock, or a capital good. One might compare it to the concept of human capital, or the stock of education that an individual acquires over her lifetime. In fact, similar to education or any capital good (e.g., computer equipment) that a firm uses in production, the stock of health requires maintenance, and depreciates, over time. And, like other capital goods, health can be characterized by its stock and its flow. The stock is an instantaneous measure of current health. The flow is the services or benefits that are generated from a stock of health.

But health cannot be purchased. An individual cannot go to the store to buy more health when her health stock falls below a particular level. Rather, an individual produces health – with the operative words being 'individual' and 'produce.' The first word signifies that each individual is responsible for their own health stock. This responsibility does not imply that accidents can always be avoided. It is quite true that unfortunate events that reduce our health stock happen, and happen with no fault of our own. Yet, our own behaviors can and do influence our health stock.

That influence is exactly where the second word 'produce' comes in. An individual chooses inputs to sustain or augment a health stock. She may also choose inputs that reduce or weaken her health stock. An obvious example of a positive input (i.e., one in which the marginal product is positive) is medical care. If one's health deteriorates, she may consume medical care to repair it. Other examples, of both positive and negative inputs, include food, cigarette, and alcohol consumption. In addition to market goods that affect health, nonmarket goods may impact health. For example, exercise, sleep, and stress are health inputs.

Given this understanding of health capital economists model it as a dynamic productive process with depreciation. That is, let H_{it} define an individual's health stock at time t. Let I_{it} denote a selected amount of health input at time t. Then, according to the description of the evolution of the health stock, a simple mathematical definition of the dynamic process might be

$$H_{i,t+1} = (1 - \delta_t)H_{it} + f(I_{it}) \quad [1]$$

where δ_t measures the depreciation rate of health capital in each period t to $t+1$ and $f(\cdot)$ is a function that converts the period t health input into health next period. Given this definition or model, it should be immediately obvious that the health production process involves time. Health in one period

 doi:10.1016/B978-0-12-375678-7.00721-5

evolves into health in the subsequent period with the help of chosen behavioral inputs during the transition period. Unfortunately, social scientists do not observe an individual's rate of health depreciation nor do they know with certainty the effectiveness of any one health input for a particular individual. Yet, these are exactly the empirical questions health economists seek to answer when studying health and health behaviors.

Example 2: Addictive Good Consumption

Cigarette consumption was described in Example 1 as a (negative) health input. To measure the effect of cigarette consumption on health outcomes, an econometrician must first understand decisions to smoke. (Smoking is an example of an action that is endogenous. Unobservable individual characteristics that affect smoking might also affect the health outcome of interest, which will bias estimates of the smoking effect.) A simple approach to modeling demand for any good is to consider the demand to be static. That is, today's demand for a good depends only on measures of specific variables today: an individual's income today, her current age, the price of the good today, and even the prices of goods that are substitutes for or complements to the good being considered, etc. However, health economists place cigarettes in a different class of goods, namely addictive goods (Chaloupka and Warner, 2000). Other examples of addictive goods might be alcohol, fast food, illicit drugs, etc.

What characterizes an addictive good? The basic premise is that past consumption of the good alters the enjoyment an individual receives from current consumption. That is, the demand for the good today depends not only on measures of specific variables today, but also on measures of specific variables in the past. Herein lies the role of time, or dynamics.

Economists have labeled three ways that past consumption of an addictive good might alter the enjoyment of the good today. The first characteristic of an addictive good is tolerance: The amount of the good consumed in the past directly affects 'happiness' today or, in the economist's terminology, contemporaneous 'utility.' Using fast food consumption as an example of a negative addiction, someone who has consumed a lot of greasy French fries in the past is unhappier (all else equal) than someone who has not developed this addiction. However, a beneficial addiction might be to exercise. Tolerance suggests that someone who has engaged in exercise routinely in the past may experience greater happiness today, all else equal.

The second characteristic of an addictive good is withdrawal: Consumption of the good provides positive utility. Put differently, an individual foregoes this utility if they choose not to smoke. Thus, a smoker, who anticipates the reduction in utility today associated with their past use, can look forward to the boost in utility they receive from continuing to smoke today. If they quit, their utility is lower.

The third characteristic is reinforcement: The marginal utility of consumption of the good today is greater when the person has a history of consumption of the good. This characteristic suggests that consumption of the good in two consecutive time periods is complementary. Adjacent complementarity also implies that reducing consumption of the addictive good (or quitting) is harder the more one has consumed in the recent past.

Each of these characteristics suggests that the demand for an addictive good depends on past behavior related to that good. That is, demand is dynamic. Note, however, that this dependence on past behavior also suggests that behavior today will impact optimal future behavior. That is, an individual deciding whether or not to smoke today (and the amount to smoke) takes into consideration their past smoking behavior, but also understands that their decision today will affect their optimization problem regarding the same behavior tomorrow. Hence, today's decision, which is based on maximizing lifetime utility from this period forward, depends not only on past behavior, but also on expected optimal behavior in the future conditional on today's choice.

Health economists theorize that the demand for an addictive good (under particular assumptions about the optimization problem of the individual) is described by the function:

$$C_{it} = c(C_{i,t-1},\, P_{it},\, C^{*}_{i,t+1},\, X_{it}) \qquad [2]$$

where C_{it} is current consumption of cigarettes, $C_{i,t-1}$ is lagged consumption, and $C^{*}_{i,t+1}$ is expected future consumption. The contemporaneous price of cigarettes is denoted as P_{it} and exogenous individual characteristics are denoted as X_{it}. Given this demand relationship for addictive goods, one can easily see how time plays a role. In particular, consumption of the addictive good in different time periods affects current consumption of the good.

Example 3: Health Insurance Selection

The optimal health insurance decision, or the demand for health insurance by an individual, is another example of a health-related behavior that involves elements of time. More specifically, an individual decides today, without perfect knowledge of her future need for medical care, whether or not to purchase a health insurance plan, which reduces the financial responsibility for medical care consumed in the near future. That is, health insurance is chosen before realization of the health state, and hence, medical care expenses. Put differently, at the point of insurance decision making, medical expenses are uncertain (Arrow, 1963).

A basic, stripped down model of optimal health insurance purchase involves choosing a plan to maximize expected utility. The uncertainty of health, and therefore medical care consumption, requires an individual to forecast – at the time of the insurance decision – the distribution of future medical care expenditure. The decision of an individual to purchase health insurance depends not only on the individual's expected medical care expenditure (i.e., some average measure), but also on the tails of that distribution. How likely am I to experience a disastrous health event that requires high medical care expenditure?

Suppose the set of health insurance alternatives differ by the level of cost-sharing or reimbursement (which ranges from 0% to 100%) and the premium (i.e., the price of the plan). That is, if an individual chooses a 30/70% plan, the insurance

company pays 70% of medical care costs during that insurance year, whereas the individual is responsible for the remaining 30% of total expenses. The premium, or price of insurance, increases with the level of coverage. Assuming utility is a function of wealth, the individual decides on the optimal level of insurance coverage that will maximize her expected utility under uncertain health or medical expense. Specifically, an individual selects insurance as if she were solving the following optimization problem:

$$\text{Max} \int U(w_{it} - D_{it} - \alpha p_{it} + \alpha D_{it})\, \mathrm{d}F(D_{it}) \qquad [3]$$

where w_{it} is the individual's wealth in period t, D_{it} is the unknown medical care expense at t with known distribution, $F(\cdot)$, and p_{it} is the insurance premium per percentage of payout, α, which measures the level of insurance coverage. In this simplified model, the individual maximizes expected utility by choosing the optimal level of coverage α (which is between 0 and 1 inclusive). Mathematically, the solution involves integrating over the medical care expense distribution, $F(\cdot)$, because costs of medical care are not known with certainty. Therefore, the optimal level of coverage depends on initial wealth, the price of insurance, the individual's level of risk aversion (captured here by the shape of the utility function $U(\cdot)$), and finally, the distribution of medical care expenses.

Here, the role of time is a bit more subtle. This example abstracts from the very realistic assumption that previous experience with particular health insurance plans or past medical care utilization may influence the current value of each health insurance alternative; in that case, the role of time mimics previous examples. However, the simple model above highlights a role of time that is different from the previous examples: Optimal decision making requires the individual to forecast future medical care expenses based on current information. Below we discuss how to solve and estimate this dynamic behavior.

The examples presented above relate to individual health behavior. Yet, there are many examples of dynamics on the supply side of health economics. Consider, for example, technology adoption or the decision of a firm to enter or exit the market. How do a firm's actions today affect the likely actions today or tomorrow of its competitors? The health literature examines these dynamic behaviors in hospital entry and exit, medical equipment adoption, and learning by firms or physicians about drug or procedure effectiveness, just to name a few areas.

Econometric Methods to Capture Dynamics

The first two examples describe the production function for health and the demand function for an addictive good. Equations [1] and [2] depict the relationships between the explained variables (i.e., health and consumption of addictive goods) and the explanatory variables. Economic theory guides the inclusion of particular explanatory, or right-hand side, variables. Then, the economist or econometrician can think about estimating an empirical model that captures the dynamic relationship.

Consider the following steps taken by the econometrician.

Step 1: Specify the Econometric Model

Given a theoretical relationship between variables of interest, the first step is to specify an appropriate econometric model. An econometrician might specify the evolution of health capital, depicted theoretically as the health production function in eqn [1], as follows:

$$H_{i,t+1} = \alpha_1 H_{it} + \alpha_2 I_{it} + u_{it} \qquad [4]$$

where α_1 and α_2 are coefficients to be estimated. They reflect the measured effect of the variation in the explanatory variables on variation in the dependent variable. In particular, α_1 measures the depreciation of the health stock between periods t and $t+1$ and α_2 measures the investment in the health stock obtained by consuming an additional unit of the input I_{it}. The u_{it} term measures the unobserved or unexplained variation in health among individuals. It captures the fact that the relationship between $H_{i,t+1}$ and H_{it} and I_{it} is not perfect or, rather, that H_{it} and I_{it} do not fully explain $H_{i,t+1}$. Error terms are always added to statistical models that describe behavior of individuals because we are social beings, and our behavior is often not as completely predictable as it might be for many natural or physical behaviors. To avoid introducing too much additional notation, the notation u_{it} is used to capture unobserved heterogeneity, generally, in all the equations that follow. The reader should understand that the amount of error and, hence, the variable that captures that error, depends on the behaviors being explained as well as the power of the observable explanatory variables.

The health production relationship can be made more realistic (and more complicated) by including additional explanatory variables that make sense theoretically. For example, Grossman himself suggests that education influences the marginal product (or effectiveness) of a health input. Let X_{it} define demographic characteristics of the individual. The econometrician might test the hypothesis that education matters by estimating the following regression:

$$H_{i,t+1} = \alpha_1 H_{it} + \alpha_2 I_{it} + \alpha_3 X_{it} + \alpha_4 I_{it} X_{it} + u_{it} \qquad [5]$$

It should be noted that the relationship in eqns [4] and [5] holds for all time periods. Thus, the model can be rewritten as

$$H_{it} = \alpha_1 H_{i,t-1} + \alpha_2 I_{i,t-1} + \alpha_3 X_{i,t-1} + \alpha_4 I_{i,t-1} X_{i,t-1} + u_{i,t-1} \qquad [6]$$

Turn now to the example of addictive good consumption. The econometrician might specify the equation that depicts the demand function described in eqn [2] as

$$C_{it} = \gamma_0 + \gamma_1 C_{i,t-1} + \gamma_2 P_{it} + \gamma_3 C^*_{i,t+1} + \gamma_4 X_{it} + u_{it} \qquad [7]$$

where the variables have been previously defined, and the marginal effects of these variables are depicted by parameters, γ, to be estimated.

Step 2: Determine the Measurable Variables to be Used in Estimation

Having specified the model, the econometrician has to determine whether data exist to estimate the model as it has been

specified. For example, how should health stock be measured? What variable exists in a data set that best captures a person's stock of health? What inputs affect health? Do measures of those inputs exist? Which inputs need to be modeled because they are chosen by the individual? For now, consider only one health input: medical care. Later, the case where multiple inputs may better explain the evolution of health will be considered.

Estimation of the empirical model in the addictive good example requires that the econometrician observe an individual's consumption and the price of cigarettes over time. In some cases where longitudinal information on individual-level consumption has not been available, econometricians have used their knowledge of the dynamic nature of addictive good consumption and the availability of cigarette prices over time to replace past and expected future consumption with the relevant price information at the time of consumption. Hence, eqn [7] becomes

$$C_{it} = \varphi_0 + \varphi_1 P_{i,t-1} + \varphi_2 P_{it} + \varphi_3 P_{i,t+1} + \varphi_4 X_{it} + u_{it} \qquad [8]$$

One can still see that 'time' plays an integral role in predicting current consumption. Specifically, prices of cigarettes yesterday, today, and tomorrow may affect cigarette consumption today.

Step 3: Evaluate the Role of Unobservables Associated with the Dynamics of the Model

Given a dependent variable (e.g., H_{it} and C_{it} in the two previous examples) and a set of observable explanatory variables, one might initially consider the use any one of the statistical estimators described in previous articles. An ordinary least squares (OLS) regression seems like an obvious candidate. However, the dynamic feature of the equations begs the question: Do unobserved individual differences (heterogeneity) that explain the observed lagged outcome ($H_{i,t-1}$), action ($I_{i,t-1}$), or event ($P_{i,t-1}$) in the past also influence the current outcome (H_{it}) or action (C_{it}) that is being explored? If so, then the error terms that explain the dependent variable and the right-hand side variable are serially correlated. It may also be the case that these unobservables are heteroskedastic (i.e., the amount of variation in the error differs by observable characteristics), but the focus in this article is on the intertemporal dependence in behaviors, outcomes, and events over time.

First, consider the desire to estimate the effect of an outcome in $t-1$ ($H_{i,t-1}$) on the same outcome in time t (H_{it}). How might the same unobservable affect both the variables H_{it} and $H_{i,t-1}$? One example must be that of unobserved health. Recall that one of the data questions above was about the measurement of health. It is hard to think of, let alone find in an available data set on individuals, a variable that fully captures one's health stock. Thus, unobserved measures of health (that are correlated with the observable measure) are captured by the error term, and these unobserved measures of health are likely correlated over time. This endogeneity produces either upward or downward bias in the OLS estimate of the endogenous variable's effect.

However, an unbiased estimate of depreciation (α_1) can be obtained only by using specific econometric techniques to account for the correlation in the unobservables. Other examples of this correlation in measures of health over time are unobserved family health history, unobserved rates of time preference that capture how forward-looking an individual may be, and unobserved health inputs. One might encounter the same problem when trying to estimate the effect of an action in $t-1$ ($C_{i,t-1}$) on the same action in time t (C_{it}). Unobservables that affect cigarette consumption today are likely correlated with consumption yesterday.

Now consider a desire to estimate the effect of an action in the past ($I_{i,t-1}$) on an outcome today (H_{it}). Might $I_{i,t-1}$ be correlated with H_{it} through unobservables that affect both? Unobserved health, for example, may be correlated with both medical care decisions and observed health outcomes. That is, lagged observed health $H_{i,t-1}$, and thus also lagged unobserved health, may affect both one's input decisions last period, $I_{i,t-1}$, and current realizations of health, H_{it}. If current unobservables 'move with' those past unobservables (i.e., are correlated), then estimated coefficients that measure the effect of these lagged observable variables are contaminated with endogeneity bias (if the econometrician does not address the correlation).

To illustrate more fully, decompose the period t error term (u_{it}) into three components. We want the first component to capture permanent unobserved differences in individuals. We label this permanent heterogeneity μ_i. These unobserved individual differences do not vary over time, but may affect observed actions or outcomes in each period. Examples of this type of heterogeneity include risk aversion, genetic characteristics, rates of time preference, or other aspects of the production process or decision-making process that remain fixed over the life cycle. The second component captures time-varying unobserved characteristics of individuals that might be correlated with the explanatory and to-be-explained variables. We label this time-varying heterogeneity v_{it}. Examples include unobserved health shocks, stress, or behaviors that may differ each period. The third component, ε_{it}, is an identically and independently distributed (iid) unobserved error term that is uncorrelated over time and uncorrelated with each of the explanatory variables of the equation. This last component does not cause any problems econometrically. The first and second must be dealt with appropriately. More formally, the general error term can be decomposed into three components:

$$u_{it} = \mu_i + v_{it} + \varepsilon_{it} \quad \text{for all } t \qquad [9]$$

As the examples suggest, the unobservables that impact estimation of variable effects may be either permanent (μ_i) or time varying (v_{it}). There are different econometric techniques that can be used to address these unobservables, depending on the type of variation/correlation.

Step 4: Determine the Appropriate Estimation Method

Economists recognize that dynamic relationships often lead to correlation in variables, or their unobserved determinants, over time. Consider, now, four different methods for addressing the econometric problems associated with unobservables that are correlated over time.

Instrumental variables

In the case of cigarette consumption (eqn [7]), economists recognize that unobservables that determine smoking behavior in the last period may affect smoking behavior in this period, which will bias the measured effects of lagged smoking on current smoking if not addressed econometrically. One solution is to find another variable that varies across individuals that explains lagged smoking behavior but, conditional on the observed lagged smoking behavior, does not independently impact current smoking behavior. Economic theory suggests that smoking decisions in each period depend on the price of cigarettes in each period. Hence, without measures of lagged smoking behavior, $C_{i,t-1}$, one can replace the behavior with its determinants, namely the price of cigarettes in the lagged period, $P_{i,t-1}$. Equation [8] above depicts the new equation using this approach.

For this variable to be a valid 'instrument' for lagged smoking behavior, the econometrician has to answer several questions. First, does this variable vary over individuals? Individual-level variation in prices is difficult to find, but price series that differ by county or state exist. And prices often vary over time. So, cigarette prices by state of residence and time of consumption usually provides enough variation to identify the effect of prices on individual behavior. Second, might $P_{i,t-1}$ be correlated with C_t through unobservables that affect both? If the 'price' of cigarettes is measured by any public policy variable such as cigarette taxes or smoking bans in public places, a public finance economist would probably answer this question with a 'yes'. The political process naturally reflects the preferences of the people, and hence these measures of the costs of smoking are correlated with demand behavior. For the purposes of this article, consider variation in the prices of cigarettes (across states and across time) to be exogenously determined. Thus, there is no need here to worry about unobserved correlation between $P_{i,t-1}$ and C_{it}.

Before stating a third question that must be answered to determine the validity of an instrument, reflect on the specification of eqn [8], which suggests that lagged prices, current prices, and future prices each affect current consumption of cigarettes in some way. It is often the case that adjacent measures of price are correlated. Such multicollinearity among variables makes estimation (and interpretation) of this model difficult. And this difficulty suggests the third question: Although the assumptions about exogeneity of cigarette prices might be valid, do individuals know prices with perfect foresight (as eqn [8] implicitly assumes). If not, then future prices, $P_{i,t+1}$, cannot be included in the equation. Rather, expectations of future prices could be included. But how do people form expectations of future prices today? They may use all information available to form an expectation equal to the true expected value of prices (i.e., rational expectations). Or they may have adaptive expectations and predict future prices using current and lagged observed prices of the good, which are already included in the equation. Regardless of one's assumption about the formation of price expectations, the role of lagged price is now twofold: it captures the effect of lagged price on both lagged smoking behavior and the distribution of future cigarette prices. Yet, theoretically, the econometric work was begun with the goal of measuring the habitual or addictive effect of lagged consumption on current consumption, measured by the coefficient on lagged price. Empirically, the revised specification (eqn [8] without the $P_{i,t+1}$ variable) no longer supports that interpretation.

Assuming that the answers to these questions support the use of an exogenous 'event' (i.e., the price of a good), as an appropriate instrument for a lagged endogenous variable, then the econometrician can proceed with estimation. Either replace the endogenous variable with the exogenous one and estimate the current smoking behavior as a function of current and lagged cigarette prices, or estimate the endogenous action (i.e., lagged smoking) as a function of lagged prices, and use the estimated predicted value of lagged smoking in place of the observed lagged smoking variable. This method requires that lagged prices have no independent explanatory power in the current smoking equation conditional on the predicted lagged smoking behavior.

One should note two things about this instrumental variables method. The former approach (i.e., replacement of $C_{i,t-1}$ with $P_{i,t-1}$) eliminates the need for panel data on individual smoking behavior. Of course, longitudinal data on cigarette prices (or taxes, or smoking bans) that vary by state of residence are needed. The second approach, which involves estimation of the lagged smoking behavior, obviously requires longitudinal data on smoking behavior. Note also that there is no need to model the permanent and/or time-varying unobserved differences among individuals with either of these approaches because the correlation is 'dealt with' by replacement of the offending variable with one that is not correlated with the error term.

Fixed effects

An alternative econometric approach is to model explicitly the permanent and time-varying unobserved differences among people that lead to correlation in variables over time. In this case, panel data on individual behaviors or outcomes is required.

First consider the case where the source of correlation across time periods is permanent unobserved individual differences that affect behavior or outcomes in all periods. That is, in the health production example, eqn [5] can be expressed as

$$H_{i,t+1} = \alpha_1 H_{it} + \alpha_2 I_{it} + \alpha_3 X_{it} + \mu_i + \varepsilon_{it} \qquad [10]$$

and in the addictive good consumption example, eqn [7] can be expressed as

$$C_{it} = \gamma_0 + \gamma_1 C_{i,t-1} + \gamma_2 P_{it} + \gamma_3 X_{it} + \mu_i + \varepsilon_{it} \qquad [11]$$

In each example above, μ_i is the permanent unobserved individual heterogeneity and ε_{it} is the serially uncorrelated iid unobserved (error) component. Conditional on μ_i, H_{it}, and I_{it} are uncorrelated with ε_{it} (and $C_{i,t-1}$ is uncorrelated with ε_{it}). Two fixed-effects methods – the within-groups estimator and the first-differencing estimator – eliminate the fixed individual unobserved effect (μ_i) by transforming the estimated equation. The former involves subtracting the mean of each variable over all years from each individual observation in each cross-section. As the mean of the fixed effect is μ_i itself, the permanent heterogeneity is eliminated. Similarly, the latter approach involves first differences ($H_{i,t+1} - H_{it}$ or $C_{it} - C_{i,t-1}$), which eliminates the permanent component.

The latter first-differencing method is used most frequently among health economists. There are tradeoffs in the econometric properties of each of these estimators.

One advantage of the fixed-effects method is that it not only addresses the serial correlation (caused by permanent heterogeneity) that creates the endogeneity bias associated with having the lag of the dependent variable as an explanatory variable, but it also addresses the correlation associated with endogenous behaviors that affect outcomes (i.e., the input behavior, I_{it}, in eqn [10]).

However, the fixed-effects methods have some disadvantages. The approach relies on changes in explanatory variables over time to identify effects of interest, eliminating time-invariant variables (e.g., gender, race) as explanations for observed behaviors. It is less efficient due to a loss in degrees of freedom in estimation. It ignores correlation generated from time-varying unobserved differences across individuals.

Random effects

The econometrician can employ another estimation tool to model the unobserved heterogeneity. Rather than treat the permanent heterogeneity as individual specific, they can treat it as random, with some distribution, and attempt to estimate the effect of explanatory variables on a behavior or outcome of interest while integrating over the distribution of the correlated unobserved heterogeneity. Sometimes the econometrician estimating with random effects will specify (or assume) the distribution of the unobserved heterogeneity. At other times, the distribution of the heterogeneity will be estimated.

An econometric approach that requires no (or few) distributional assumptions on the unobservables is called the discrete factor random-effects (DFRE) estimator. The random-effect specification introduces an unobservable, μ, that takes on the estimated discrete values $\mu_1,\ldots,\mu_k$ (rather than individual specific values indicated by an i subscript in the fixed-effect specification), with estimated probabilities $\varphi_1,\ldots,\varphi_k$, and $\Sigma_k\ \varphi_k=1$. In this case, consumption behavior in periods $t=2,\ldots,T$ are estimated with the dynamic equation:

$$C_{it}=\gamma_0+\gamma_1 C_{i,t-1}+\gamma_2 P_{it}+\gamma_3 X_{it}+\mu+\varepsilon_{it} \qquad [12]$$

and estimation involves integration of the likelihood function over the estimated discrete distribution of μ.

The DFRE procedure also allows for the introduction and estimation of time-varying unobserved heterogeneity (i.e., the v_{it} term in eqn [9]). One simply needs to also estimate the mass points and probabilities of the mass points associated with this type of heterogeneity.

Another advantage of the DFRE approach is the ease with which an econometrician can jointly estimate two (or more) behaviors of interest. Referring to the health production function example, a source of correlation could be between the lagged health outcome and current health outcome, but it could also be between the input behavior and the health outcome. Modeling the latter correlation explicitly requires jointly estimating the input behavior (i.e., medical care consumption, cigarette consumption, etc.) with the health production function. The linear DFRE version of the multiple equation case would also require the estimation of factor loadings on the unobserved heterogeneity components in each equation to capture different effects of the heterogeneity on different outcomes. There is a nonlinear approach where the joint probabilities of each of the two types of heterogeneity are modeled and estimated.

Note that in the jointly estimated set of equations, as in the health production function example where both the input behavior and the subsequent health outcome are modeled jointly, identification requires that there exists a variable that impacts input behavior but, conditional on the input, does not also affect health outcomes. Theory suggests such variables. For example, if medical care is the only input to health production, prices of medical care (captured perhaps by health insurance cost-sharing characteristics, distance to the physicians office or hospital, supply of doctors, etc.) affect demand for medical care, but do not independently affect health transitions conditional on medical care consumption.

However, it cannot be denied that health is a function of more than medical care inputs. As stated earlier, health depends on different types of medical care inputs (e.g., preventive care, curative care) and nonmedical care inputs (e.g., cigarette consumption, alcohol consumption, physical exercise, nutrition, sleep, stress, etc.). If any of the omitted inputs are complements to or substitutes for the included (i.e., observed) input, then they are necessarily jointly chosen with the included input and hence a function of the same explanatory variables. One can also prove that the income effect associated with a fixed budget set, irregardless of a cross-price relationship between inputs, suggests that the model is not identified as specified (Mityakov and Mroz, 2012). Hence, strong assumptions are necessary for estimation of unbiased effects of health inputs on health outcomes.

Additionally, it is necessarily the case that consumption in every period is correlated with the discrete factor, or permanent heterogeneity term, μ. But, consumption for the first period of observation in the data cannot be explained by the dynamic equation [12] because the econometrician does not observe smoking behavior before period one. Hence, an initial condition (i.e., smoking in the first observed period) can be specified in its reduced form and must be jointly estimated with the dynamic equation to obtain the correct distribution of the unobserved permanent heterogeneity. It is also necessary that the econometrician be able to identify this initial condition. That is, there must be a variable that explains the initially observed behavior (or outcome) that does not also explain subsequent behaviors (or outcomes) conditional on the lagged behavior (or initial condition in this case).

Generalized method of moments

Finding appropriate instruments (or identification variables) for estimation of these dynamic models is a big hurdle for econometricians. To address this difficulty, the methods have exploited the dynamic variation in behaviors, outcomes, and events over time in search of an instrument. As another example, the first-differenced generalized method of moments (GMM) estimator uses the twice-lagged dependent variable as the instrumental variable. Additional lags can also be used.

Identification in this context is similar to that in the DFRE approach. Both are identified through the variation in complete histories of the exogenous variables in the equations

being jointly estimated. Think of it this way: If cigarette consumption in period t depends on cigarette consumption in period $t-1$ (and period t prices of cigarettes), and one wants to model cigarette consumption using a dynamic equation every period, then the entire history of cigarette prices explains current smoking behavior. At the individual level (or state level) there is likely to be additional variation in this history of cigarette prices relative to the variation in the last period's cigarette price.

When multiple behaviors need to be modeled (i.e., the health outcome and the endogenous health input), GMM estimation can combine the set of moment conditions specified for the equations in levels with additional moment conditions specified for the equations in first differences. In this case, twice-lagged variables serve as instruments because they are uncorrelated with the differenced time t and $t-1$ error terms.

Up to now, it has been assumed that these time-varying unobservables are drawn every period, from the same distribution, where, by assumption, these errors are not persistent. That is, a draw in one period does not depend on the draw in the previous period. However, it may be the case that this time-varying heterogeneity is not completely subsumed (or reflected) by the observed period t behavior or outcome. Rather, the disturbance term may be autoregressive (i.e., $v_{i,t+1}=\lambda v_{it}+e_{it}$). With the differenced GMM estimator, the coefficient λ can be estimated. An econometrician could also use copula functions to explicitly model the serial correlation in nonpermanent, time-varying error terms.

Solution and Estimation of Dynamic Theoretical Models

This article has discussed econometric methods that attempt to recover the causal effect of variables of interest on outcomes of interest in a dynamic setting. Often, however, one may want to measure (or estimate) the value of a parameter of interest for which we do not have a corresponding observable variable. Consider the third example of dynamics in health-related decision making presented above: the optimal health insurance selection. A health economist may desire to understand what determines observed insurance choices. Theory suggests that a person's risk aversion (or aversion to the financial loss associated with reduced health and subsequent medical care consumption) plays a role in determining how much health insurance is optimal for him. Economists capture risk aversion with the shape of the utility function. A linear function, for example, reflects no risk aversion: given the risk of poor health (or medical care expenditure) an individual would be indifferent between having health insurance coverage and financing the full cost of care out of pocket. A concave utility function reflects risk aversion (or risk avoidance). But how can an econometrician use observed data to recover this unobserved risk preference?

This question requires that the econometrician parameterize and solve the individual's optimization problem (eqn [3]) and use data on observed health insurance choices, medical care utilization (or expenses), individual characteristics, and prices of insurance to estimate the parameters of the model. Rather than measuring correlations or causal effects in linearized demand functions (Example 2) or stand-alone production functions (Example 1), solution and estimation of the parameterized optimization problem (Example 3) recovers the preferences, constraints, technologies, and expectations of forward-looking individuals. Not only are the recovered parameters easily interpreted as common constructs used by economists, the estimated model can also be used to evaluate interesting health policy alternatives when variation in such policy parameters are not available in the observed data.

Looking specifically at the optimal insurance selection example, an econometrician solves an expected utility maximization problem. The shape of the nonlinear utility function, which depends on 'disposable wealth' (because, by assumption, the individual gets happiness from consumption which costs money), and the shape of the distribution of financial risk that an individual faces (i.e., medical care expenses), are critical components of the optimal solution. To understand optimal behavior, the econometrician must accurately capture both aversion to risk and the risk distribution. Of much importance is accurately capturing the tail of the expenditure distribution, for it is the rare or unlikely, large financial loss events that reduce happiness (or utility) the most. To complicate things further, these constructs may depend on individual unobservables that are likely correlated over time.

There is not enough space in this article to detail the methods used to solve and estimate such dynamic discrete choice problems in health economics. The methods that recover structural or primitive parameters, like those that recover reduced-form parameters, also require identification. The econometrician must be very specific about the observed behavior that helps to estimate the parameters of interest. Nobel prize winning economist Jim Heckman describes the problem of identification that econometricians want to avoid as occurring when 'many different theoretical models and hence many different causal interpretations may be consistent with the same data' (Heckman, 2000). Thus, no matter which estimation procedure is adopted, the econometrician must be clear about what assumptions are being made to justify identification, because the assumptions will inevitably affect interpretation of the estimated parameters. And after all, it is these measured parameters that form the basis of the answers to our originally posed questions.

Summary

This article introduces the concept of dynamics in an economist's model of behavior. The three examples depict economic relationships between variables over time using a demand function, a production function, and full solution of an individual's optimization problem. The reader is introduced to the main econometric problems associated with estimating models with dynamics. The article briefly discusses some econometric methods used to address the intertemporal dependence exhibited by dynamic empirical relationships. The article concludes by explaining the importance of theory in supporting both the justification for causal empirical interpretations as well as the understanding of dynamic health-related relationships over time.

See also: Abortion. Addiction. Advertising as a Determinant of Health in the USA. Alcohol. Education and Health. Health and Its Value: Overview. Health Care Demand, Empirical Determinants of. Illegal Drug Use, Health Effects of. Macroeconomy and Health. Nutrition, Economics of. Peer Effects in Health Behaviors. Pollution and Health. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Sex Work and Risky Sex in Developing Countries. Smoking, Economics of

References

Arrow, J. K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5), 941–973.

Chaloupka, F. J. and Warner, K. E. (2000). The economics of smoking. In Anthony, J. C. and Joseph, P. N. (eds.) *Handbook of health economics* (Part B), vol. 1, pp. 1539–1627. Elsevier. ISSN 1574-0064, ISBN 9780444504715, 10.1016/S1574-0064(00)80042-6.

Heckman, J. (2000). Causal parameters and policy analysis in economics: A twentieth century. *Quarterly Journal of Economics* **115**(1), 45–97.

Mityakov, S. and Thomas, M. (2012). Economic theory as a guide for the specification and interpretation of empirical health production functions. Working paper.

Further Reading

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277–297.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115–143.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy* **80**(2), 223–255.

Economic Evaluation of Public Health Interventions: Methodological Challenges

HLA Weatherly, RA Cookson, and MF Drummond, University of York, York, UK

Glossary

Budget constraint The limit to expenditure imposed by a cash-limited budget.

Efficiency A resource allocation is efficient if it is not possible to reallocate resources so as to increase one person's utility (or health, or output) without decreasing another person's utility (or health or output). In health economics the entity maximized is generally assumed to be utility, health, or welfare.

Extra-welfarism or Nonwelfarism A normative framework of economics which holds the evaluation of a policy or resource allocation should be based on a larger set of information than solely the utilities attained by members of society.

Matching (in biostatistics) Selecting a control population that is matched on some characteristics that may influence the outcome of interest independently of the disorder in question. Also a process through which pairs of individuals are brought together in order to trade, share, or otherwise engage in some mutual activity.

Partial equilibrium analysis Classic demand and supply analysis in which each market is treated in isolation from all others, compared with general equilibrium analysis.

Quality-adjusted life-year The quality-adjusted life-year (QALY) is a generic measure of health-related quality of life that takes into account both the quantity and the quality of life generated by interventions.

Value-based pricing A method of pricing pharmaceuticals that links their prices to the estimated value of the health benefits they generate.

Welfarism A tenet of welfare economics which holds that the evaluation of a policy or resource allocation should be based solely on the utilities attained by members of society.

Willingness to pay The maximum sum an individual (or a government) is willing to pay to acquire some good or service, or the maximum sum an individual (or government) is willing to pay to avoid a prospective loss. It is usually elicited from controlled experiments in which subjects reveal their willingness to pay.

Introduction

There has long been an aspiration to invest in promoting health, preventing ill health, and reducing health inequality. This aspiration can be realized through a wide variety of public health interventions, including not only screening, vaccination, and other preventive activities undertaken by healthcare professionals but also a broad range of fiscal and social programs and regulations beyond the healthcare sector with impacts on the health of the population.

Economic evaluation is increasingly used to inform decisions about which public health interventions to fund from scarce resources. However, there remains a dearth of evidence on the effectiveness and cost–effectiveness of public health interventions and the evidence that is available tends to be relatively weak – at least compared with evidence for healthcare interventions – with important methodological challenges remaining. In the UK, for example, the Wanless Report of 2004 noted the lack of evaluations undertaken in the public health field and the lack of methods development, expertise and funding available to generate the evidence.

Health economic guidelines for assessing the cost–effectiveness of healthcare technologies – such as new drugs, devices and medical procedures – have existed in many jurisdictions across the developed world for well over a decade. By contrast, methods for the economic evaluation of public health interventions are less well-established. Healthcare technology assessment guidelines use an economic evaluation framework to provide a clear and transparent approach for assessing the relative costs and benefits of alternative options with the aim of achieving an efficient allocation of resources. Typically, in relation to healthcare technology assessment, this framework focuses on the decision-making objective of maximizing health gain subject to an exogenously fixed healthcare budget constraint.

The economic evaluation of many public health interventions raises additional methodological challenges. As with the evaluation of standard 'clinical' healthcare technologies, it is important to determine effect estimates for use within economic evaluations of public health interventions. However, public health interventions tend to be directed at populations or communities rather than specific individuals, and can be less suited to evaluation through randomized controlled designs: the gold standard of study design for obtaining unbiased estimates of effect. In addition, public health interventions tend to generate a broad range of nonhealth benefits and opportunity costs, which may extend beyond the healthcare sector, with implications for other sectors subject to different objectives and budget constraints. Lastly, although standard health economic evaluation methods focus on maximizing health gain, a particular feature of many public health interventions is a concern to reduce health inequalities and so decision makers may be interested in information about the distribution of health levels, gains and opportunity costs within the general population as well as the average health gain for recipients of the intervention. It is, therefore, not clear how far standard methodological guidelines for healthcare technology assessment are appropriate in public health, and some public health scholars have argued

that over-zealous application of standard health technology assessment (HTA) evaluation processes and criteria in public health can lead to systematically misleading conclusions.

This article briefly reviews standard methods for the economic evaluation of healthcare interventions before identifying key methodological challenges for the economic evaluation of public health interventions. To illustrate the methodological issues, it contrasts National Institute for Health and Clinical Excellence (NICE) (http://www.nice.org.uk/) methods guides for economic evaluation of healthcare technologies and public health interventions, respectively, linking this to the methodological challenges of undertaking economic evaluation of public health interventions. Finally, it explores ways forward, noting some recent methods developments in the field.

Methods for the Economic Evaluation of Healthcare Interventions

Economic evaluation offers a clear analytical framework for assisting decision making. In the presence of limited resources and a fixed healthcare budget, economic evaluation offers a transparent approach, underpinned by explicit social value judgments, for choosing how to allocate society's scarce healthcare resources. To do this, decision-making objectives and comparator interventions are identified and the opportunity cost of selecting a particular intervention is assessed by considering whether its' value exceeds the value that would have been achieved if the next best alternative intervention were selected, given available resources. Costs and consequences of relevant alternative activities are compared, with the most efficient use of resources being the option that provides the best outcome.

There are two main philosophical approaches underpinning economic evaluation: the 'welfarist' approach and the 'non-welfarist' approach. Each has implications for the economic evaluation methods of choice. The key outcome in the welfarist approach is the satisfaction of individual preferences, typically measured using willingness to pay (WTP) reflecting the maximum amount an individual would pay for a particular intervention. In contrast, the nonwelfarist approach focuses on some other measure of outcome reflecting the decision-maker's objective, such as the quality-adjusted life-year (QALY) which can be used as a summary measure of total population health benefit. In practice, the nonwelfarist approach and the use of cost–effectiveness analysis (CEA) based on QALYs predominates in the healthcare sector, whereas the welfarist approach and the use of cost–benefit analysis (CBA) based on WTP estimates predominates in other areas of social policy such as transport, occupational safety, the environment, employment, housing, and so on. However, each approach can be applied in different ways and both CEA and CBA could in principle be conducted using different outcome measures.

One of the most comprehensive and commonly referred to set of guidelines for health economic evaluation methods is the NICE health technology 'reference case' from the UK. These guidelines are periodically revised – they were first issued in 2004, updated in 2008, and a third edition is currently being prepared that may incorporate substantial revisions related to a new system of 'value-based pricing' due to be introduced from 2014. The 2008 version, which sets out a thoroughly 'non-welfarist' approach to undertaking economic evaluations of healthcare interventions, is described in Box 1. Under this approach, the aim is to maximize health given a fixed budget constraint, whereby funding the new intervention involves displacing one or more existing interventions. A new intervention is considered cost–effective if the extra cost incurred to gain an extra one QALY, relative to the next best intervention, is less than approximately £20 000 to £30 000. This is the cost–effectiveness threshold value and represents the opportunity cost or health forgone by the displaced activity.

Methods Challenges for the Economic Evaluation of Public Health Interventions

Unlike clinical healthcare interventions, public health interventions tend to be directed at populations or communities

Box 1 Summary of the HTA and the public health reference cases (National Institute for Health and Clinical Excellence, 2008, 2009)

Element of assessment	*NICE HTA reference case*	*NICE public health reference case*
Defining decision problem	The scope developed by NICE	The scope developed by NICE
Comparator	Therapies routinely used in NHS	Therapies routinely used in Public sector
Perspective on costs	NHS and PSS	Public sector, including the NHS and PSS
Perspective on effects	All health effects on individuals	All health effects on individuals
Type of economic evaluation	CEA	Primary analysis CEA Secondary analysis CCA, CBA
Synthesis of evidence on outcomes	Based on a systematic review	Based on a systematic review
Measure of health effects	QALYs	QALYs
Source of data for measurement of HRQoL	Reported directly by patients and/or carers	Reported directly by patients and/or carers
Source of preference data for valuation of changes in HRQoL	Representative sample of the public	Representative sample of the public
Discount rate	Annual rate 3.5%, costs and health effects	Annual rate 3.5%, costs and health effects
Equity position	Additional QALY same weight regardless of other characteristics of individuals receiving health benefit	Additional QALY same weight regardless of other characteristics of individuals receiving health benefit

rather than specific individuals. One implication of this for evaluation is that public health interventions often have relatively small and hard-to-detect effects at the level of the individual, which can nevertheless sum to large effects at population level. Public health interventions also tend to generate a broader range of costs and nonhealth benefits, including costs falling on private consumption and public sector budgets beyond the healthcare sector. Finally, whereas standard economic evaluation methods focus on efficiency with the aim of maximizing health gain, typically the aim of public health interventions extends further to include a concern for reducing unfair health inequalities. Indeed, some public health professionals would go so far as to say that the primary goal of public health interventions is to generate a more equitable distribution of health in society. Based on these considerations, adjustments to standard health economic evaluation methods may be required to assess public health interventions. The formulation of social objectives, the range of outcomes and opportunity costs to be quantified, and the methods for evaluating those outcomes and opportunity costs may all need to be reconsidered, to align the methods with the broader scope and goals of public health interventions.

A number of reviews have explored the challenges of economic evaluation in public health. Four key methods challenges can be identified in applying standard health economic evaluation methods to public health interventions. These include (1) attributing outcomes to interventions; (2) measuring and valuing outcomes; (3) incorporating equity considerations; and (4) identifying inter-sectoral costs and consequences, as developed further below.

Attributing Outcomes to Interventions

Before undertaking CEA, it is essential to determine the effectiveness of relevant comparator interventions. Most healthcare interventions are directed at identified groups of individuals, and the randomized controlled trial (RCT) design is typically used as the 'gold standard' study design for primary data collection. Most published guidelines in the healthcare field, including the current NICE reference case, indicate a preference for using RCT evidence to identify and measure the effectiveness of relevant comparators. Some individually focused, face-to-face public health interventions may be suitable for evaluation using an RCT. However, this might not be feasible for more community-based public health interventions, and other forms of experimental data, such as cluster randomized trials, may not be available for obtaining effect estimates. Because there are likely to be fewer controlled trials of public health programs, it will often be necessary to use other approaches for obtaining an unbiased estimate of the intervention effect. Natural experiments and the use of nonexperimental data can be used to fill some gaps in the public health evidence base. However, evidence of this kind is vulnerable to selection bias. Where study participants are not randomized to the interventions, the effects estimated for different interventions may be biased because of confounding between assignment to the intervention and the study participant characteristics. This implies that effectiveness estimates may, in part, be caused by differences in population characteristics instead of the intervention of interest. More use may be made of statistical techniques that have been developed to analyze nonexperimental data, including a range of econometric methods and simulation modeling methods. Methods are available to improve the validity of the comparator groups through study design, such as matching patients across interventions and in the analysis of results by statistically adjusting for case mix, assuming sufficient data is available to do so.

Given the dearth of RCT evidence for many public health interventions, systematic reviews of evidence which exclude all non-RCT design evidence may not yield parameter estimates that could be used in economic evaluations. Instead, a common outcome would be that there was insufficient research evidence available for assessing effectiveness. Other methods such as narrative review summaries are not particularly helpful for analysts requiring empirical estimates. Instead, broader evidence synthesis techniques are required, which enable the analyst to include all relevant evidence within an economic evaluation, including robust nonexperimental evidence from natural experiments as well as RCTs. Modeling is also required to extend the analysis to an appropriate time horizon. This may be of particular importance for public health interventions which impact health over the longer term. Modeling is also required to indicate how uncertainty in the available evidence translates to the probability that a particular decision is the correct one.

Measuring and Valuing Outcomes

Typically the main aim of a new healthcare technology is to improve health. By contrast, public health interventions tend to generate a broad range of health and nonhealth benefits, which may extend beyond the healthcare sector. Many health economic guidelines, including the NICE reference case for HTA, recommend that health outcomes are measured in QALYs. For public health interventions, however, a range of nonhealth outcomes may also be relevant – including fairly tangible crime, education and employment outcomes as well as harder-to-measure outcomes such as public reassurance, the empowerment of citizens to make informed choices, and community cohesion. Some of these outcomes may be possible to incorporate within a QALY-type framework, others not. Therefore, the use of other outcome measurement and valuation methods may be appropriate – for example, subjective well-being scores, or multidimensional indices of well-being in which health is only one component, or WTP-based methods including the possibility of using some form of 'adjusted' WTP after purging the influence of income, incomplete information, misperceptions of risk, protected characteristics under equalities legislation and/or other determinants of 'raw' WTP that social decision makers may consider inappropriate reasons for public resource allocation decisions.

Intersectoral Costs and Consequences

Public health interventions may have impacts that extend beyond the healthcare sector. Costs and benefits associated with public health interventions can fall on different sectors of the public sector. For example, a public health intervention to reduce substance abuse may reduce criminal justice costs.

Interventions that are implemented in other sectors of the economy may also have public health implications. For example, improvements in housing could reduce illness and injuries, with consequent reductions in healthcare utilization and expenditure. In addition, individuals may incur out-of-pocket costs associated with accessing and using interventions. There may be ripple effects associated with an intervention that could extend across other sectors of the economy, including the private and voluntary sectors. An obvious way of addressing this challenge would be to monetize benefits. However, this still raises practical questions about how different health and nonhealth outcomes are to be valued and how to address the issue of fixed budget constraints faced by healthcare and other public sector decision makers, rather than assuming that all costs and outcomes are costlessly exchangeable between different policy sectors. It also raises deeper theoretical questions about whether and how it is possible to integrate 'welfarist' and 'nonwelfarist' approaches.

If a healthcare and personal social services (PSS) perspective is chosen, as for the NICE HTA reference case, resource use and costs falling on the healthcare and PSS sector are evaluated but impacts falling on other sectors are not. Where the healthcare sector budget is fixed, a new intervention can only be funded if other activity is displaced within the sector. There is an opportunity cost incurred in investing in the new intervention, i.e., the health forgone among the group of patients whose intervention is displaced and therefore no longer available. Under the NICE HTA reference case, this health opportunity cost is approximated by the cost–effectiveness threshold value. Using this approach, for approximately every £20 000 of expenditure the opportunity cost is one QALY lost through displacing existing interventions. However, the relevant opportunity costs and threshold values are likely to differ across sectors – with different sectors having different threshold values for both health and nonhealth opportunity costs.

The NICE HTA reference case recommends that analysts undertake CEA on the basis of benefits measured in QALYs and costs covering National Health Service (NHS) and PSSs resource use. To identify possible inter-sectoral impacts of public health interventions, the costs and benefits falling on other sectors could be considered for each comparator intervention. The cost–consequence analysis (CCA) approach, whereby a range of sector-relevant costs and consequences are measured and reported separately, could be informative. In addition, it might be appropriate to account for these impacts using real or hypothetical budgetary transfers across sectors whereby sectors that gain net benefits can 'compensate' sectors that lose net benefits – although the feasibility of this approach requires further investigation. For example, if a generic but sector-specific measure of outcome such as the QALY were identified for each sector, this could be used in reference to the relevant cost–effectiveness threshold value for the sector to compute net benefits for each sector.

Incorporating Health Equity Considerations

The final key methods challenge identified for the economic evaluation of public health interventions is a concern to reflect the health equity implications of public health interventions The importance of achieving health equity is recognized in many published guidelines for economic evaluation. However, in practice, health equity considerations are rarely quantified. In terms of health outcomes, it is typically assumed that the value of a QALY is the same whoever receives it. The analysis will also contain some judgment about which types of resource use to cost, and this can be influenced by equity considerations – for example, considerations of non-discrimination may influence judgments about how far to count productivity costs, which can be much higher for highly paid workers compared with those on low pay and economically inactive groups such as children and pensioners. However, current analyses do not examine health inequality issues – in particular the distribution of QALY levels or gains between population subgroups, for example, by socioeconomic status, ethnicity and gender – which are of particular interest in public health.

To reflect health inequality considerations, data on equity-relevant subgroups need to be identified, collected, and analyzed. Assuming the decision maker has the twin objectives of both reducing health inequality and improving health, if the cost-effective intervention is the option that also minimizes health inequality then the decision is clear. However, if one intervention achieves greater health outcome and the comparator intervention achieves greater health equality then under standard cost-effectiveness decision rules it is not clear which intervention to choose. Some methods have been proposed for dealing with this trade-off issue, as reviewed in the section on methods developments below.

Recent Guidance for the Economic Evaluation of Public Health Interventions

Before reviewing recent methods developments in the field, it is useful to refer to NICE guidance that has been developed to facilitate a consistent and transparent approach to undertaking good quality economic evaluations of public health interventions in the UK. The NICE guide to methods for the development of public health guidance is periodically updated: it was first issued in 2006, and has been updated in 2009 and 2012. Described below is the 2009 version, which is the most directly comparable to the 2008 NICE 'reference case' for technology appraisal. The main relevant changes in the 2012 edition are a reduced discount rate for costs and benefits and an even stronger emphasis on conducting CCA and CBA as well as CEA using QALYs, following a recent shift of public health budgets in England to local government and away from the healthcare sector. As detailed in **Box 1**, the NICE public health reference case (right-hand column) differs in a few characteristics compared to the NICE HTA reference case (left-hand column). These differences illustrate that elements of assessment have been adapted to reflect the characteristics of public health interventions.

The NICE public health reference case reflects the fact that public health interventions may involve resources, costs, and outcomes beyond the healthcare sector. It recommends that "important health effects and resource costs are all included"

and "effects and outcomes not related to health are included (if they are important for the public sector)." Therefore, NICE recommends that analysts include this information in addition to the information recommended for the NICE HTA reference case. As comparison of the two approaches shows, for the NICE HTA reference case the perspective on cost is fairly restrictive, including only NHS and PSSs costs. Also, the prescribed measure of health benefit is the QALY. Each QALY gained is assumed to have the same weight regardless of the other characteristics of the individuals receiving the health benefit (e.g., their age, socioeconomic status, or severity of their health condition). The NICE methods for developing public health guidance differ in that the perspective on costs is extended to encompass all costs falling on the public sector, recognizing the broader, cross-sector nature of most public health interventions. To facilitate comparability between NICE decisions, the QALY remains the primary measure of health outcome in the 'reference case.' However, for the public health reference case it can be supplemented by CCA and CBA approaches in order to take account of the broader aims and scope of public health interventions. This allows explicit consideration of multiple, nonhealth related and/or outcomes that are difficult to quantify. It also means that the impact of the interventions on the distribution in health gains can be evaluated to inform public health policy.

Methods Developments in the Economic Evaluation of Public Health Interventions

Given increasing interest in the economic evaluation of public health interventions and current public health economic evaluation guidance, it is useful to review some of the ongoing methods requirements and developments in the area that might be used in future evaluations.

Attributing Outcomes to Interventions

In terms of primary data collection to assess the relative effectiveness of public health interventions, it is often feasible to undertake individual or cluster randomized RCTs and where possible this is recommended for measuring outcomes, although it is likely to be possible only over the short term. Where this is not feasible, nonrandomized trials may be undertaken and use of methods to restrict entry to the interventions based on those with particular characteristics or selecting controls that match the cases in terms of the confounding factor(s) may prove useful. As it is likely that outcomes of interest will extend beyond the length of trial follow-up, it is useful if outcomes measured match those available in longer term observational studies. Analytical techniques may be used to analyze nonexperimental data including econometrics. Economics has a long tradition of analyzing nonexperimental data for deriving effects attributable to a range of public health interventions. These include various matching techniques such as propensity scores, difference in differences techniques, time series analyses of natural experiments, and, where appropriate, more sophisticated econometric modeling and structural modeling. In addition, Bayesian methods may be useful in synthesizing data (modeling) including in examining the data where participants in studies do not match typical NHS patients, where intermediate outcomes are used, where relevant comparators have not been used, where long-term costs and benefits extend beyond the trial period and in quantifying the decision uncertainty and variability around the estimates. Further research might be undertaken to develop the methods for synthesizing all relevant data, experimental and nonexperimental and aggregate and individual-level data, for use in economic evaluations of public health interventions.

Measuring and Valuing Outcomes

Given that the aims and scope of public health interventions tend to be broader than for standard healthcare interventions, the measures of outcome chosen need to reflect this. As discussed above, the NICE reference case recommends CEA using QALYs as the primary form of analysis, with patients and/or carers as the source of data for measurement of health-related quality of life (HRQoL) and with values based on a representative sample of the views of the public. CCA and CBA are recommended as a secondary analysis to include other measures of outcome appropriate to decision making given the interventions evaluated.

Ongoing research includes development of sector-specific generic outcomes based on the QALY approach, for example, a social care QALY, development of a nonsector specific multidimensional measure such as a well-being index and a nonsector specific unidimensional measure such as happiness. The choice of outcome measure reflects the normative foundation underpinning the analysis and can also reflect the impact of the intervention on a particular sector or across multiple sectors of the economy.

Inter-Sectoral Costs and Consequences

Where the costs and consequences of public health interventions extend beyond the healthcare sector, the NICE HTA reference case methods would need to be extended to demonstrate this impact. The NICE public health reference case accounts for broader outcomes in the sense that the use of the cost–consequence or cost–benefit approach enables the analyst to describe other outcomes beyond the QALY and the healthcare sector. In terms of costs, besides NHS and PSSs costs, other public sector costs may be considered. Use of the cost–consequence approach could, however, mean that decision rules are not explicit as there are no standard decision rules using this approach and it is not clear the value decision makers would attach to different impacts in order to come to a decision about the cost–effectiveness of an intervention. The use of CBA may require a shift in the normative position to a more 'welfarist' perspective, though it may also be possible to monetize at least some nonhealth outcomes using methods other than estimating 'raw' WTP for those outcomes – for example, using 'adjusted' WTP estimates, sector-specific threshold values, and relative valuations of those outcomes in terms of other outcomes that are more readily monetized.

Research is ongoing to assess the practicalities of evaluating possible budgetary transfers across different sectors of the

economy. In particular, an inter-sectoral compensation test approach to analyze the net benefit of costs, which fall on different sectors of the economy is being explored, and a stochastic mathematical programming approach is being developed to explore how to allocate resources in the context of different budgets and different budgetary policies across sectors.

Finally, research is also being undertaken to assess the use of a general equilibrium approach to simultaneously consider the impact of interventions across all sectors of the economy. The large majority of health economic evaluations undertaken to date take a partial equilibrium approach to analysis by assuming all other costs (and benefits) remain the same apart from those being evaluated. This is appropriate for evaluating the impact of most healthcare interventions. However, some health issues such as antimicrobial resistance and potentially pandemic diseases (e.g., seasonal flu, severe acute respiratory disease (SARS)) might have a macroeconomic impact that alters broader resource use and costs in the economy as a whole.

Incorporating Equity Considerations

The NICE public health reference case makes no explicit mention of equity considerations. However, the use of CCA, and assessment of subgroups in sensitivity analysis, assuming sufficient individual-level data on equity-related subgroups, may enable the analyst to include health equity issues. However, as a starting point, relevant health equity characteristics need to be identified and could include a whole range of possibilities such as socioeconomic status, degree of voluntariness or personal responsibility for health risk, and the value of treating current ill health versus preventing future health risk. If, following evaluation, the most cost–effective option is likely to be judged inequitable, either on the grounds of health inequality impact or procedural justice, it would be possible to assess the opportunity cost of not selecting that option, in terms of aggregate health gain forgone or additional resources used. Another approach that has been suggested is quantitative health impact assessment, allowing for health opportunity costs as well as health gains. Here, once a health inequality or a set of health inequalities have been determined, the distribution of net health impacts of the intervention is assessed by different equity-relevant groups. Building on this approach, it may be possible to assess the magnitude of any reduction in health inequality following adoption of the intervention and to clarify trade-offs with the objective of maximizing population health improvement.

The NICE reference cases state that an additional QALY is given the same weight regardless of other characteristics of individuals receiving health benefit. There has been some research into public and stakeholder views on equity weighting in a public health context and considerable additional research to overcome technical and practical issues is required to examine how much sacrifice in total population health is merited in order to pursue particular equity goals.

Summary

Economic evaluation provides a clear analytical framework for combining evidence and explicit social value judgments to inform decisions about how far investments of scarce resources in public health interventions are worthwhile. Given tight budgets and ballooning healthcare costs worldwide, policymakers are increasingly interested in ways of shifting the balance of effort toward preventative activity that has the potential to both improve health and reduce healthcare cost. So policymakers are likely to have an ongoing interest in assessing how far investments in public health interventions represent good value for money. Methods for the economic evaluation of healthcare interventions need to be adapted and refined in relation to the four methods challenges identified, in order to help analysts undertake good quality, relevant economic evaluations of public health interventions. Methods development in the field is still at an early stage and further research is required to improve the usefulness of methods and to pilot new methods with the aim of providing more useful information to support decisions about the investment of scarce resources into public health interventions.

See also: Adoption of New Technologies, Using Economic Evaluation. Dynamic Models: Econometric Considerations of Time. Economic Evaluation, Uncertainty in. Health Econometrics: Overview. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Latent Factor and Latent Class Models to Accommodate Heterogeneity, Using Structural Equation. Nonparametric Matching and Propensity Scores. Primer on the Use of Bayesian Methods in Health Economics. Public Health in Resource Poor Settings

References

National Institute for Health and Clinical Excellence (2008). *Guide to the methods of technology appraisal*. London: NICE. Available at: www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (accessed 07.02.13).

National Institute for Health and Clinical Excellence (2009). *Methods for development of NICE public health guidance*. 2nd ed. London: NICE. Available at: www.nice.org.uk/phmethods (accessed 07.02.13).

Further Reading

Blundell, R. and Costa, D. M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies* **21**(4), 427–468.

Claxton, K., Sculpher, M. and Culyer, A. (2007). *Mark versus Luke? Appropriate methods for the evaluation of public health interventions*. Centre for Health Economics Research Paper 31, University of York.

Cookson, R., Drummond, M. and Weatherly, H. (2009). Explicit incorporation of equity considerations into economic evaluation of public health interventions. *Health Economics Policy and Law* **4**, 231–245.

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Cabinet Office, Behavioural Insights Team. Available at: SSRN: ssrn.com/abstract=2131581 or dx.doi.org/10.2139/ssrn.2131581 (accessed 08.05.13).

McKenna, C., Chalabi, Z., Epstein, D. and Claxton, K. (2010). Budgetary policies and available actions: A generalisation of decision rules for allocation and research decisions. *Journal of Health Economics* **29**(1), 170–181.

Medical Research Council (2008). *Developing and evaluating complex interventions: New guidance*. London: UK Medical Resource Council. Available at: www.mrc.ac.uk/complexinterventionsguidance (accessed 08.01.13).

Medical Research Council (2012). *Using natural experiments to evaluate population health interventions: Guidance for producers and users of evidence*. London: UK

Medical Research Council. Available at: www.mrc.ac.uk/naturalexperimentsguidance (accessed 08.01.13).

Smith, R., Yago, M., Millar, M. and Coast, J. (2005). Assessing the macroeconomic impact of a healthcare problem: The application of computable general equilibrium analysis to antimicrobial resistance. *Journal of Health Economics* **24**, 1055–1075.

Vining, A. and Weimer, D. L. (2010). An assessment of important issues concerning the application of benefit–cost analysis to social policy. *Journal of Benefit–Cost Analysis* **1**(1), Article 6. Available at: www.bepress.com/jbca/vol1/iss1/ (accessed 08.01.13).

Wailoo, A., Tsuchiya, A. and McCabe, C. (2009). Weighting must wait: Incorporating equity concerns into cost effectiveness analysis may take longer than expected. *Pharmacoeconomics* **27**, 983–989.

Wanless, D. (2004). Securing good health for the whole population. *Final Report*. London: HM Treasury.

Weatherly, H., Drummond, M., Claxton, K., et al. (2009). Methods for assessing the cost–effectiveness of public health interventions: Key challenges and recommendations. *Health Policy* **93**, 85–92.

Economic Evaluation, Uncertainty in

E Fenwick, University of Glasgow, Glasgow, Scotland, UK

Sources of Uncertainty

Uncertainty exists wherever the truth is unknown either due to imperfect information or imperfect measurement. Within economic evaluations of healthcare, there are a number of sources of uncertainty. Methodological uncertainty relates to the analytic methods used to undertake an economic evaluation. Sources of methodological uncertainty include whether discounting should be employed and, if so, at what rate or rates, and whose preferences should be used to value health outcomes (those of the patient, public, or professional). Structural uncertainty relates to the structure and assumptions employed within an analysis. For example, the assumptions underlying the extrapolation of outcomes from a trial or the choice of the number of health states in a Markov model. This type of uncertainty is particularly relevant to (although not limited to) model-based analyses. It is often overlooked within analyses, largely due to the complexities of incorporating changes to structure and assumptions, despite the potential for considerable impact on results. Stochastic (or first order) uncertainty reflects differences in how interventions are experienced and impact within a population. For example, the different length and/or severity of adverse events experienced by different patients with the same prognosis receiving the same intervention. Stochastic uncertainty reflects random variation between people within the population and is represented by the sample variance (in trial-based studies) or the dispersion in the output from first order Monte Carlo simulation (in model-based studies). Uncertainty within the population is not the main focus for economic evaluation which is concerned, instead, with uncertainty at the population level. As such, stochastic uncertainty will not be covered in this article. Note that stochastic uncertainty is fundamentally different to heterogeneity which reflects the variation between people that can be explained by their specific identifiable characteristics. These characteristics might include, for example, age, gender, ethnicity, geographical location. Heterogeneity is best handled through the use of subgroups within the analysis, with results either presented independently for each subgroup or, if required, included in a weighted analysis for an aggregate group. Finally, parameter uncertainty reflects the uncertainty associated with specific parameters employed within an analysis. For example, the uncertainty surrounding the effectiveness of an intervention or the utility value associated with a particular health state.

Incorporating Uncertainty within Analyses

The existence of these various uncertainties within an economic evaluation inevitably leads to uncertainty in the estimation of the costs, effects, and cost-effectiveness associated with the health intervention and ultimately to uncertainty in the decision about whether or not to fund the intervention. Undertaking an analysis of these uncertainties allows an assessment of the impact that they have on the results; illustrating the robustness of the results to changes in the inputs used in the analysis and assessing confidence in decisions. An analysis of uncertainty can also contribute to an assessment of the value of undertaking further research through a formal value of information analysis. According to the recent joint International Society for Pharmacoeconomics and Outcomes Research and Society for Medical Decision Making Modeling Good Research Practices Task Force Working Group Guidelines, "(t)he systematic examination and responsible reporting of uncertainty are hallmarks of good modeling practice. All analyses should include an assessment of uncertainty and its impact on the decision being addressed" (Briggs *et al.*, 2012). This assessment of uncertainty usually takes the form of sensitivity analysis (SA), where assumptions or parameter values used in the economic evaluation are systematically varied to observe the impact on the results. Within deterministic SA (DSA), this systematic variation is performed manually to ascertain the impact associated with specific combinations of assumptions and/or parameters (see Section Deterministic Sensitivity Analysis). In contrast, probabilistic SA (PSA) involves repeatedly varying all of the uncertain parameters simultaneously, in order to get an overall assessment of the impact of the uncertainty. Of the sources of uncertainty described in Section Sources of Uncertainty, only parameter uncertainty can be assessed using either DSA or PSA. Methodological and structural uncertainties should not be assessed within a PSA. In addition, in certain circumstances scenario analyses are employed (e.g., when investigating heterogeneity). Here, alternative assumptions or parameter values associated with specific subgroups are substituted into the economic evaluation to examine the impact on the results.

Deterministic Sensitivity Analysis

DSA involves manually varying the parameter values or assumptions employed within the economic evaluation to test the sensitivity of the results to these values. There are a number of methods available to undertake DSA. One-way, two-way, and multiway SA involve substituting different values for one, two, or more parameter(s), method(s), or assumption(s) at a time and examining the impact on the results. The results of DSA can be displayed either graphically or through the use of tables, or conversely the results can be summarized in the text. This is fairly straightforward for one-way, two-way, and even three-way SA (which can employ contour plots) but becomes more of a challenge with multiway SA when more than three parameters are changed simultaneously. Analysis of extremes involves changing all parameters and/or assumptions to their most extreme values (which can be either best or worst case values) simultaneously and assessing the impact on the results. The results can be reported in the text. All these methods of DSA require that the range of values that the

 doi:10.1016/B978-0-12-375678-7.01419-X

parameter(s) or assumption(s) can take is specified before the analysis. These ranges should be informed by and incorporate the available evidence base. In contrast, the final method of DSA, threshold analysis requires no such information. Here, the levels of one or more parameters, assumptions or methods are varied to identify the point at which there is a significant impact on the results, for example, the intervention becomes cheaper, more effective, or cost-effective. Again, the results can be displayed graphically, in tables or in the text. It is then left up to the user of the results to interpret and determine whether the values identified constitute reasonable levels for the parameter, assumptions, or methods.

Probabilistic Sensitivity Analysis

PSA involves repeatedly varying all of the uncertain parameters employed within an economic evaluation simultaneously, to get an overall assessment of the impact of the uncertainty. As such, PSA requires the specification of probability distributions for each parameter to fully reflect the parameter uncertainty. Each of these probability distributions represents both the range of values that the parameter can assume and the likelihood that the parameter takes any specific value within the range.

Assigning probability distributions to parameters

Within a PSA there are three main methods for assigning probability distributions to parameters:

1. Using patient-level data
2. Using secondary data from the literature
3. Assessing and incorporating expert opinion

Where sample data are available (e.g., from a clinical study) it can either be incorporated directly into the analysis through the use of bootstrapping (see Section Propagating uncertainty – bootstrapping) or the moments of the data can be used to fit a probability distribution.

Where historical data are available from previously published studies, this should be used to specify the probability distribution for the parameter. Here the premise is to match what is known about the parameter in terms of its logical constraints, behavior etc. with the characteristics of the distribution. As such, particular distributions are the most appropriate for specific parameters. For example, beta distributions should be used to specific uncertainty in probabilities, log-normal distributions should be used for relative risks or hazard ratios and gamma or log-normal distributions should be used for right-skewed parameters such as costs.

Where there are no primary or historical data available from which to specify the probability distribution for a particular parameter, then expert opinion can be used. However, care must be taken when eliciting opinions from experts, to ensure that it is the uncertainty in the parameter that is captured rather than various estimates of the mean. For example, the Delphi method is commonly used when eliciting expert opinion, however, this approach generally produces a single point estimate through consensus and therefore does not capture uncertainty. It is important that parameters are not excluded from the analysis of uncertainty because they have little information with which to estimate the parameter – these are precisely the parameters that need to be included, and with a wide distribution to represent the uncertainty.

Propagating uncertainty – Monte Carlo simulation

Once probability distributions are assigned to the parameters, the uncertainty is propagated through the use of (second order) Monte Carlo simulation. Here, a value is selected for each parameter from its probability distribution and the associated cost and effects are estimated based on these specific parameter values. These selections are most commonly made randomly from each probability distribution. Although recently, latin hypercube or orthogonal sampling (where selections are sampled from a specific section of the probability distribution) have been suggested to improve efficiency in sampling. The process is repeated thousands of times and a distribution of expected costs and effects is generated. These distributions reflect uncertainty at the population level, with each iteration representing a possible realization of the uncertainty that exists in the analysis, as characterized by the probability distributions.

Propagating uncertainty – bootstrapping

Within a trial-based study, an estimate of the population-level uncertainty can be obtained through bootstrapping the sample data. Here, samples are repeatedly taken at random from the original sample. These samples are each the same size as the original sample and are drawn with replacement. As with a (second order) Monte Carlo simulation, the bootstrap provides a distribution of the expected costs and effects associated with the intervention.

Presenting Uncertainty

Tornado Plots

Tornado plots can be used to illustrate the impact on the results (i.e., costs, effects, or cost-effectiveness) associated with a series of one-way SA involving different parameters (Figure 1). Here, the uncertainty in the results associated with the uncertainty in each parameter is illustrated in a series of stacked bars (one per parameter). The length of each bar illustrates the extent of the uncertainty in the results associated with the uncertainty in that particular parameter. The parameters (bars)

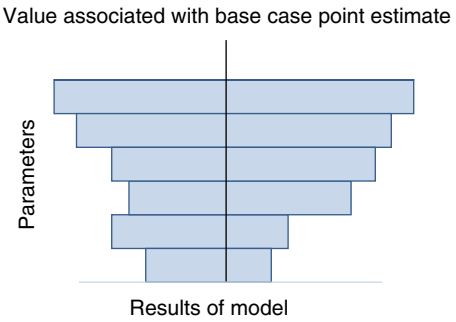


Figure 1 Tornado diagram illustrating the impact on the results of uncertainity in each parameter.

are stacked in order of length from smallest to longest (i.e., the parameters for which uncertainty in the parameter has the smallest impact on uncertainty in the results are at the bottom) forming a funnel or tornado shape. All of the bars are aligned around the result (cost, effect or cost-effectiveness) corresponding to the base case value for the parameter, hence the bars are not necessarily symmetrical and the funnel/tornado is not necessarily smooth.

Cost-Effectiveness Planes

Uncertainty in the costs and effects associated with an intervention, generated either from a probabilistic SA or from bootstrapping trial data, can be graphically represented on a cost-effectiveness plane. Where the decision involves only two interventions, the incremental costs associated with the intervention of interest are plotted against the incremental effects for each iteration from the simulation, as a series of incremental cost-effect pairs, on an incremental cost-effectiveness plane. Incremental costs are conventionally plotted on the *y*-axis with incremental effects on the *x*-axis. As such, the slope between any specific cost-effect pair in the plane and the origin represents the incremental cost-effectiveness ratio (ICER) associated with that cost-effect pair (i.e., the incremental cost/incremental effect). The plane is split into four quadrants by the origin (which represents the comparator). The NE and SE quadrants involve positive incremental effects associated with the intervention of interest, whereas the NE and NW quadrants involve positive incremental cost. Figure 2 illustrates the joint distribution of incremental costs and effects as a cloud of points on the incremental cost-effectiveness plane.

The location of the incremental cost-effect pairs in relation to the *y*-axis indicates whether there is uncertainty regarding the existence, or not, of cost-savings. For example, if all of the incremental cost-effect pairs are located above the origin (in the NE and/or NW quadrants) then the intervention is definitely more expensive. The spread of the incremental cost-effect pairs in relation to the *y*-axis indicates the extent of the uncertainty regarding the magnitude of incremental costs. For example, in Figure 2, the incremental cost-effect pairs are plotted closely together in terms of incremental cost indicating that there is little uncertainty surrounding the magnitude of the incremental cost. The same holds for the location and spread of the incremental cost-effect pairs in relation to the *x*-axis and the existence and extent of uncertainty in the incremental effects. For example, in Figure 2, the location and spread of the incremental cost-effect pairs indicate that there is no uncertainty regarding the existence of an effect benefit associated with the intervention of interest (in comparison to the alternative) but that there is considerable uncertainty regarding the size of the effect benefit.

The incremental cost-effectiveness plane provides a good visual representation of the existence and extent of the uncertainty surrounding the incremental costs and effects individually. In addition, the location of the joint distribution of incremental costs and effects (the cloud of incremental cost-effect pairs) within the four quadrants of the incremental cost-effectiveness plane can provide some information about the cost-effectiveness of the intervention. If the cloud is located completely in the SE quadrant (or the NW quadrant) then there is no uncertainty in the cost-effectiveness; the intervention dominates (is dominated by) the alternative. Where the cloud of incremental cost-effect pairs falls into the NE or SW quadrants or straddles more than one quadrant, the incremental cost-effectiveness plane does not provide a useful summary or assessment of the uncertainty in the cost-effectiveness. In addition, a distinction must be made between uncertainty in the cost-effectiveness of the intervention and uncertainty in the

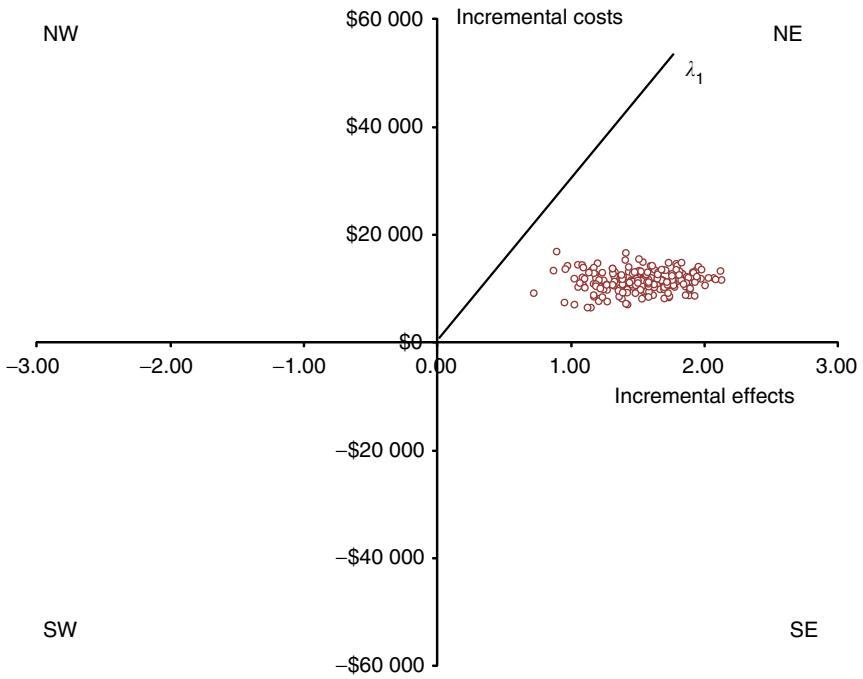


Figure 2 Incremental cost-effectiveness plans.

decision to adopt the intervention based on the current information about costs, effects and cost-effectiveness (decision uncertainty). Decision makers using the results of economic evaluations to guide decisions about whether to adopt new interventions are interested in the latter. An assessment of the decision uncertainty requires the comparison of the joint distribution of the incremental costs and effects with a predetermined, external threshold level representing the willingness to pay for the effects (λ) to determine the proportion of the joint distribution that falls below the threshold. The assessment of the decision uncertainty is not too daunting when the cloud of incremental cost and effect pairs falls into just one or even two quadrants, or when the cost-effectiveness threshold is known with certainty. Returning to Figure 2, at a willingness to pay threshold of λ_1 there is no uncertainty associated with the adoption of the intervention despite the considerable uncertainty in the cost-effectiveness of the intervention. This is because the entire joint distribution of incremental costs and effects falls below (to the South and East of) the cost-effectiveness threshold (λ_1). Where the cloud of incremental cost-effect pairs falls into the SE or NW quadrants there is also no decision uncertainty; the intervention is definitely cost-effective (SE) or definitely not cost-effective (NW). When the joint distribution of costs and effects covers three or all of the quadrants, or the cost-effectiveness threshold is unknown then the assessment of the decision uncertainty will involve considerable computation, and the incremental cost-effectiveness plane will not provide a useful summary of the decision uncertainty.

Where the decision involves more than two interventions, the costs and effects for each intervention are plotted (for each iteration from the simulation) as a series of cost-effect pairs, on a cost-effectiveness plane (see Figure 3). Here, the spread of the cost-effect pairs for an intervention in the y-axis (x-axis) provides information on the extent of the uncertainty in the costs (effects). In addition, the location of the cost-effect pairs for an intervention in comparison to the cost-effect pairs for other interventions provides some information about the existence of uncertainty in the incremental costs and incremental effects. For example, in Figure 3, it is possible to determine that intervention B is definitely more expensive than intervention A (incremental cost is positive), but it is not possible to determine that it is more effective than intervention A. In contrast, intervention C is both more costly and more effective than both A and B. For decisions involving more than two interventions, the cost-effectiveness plane can not provide an assessment of the uncertainty in the cost-effectiveness or an assessment of the decision uncertainty. For these assessments, knowledge is required regarding the relationship between each of the cost-effect pairs for each intervention (i.e., which cost-effect pair for intervention A relates to which cost-effect pair for intervention B and which cost-effect pair for intervention C). This information is not easily presented or computed in the cost-effectiveness plane.

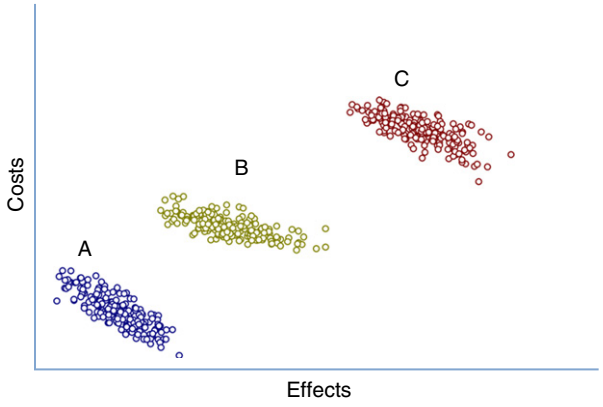


Figure 3 Cost-effectiveness plane.

Cost-Effectiveness Acceptability Curves

Cost-effectiveness acceptability curves (CEAC) provide a graphical representation of the decision uncertainty associated with an intervention. They present the probability that the decision to adopt an intervention is correct (i.e., that the intervention is cost-effective compared with the alternatives given the current evidence) for a range of values of the cost-effectiveness threshold (λ). This probability is essentially a Bayesian definition of probability (i.e., the probability that the hypothesis is true given the data), although some commentators have given the CEAC a Frequentist interpretation.

Where the decision involves only two interventions, the decision uncertainty is derived from the joint distribution of incremental costs and effects, as the proportion of the incremental cost-effect pairs that are cost-effective. In an incremental cost-effectiveness plane, this can be identified as the proportion of cost-effect pairs that fall below a specific cost-effectiveness threshold (as described above). The CEAC is then constructed by quantifying and plotting the decision uncertainty for a range of values of the cost-effectiveness threshold (λ). As noted in Section Cost-Effectiveness Planes, incremental cost-effect pairs that fall in the SE (or NW) quadrant are always (never) cost-effective, as such these incremental cost-effect pairs are always (never) counted in the numerator of the proportion. Incremental cost-effect pairs that fall in the NE and SW quadrants are either considered cost-effective or not depending on the cost-effectiveness threshold (λ). When the cost-effectiveness threshold (λ) is zero (i.e., the decision maker places no value on effects), only incremental cost-effect pairs in the SE and SW quadrants will be considered cost-effective (i.e., those with negative incremental costs). When the cost-effectiveness threshold (λ) is infinite (i.e., the decision maker only values effects and places no value on costs), only incremental cost-effect pairs in the NE and SE quadrants will be considered cost-effective (i.e., those with positive incremental effects). Between these two levels, as the cost-effectiveness threshold (λ) increases (i.e., the decision maker increasingly values effects), incremental cost-effect pairs in the NE (SW) quadrant are added to (removed from) the numerator. This reflects the fact that incremental cost-effect pairs in the NE quadrant (i.e., positive cost, positive effect) increasingly provide effects at a cost lower than the decision maker would be prepared to pay, whereas those in the SW quadrant involve a loss of effects without the level of savings that the decision maker would require. As a result, the CEAC does not represent a cumulative distribution function; its

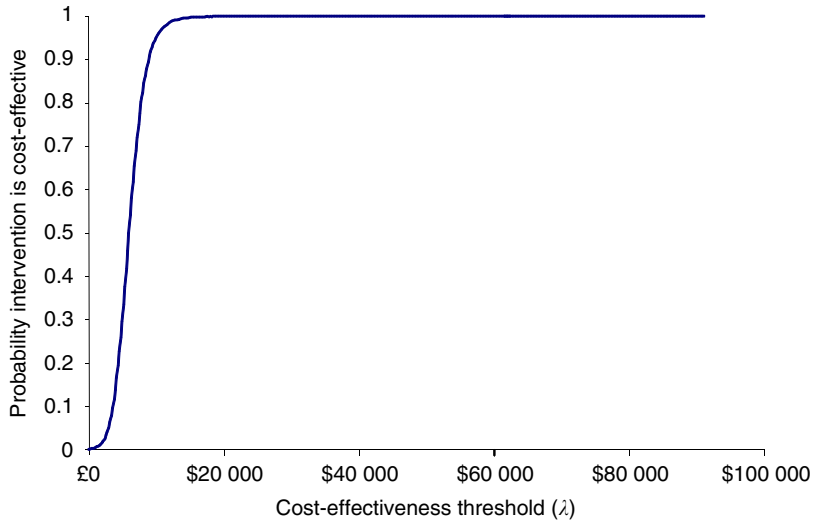


Figure 4 Cost-effectiveness acceptability curve.

shape and location will depend solely on the location of the incremental cost-effect pairs within the incremental cost-effectiveness plane. Figure 4, presents a CEAC for a decision involving two interventions. By convention, for decisions involving only two interventions, the CEAC is only shown for the new intervention of interest, however, the CEAC for the alternative could also be presented. Given that the interventions are mutually exclusive and collectively exhaustive (i.e., for each incremental cost-effect pair the new intervention is either cost-effective or the alternative is cost-effective) then the CEAC for the alternative has the opposite shape and location, with the curves crossing at a probability of .5.

Where the decision involves more than two interventions, CEACs can be constructed for each intervention by determining the decision uncertainty associated with each intervention compared to all of the alternatives simultaneously (i.e., the probability that the intervention is cost-effective compared with all of the alternatives given the current evidence). Again, as the interventions are mutually exclusive and collectively exhaustive (i.e., for each cost-effect pair intervention only one of the interventions A, B, or C is cost-effective) then the CEAC for every intervention will vertically sum to one. It is inappropriate to present a series of CEACs that compare each intervention in turn to a common comparator, as this provides no indication of the uncertainty surrounding the decision between the interventions. Figure 5 presents a series of CEACs associated with a decision involving more than two interventions.

It is very important to stress that the CEAC simply indicates the decision uncertainty associated with an intervention for a range of values of λ. Thus, in the context of expected value decision making (where the decision is made on the basis of the expected costs, effects, and cost-effectiveness) the CEAC does not provide any information to aid the decision about whether to adopt the intervention or not. Therefore statements concerning the CEAC should be restricted to statements regarding the uncertainty surrounding the decision to select a particular intervention, or the uncertainty that the intervention is cost-effective, compared with the alternatives given the current evidence. Information from the CEAC should not be used to make statements about whether or not to adopt an intervention.

The cost-effectiveness acceptability frontier (CEAF) has been suggested to supplement the CEAC in the context of expected value decision making. The CEAF provides a graphical representation of the decision uncertainty associated with the intervention that would be chosen on the basis of expected value decision making. As such, the CEAF provides no additional information about the decision uncertainty, it simply replicates the CEAC for the intervention that would be selected by the decision maker at each value of the cost-effectiveness threshold (λ). As such, discontinuities occur in the CEAF at values of the cost-effectiveness threshold (λ) at which the decision alters (see Figure 5).

Intervals and Distributions for Net Benefits

Net benefits (NB) have been suggested as an alternative method to present the results of economic evaluations. In this framework, the issues associated with ICERs are overcome by incorporating the cost-effectiveness threshold (λ) within the calculation to provide a measure of either the net health benefit or the net monetary benefit. Here, following a probabilistic SA or bootstrap of trial data, the cost and effect pairs for every iteration are replaced by an estimate of NB; generating a distribution of net benefit. Where the decision involves two alternatives, the incremental net benefit (INB) can be used. The uncertainty can be either be summarized and presented as a confidence interval for (I)NB or presented in full as a distribution of (I)NB. Given that the net benefit measure incorporates the cost-effectiveness threshold, where the threshold is unknown the results must be provided for a range of values of the threshold. Figure 6 presents the confidence interval for the INB for a range of values of the cost-effectiveness threshold as an INB curve. This curve provides information

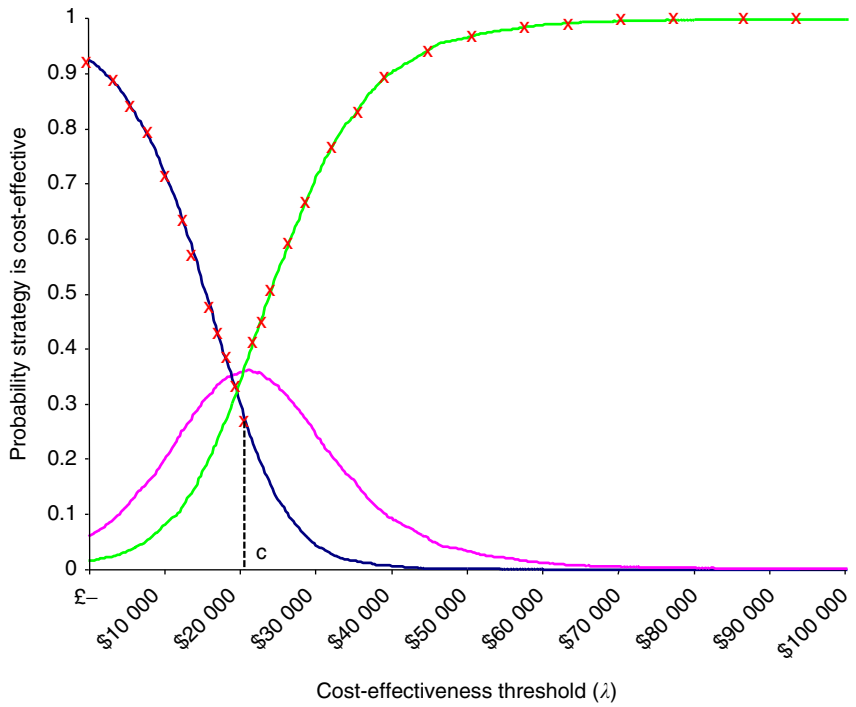


Figure 5 Cost-effectiveness acceptability curves and cost-effectiveness acceptability frontier. *Note*: c represents the value of the ICER where the decision switches; the shape of the CEAF is identified by **X**.

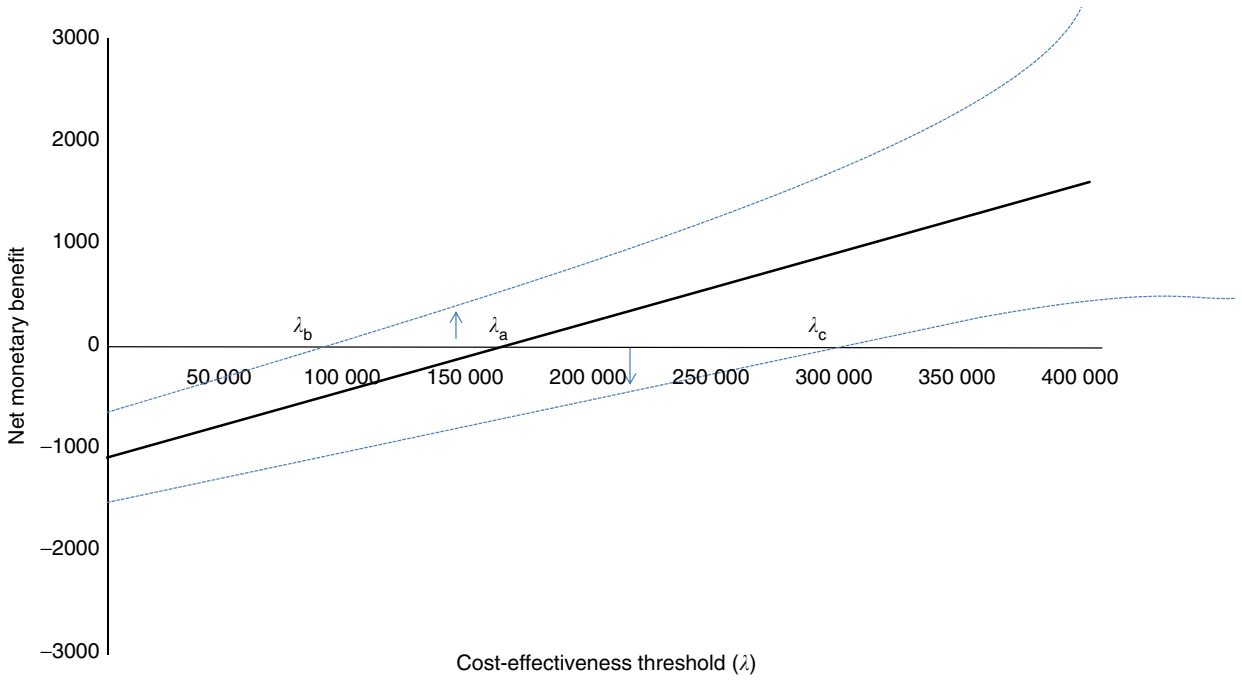


Figure 6 INB curve.

about the extent of the uncertainty in INB as well as identifying which intervention to adopt (on the basis of expected value decision making) for every value of the cost-effectiveness threshold (λ). For example, in Figure 6 for values of the threshold above λ_a the intervention should be adopted (as the INB > 0), below λ_a the alternative should be adopted. With regard to the decision uncertainty associated with the intervention, at values for the threshold below λ_b there is no

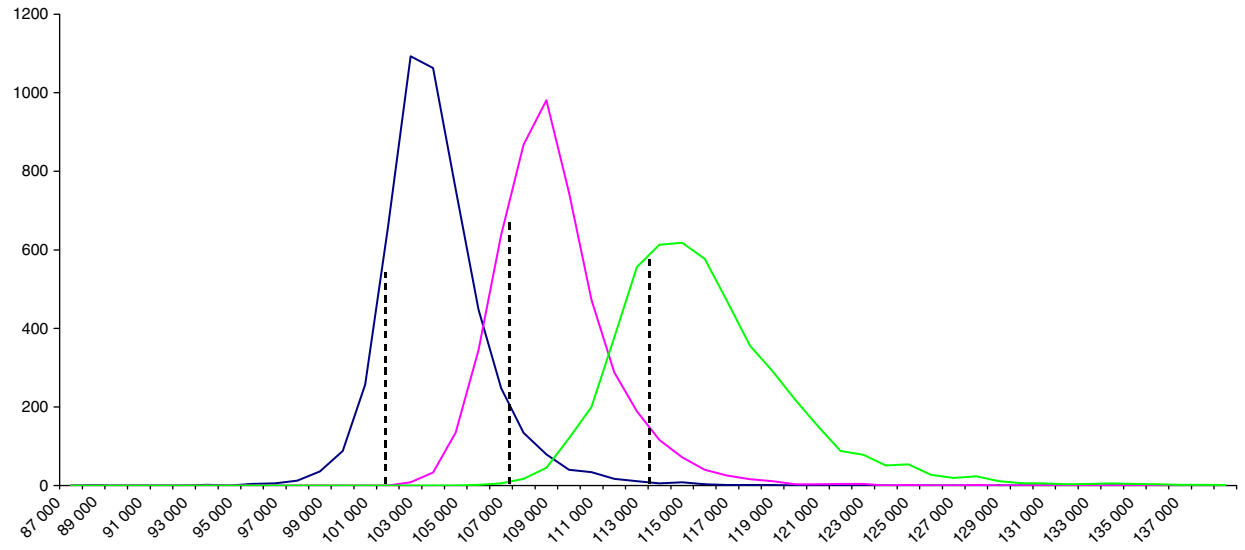


Figure 7 Distributions of net monetary benefit.

decision uncertainty; the intervention is not cost-effective. At values for the threshold above λ_c there is no decision uncertainty; the intervention is cost-effective. For values of the threshold between λ_b and λ_c assessment of the decision uncertainty associated with the intervention requires an evaluation of the proportion of the distribution of INB that falls above zero (i.e., the vertical distance from the *x*-axis to the 95% line). The decision uncertainty associated with the comparator is given by the proportion of the distribution of INB that falls below zero (i.e the vertical distance from the 5% line to the *x*-axis). Figure 7 presents distributions of NB for a particular value of the cost-effectiveness threshold (λ). As noted earlier, where the threshold is unknown the distributions would have to be provided for a range of values of the threshold. Figure 7 provides information about the extent of the uncertainty in the NB associated with each intervention as well as identifying which intervention to adopt (on the basis of expected value decision making) for a specific value of the cost-effectiveness threshold (λ). An assessment of the decision uncertainty associated with an intervention would require an evaluation of the proportion of the distribution of NB that overlaps with the NB distributions associated with the other interventions. Where the decision involves more than two interventions, this evaluation is not straightforward. Therefore it is only in the situation that the NB distributions are distinct (i.e., do not overlap) and there is no decision uncertainty, that the figure provides any information about the decision uncertainty associated with the interventions.

Linking Analysis of Uncertainty to Decision Making

The presence of decision uncertainty means that there is inevitably some possibility that decisions made on the basis of the available (uncertain) information will be incorrect and introduces the possibility of error into decision making. Where the decision maker has the authority to delay or review decisions (based on either additional evidence that becomes available, or that they request) an analysis of uncertainty is important because it links to the value of additional research.

See also: Adoption of New Technologies, Using Economic Evaluation. Analysing Heterogeneity to Support Decision Making. Information Analysis, Value of. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Statistical Issues in Economic Evaluations. Value of Information Methods to Prioritize Research

Reference

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., et al. (2012). Model parameter estimation and uncertainty analysis. A report of the ISPOR-SMDM modeling good research practices task force working group. *Medical Decision Making* **32,** 722–732. Available at: http://mdm.sagepub.com/content/32/5/722.full (accessed 24.07.13).

Further Reading

Briggs, A. H. (2001). Handling uncertainty in economics evaluation and presenting the results. In Drummond, M. F. and McGuire, A. (eds.) *Economic evaluation in health care: Merging theory and practice,* ch. 8, pp. 172–215. Oxford, UK: Oxford University Press.

Briggs, A. H. and Gray, A. M. (1999). Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technology Assessment* **3**(2), 1–63. Available at: http://www.hta.ac.uk/fullmono/mon302.pdf (accessed 24.07.13).

Briggs, A. H., Sculpher, M. J. and Claxton, K. P. (2006). *Decision modelling for health economic evaluation. Handbooks in health economic evaluation*. Oxford, UK: Oxford University Press.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., et al. (2012). Model parameter estimation and uncertainty analysis. A report of the ISPOR-SMDM modeling good research practices task force working group. *Value in Health* **15,** 835–842. Available at: http://www.ispor.org/workpaper/Modeling_Methods/Model_Parameter_Estimation_and_Uncertainty-6.pdf (accessed 24.07.13).

Claxton, K. (2008). Exploring uncertainty in cost-effectiveness analysis. *Pharmacoeconomics* **9**, 781–798.

Fenwick, E., Claxton, K. and Sculpher, M. (2001). Representing uncertainty: The role of cost-effectiveness acceptability curves. *Health Economics* **10**, 779–787.
Fenwick, E., O'Brien, B. and Briggs, A. H. (2004). Cost-effectiveness acceptability curves: Facts, fallacies and frequently asked questions. *Health Economics* **13**, 405–415.
Hunink, M., Glasziou, P., Siegel, J., et al. (2001). Decision making in health and medicine. Integrating evidence and values. In Hunink, M., Glasziou, P., Siegel, J., et al. (eds.) *Variability and uncertainty*, ch. 11, pp. 339–363. Cambridge, UK: Cambridge University Press.

Education and Health

D Cutler, Harvard University and NBER, Cambridge, MA, USA
A Lleras-Muney, UCLA, Los Angeles, CA, USA

In their seminal 1965 study, Kitagawa and Hauser documented that mortality in the US fell with education. Since then a very large number of studies have confirmed that the well-educated enjoy longer lives: for example, in 1980, individuals with some college education at the age of 25 years could expect to live another 54.4 years, whereas life expectancy at the age of 25 years for those without any college education was only 51.6 years.

Not only are the differences in health by education large, but also, by most measures, these differences have been growing in recent years. For instance, in 2000, those with some college education lived 7 years longer than those without any college education – thus the gap increased by 4 years since 1980. Education not only predicts mortality in the US but also is an important predictor of health in most countries, regardless of their level of development. Furthermore, the life expectancy gaps are growing around the world. Education gradients in mortality since 1980 are also known to have increased in Estonia, Sweden, Finland and Norway, Russia, Denmark, England/Wales, and Italy – although caution must be exercised as the number and composition of individuals within education categories has also changed substantially over time. The more educated are also noticeably healthier while they are alive, as they report being in better health, having fewer health conditions and limitations. Children of educated parents are also in better health in both developed and developing countries.

This review synthesizes what is known about the relationship between education and health in both developed and developing countries. Although previous work has thought of the effect of education separately for richer and poorer countries, there are insights to be gained by integrating the two. For example, education is associated with lower mortality in most developed countries, and this relationship is similar regardless of the generosity of the social protections and health insurance systems that are in place. This suggests that access to care is not the main reason for the association in the first place. This approach is illustrated by comparing the effects of education on various health and health behaviors around the world to generate hypotheses about why education is so often (but not always) predictive of health.

The review then goes on to examine theories for the relation between education and health and then review the empirical evidence on this relationship paying particular attention to causal evidence and evidence on mechanisms linking education to higher health.

Stylized Facts about Education and Health

To examine the link between education and health across countries, data from three sources are combined. Data for most developing countries come from the Demographic and Health Surveys (DHS) for years between 2004 and 2009. Data for the US come from the Behavioral Risk Factor Surveillance System (BRFFS) for 2005. Data for Europe come from the Eurobarometer Surveys (2005 and 2009). Data for a total of 61 countries are known. Each country was matched to its per capita level of gross domestic product (GDP) in the current US dollars reported by the World Bank. To create a consistent sample, the attention is restricted to women aged 15–49 years (the DHS does not collect data on men or older women). More details on the data construction are in the Data Appendix.

Education is measured as years of school in the DHS and the BRFFS, but the Eurobarometer only asks about the age at which a person finished schooling. It is assumed that years of schooling in the Eurobarometer data are 5 years less than the age at which schooling was finished. As some people take significant time off before finishing schooling, the authors truncate schooling at 25 years. Although not ideal, this is the only standardized data source with a large number of countries.

For all of these countries, the measures of height (in centimeters) and weight (in kilograms) are known, which are used to construct body mass index (BMI = weight/height squared), an indicator for being underweight ($BMI \leq 18.5$) and an indicator for being obese ($BMI \geq 30$). The data from the DHS come from actual measures, whereas the data for the US and Europe are self-reported. For all of the countries, it is also known whether the person is currently a smoker. For a few developing countries and all developed countries, it is known whether the person drinks alcohol. Finally, only for developing countries, measures of hemoglobin levels (HbA1c) are known, which is a key indicator of diabetes and a measure of whether the person had a sexually transmitted disease in the past year.

To document basic patterns in the relationship between education and health, the following ordinary least squares (OLS) regression for each country in the sample is estimated:

$$H_{ic} = \beta_0 + \beta_{lc} * Education_{ic} + X_i\alpha + \varepsilon_i \quad [1]$$

where H_i is a health or health behavior indicator of individual i in country c, *Education* is measured in years, and X_i contains basic demographics: age, age squared, marital status, ethnicity, race, and religion dummies. For each country and outcome, the regression coefficient β_{lc} is obtained, which is plotted by the level of GDP (in logs). All of the surveys have complex sampling design schemes, and the weights provided by the survey are used to compute means and weight regressions.

It is difficult to interpret the coefficient of education in these regressions as causal because education and health could be both determined by unobservable factors. Also the coefficient on education might reflect the effect of health on schooling rather than the reverse. These issues are discussed in the Section Evidence on the Causal Effect of Education. For the time being, the correlations that are observed are

 doi:10.1016/B978-0-12-375678-7.00309-6

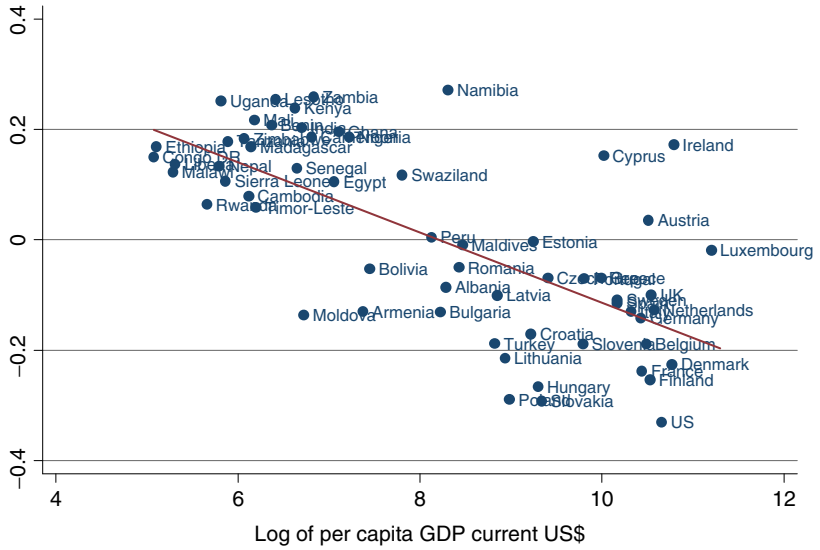


Figure 1 Coefficient of education on BMI by GDP.

described and the reasons for the patterns across countries are hypothesized.

The effect of education is estimated for each country in a linear model that includes years of education. It is not clear whether years of schooling are comparable across countries because the quality of education differs widely by country and thus the actual education of individuals might differ even when years of school are comparable. Ideally, one would use test scores or other measures of achievement (such as literacy and numeracy), but these are not available here or in most surveys. Also, one might prefer to look at nonlinear models, where the effect of education is allowed to vary depending on the level of education. Previous research has generally found that linear models are good approximations, although this refers to high-income countries. Nevertheless, the estimates are of interest because they mirror the standard estimates that are produced when looking at specific countries and times. The results presented here are restricted to women because the DHS surveys collect information systematically on them but not necessarily for men. Previous research documents that correlations between education and health are similar for men and women, although in general, correlations are stronger for men, but this varies depending on the outcome.

Figure 1 shows the education gradient in BMI as it relates to average income – each dot in the graph corresponds to the coefficient of education on BMI obtained from a separate regression for each country. BMI is generally taken as an indicator of short-term nutrition. The figure suggests a clear pattern by income: in poorer countries, those with more education have higher BMIs, whereas the opposite is true in richer countries. The crossover point is income of approximately US$3000 per capita, roughly the income of Bolivia and Peru. However, the relationship between health and BMI is not monotonic: higher weight (given height) is associated with lower mortality at low levels of weight, but after some threshold, increased weight is associated with larger mortality. To disentangle these effects, the next set of estimates reports the effect of education on the likelihood of being underweight and on the likelihood of being obese: both of these are indicators of poor health.

Figure 2 shows the patterns for being underweight. Overall, education is associated with a decrease in undernutrition: most coefficients are either negative and statistically significant or essentially zero (although there are a few exceptions). The effect of education is largest for the poorest countries and then becomes zero (or positive) as GDP rises. This is essentially due to the fact that there is very little undernutrition in countries that have reached middle levels of income, and there is no effect of education on malnutrition when the prevalence rates are low. This is more evident in Figure 3, which plots education coefficients against levels of malnutrition (the share of the population that is underweight).

Figure 4 shows the patterns for obesity. These patterns are very similar to the patterns for BMI: In poorer countries, the effect of education on obesity is positive and significant, whereas it becomes negative and significant for richer countries. This pattern has been noted before and it is more marked for women than men (The graphs presented here only show patterns for women).

Thus, it is observed that around the world the more educated avoid malnutrition, but not always obesity. It is possible that when levels of nutrition are low, obesity is associated with increased survival because people are better able to fight infectious disease, and chronic problems are not large killers. But once infectious diseases fall and chronic conditions become more important, the pattern reverses (conditional on knowledge that obesity is bad). It is also possible that girth is a status symbol or symbol of wealth in societies that are poor; but the same in rich societies where knowledge of the health consequences is widespread, the opposite becomes true, as rich individuals will devote their resources to staying thin and fit. But the data strongly suggest that the effect of education depends on the level of development and the position of the countries in the 'nutrition transition' in particular: as countries

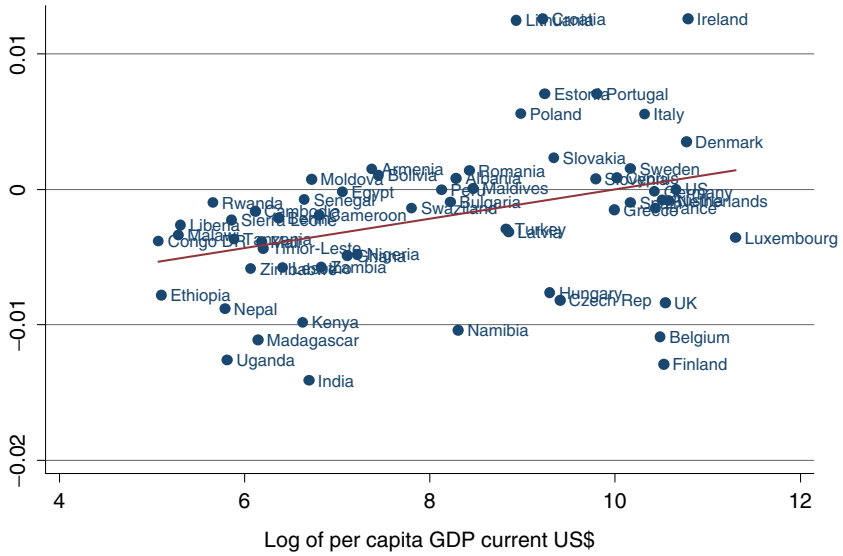


Figure 2 Coefficient of education on underweight by GDP.

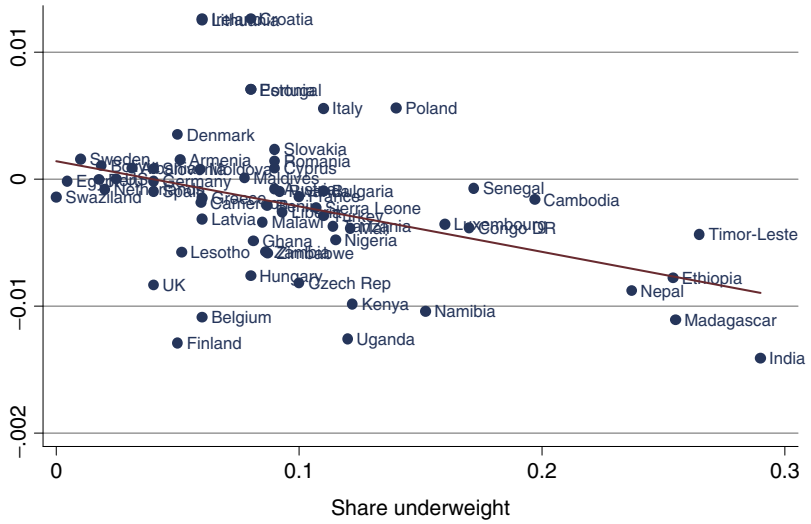


Figure 3 Coefficient of education on underweight by underweight level.

develop, the types of food available (high-fat, high-sugar, and high-density processed foods in particular) and their costs change substantially.

Figure 5 shows the patterns for hemoglobin levels by income – although only for women in developing countries. Again, it is found that the effect of education is protective at low levels of income, and then decreases with GDP; this is again a function of the fact that on average hemoglobin levels rise with GDP. So in poorer countries, the more educated avoid malnutrition. But Figure 6 shows that they do not always avoid disease; among very low-income countries, there are more countries where education is associated with a higher incidence of sexually transmitted infections (STIs) than countries where education is protective. But there is a trend by income again: education is more likely to be protective for higher levels of GDP. Recent work that looks at sexual behavior responses by education level in Africa also reports that the 'effect of education' varies depending on the stages of the human immunodeficiency virus (HIV) epidemic.

Figure 7 shows the patterns for the effect of education on smoking, the leading cause of preventable deaths worldwide. In general, the effect of education on smoking is negative, but for the poorest countries the coefficients tend to be very small. Also, for many middle-income countries, there is a positive effect of education. It is unlikely that this reflects differential knowledge of the harms of smoking among the better educated. The danger of cigarette smoking is well known around the world even in the poorest countries: for example, in Bangladesh, 93% of smokers report that smoking causes lung cancer (International Tobacco Control Policy Evaluation Project). Rather, it may reflect the social acceptability of smoking as income increases or the onset of public policies to reduce smoking at very high incomes. It is also possiblethat in some countries the effects of knowledge are counteracted by the

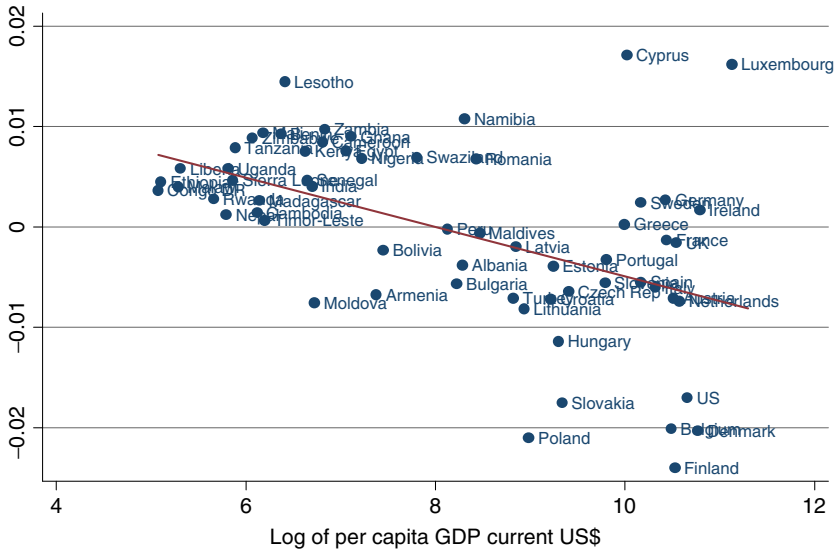


Figure 4 Coefficient of education on obesity by GDP.

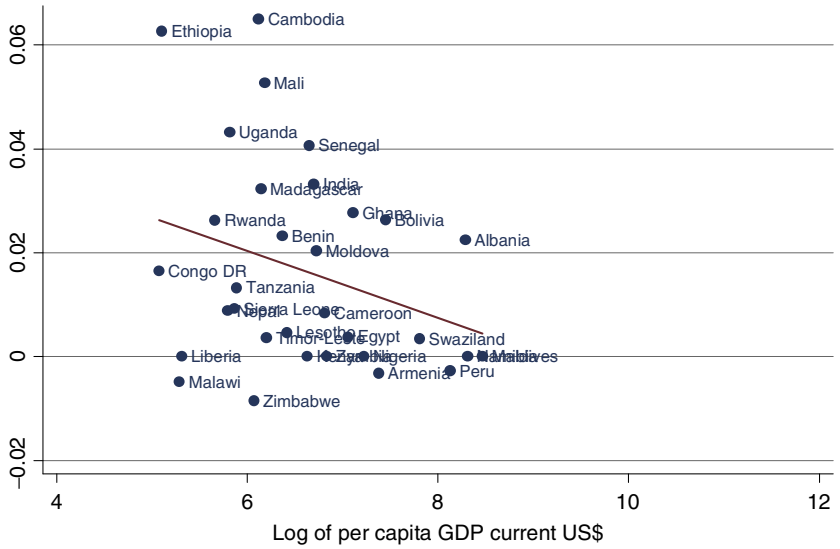


Figure 5 Coefficient of education on hemoglobin by GDP.

effects of higher incomes, because smoking is a normal good. Again these patterns suggest that the effect of education on smoking depends on the level of development defined both in terms of income and knowledge and will, therefore, vary over time and space. Table 1 presents some evidence of this 'smoking transition' for the US. In 1949, high school dropouts were less likely to smoke than high school graduates or individuals with higher education – the opposite of what is observed today. In 1949, dropouts were also more likely to think smoking was harmful. But between 1950 and 1970, the more educated became more likely to think that smoking was harmful as knowledge of the harms of smoking emerged; and by 1969 they were also less likely to smoke.

Figure 8 shows the patterns for drinking. Data on drinking for many developing countries are not known, so somewhat higher income countries are examined. Alcohol appears to be a normal good. Education increases the odds of drinking alcohol in almost all the countries that are examined. Modest alcohol consumption might not be detrimental to health, so it is not necessarily clear that these coefficients have the 'wrong' sign. Ideally, it would be better to determine whether education lowers heavy drinking, which does fall with education levels in the US and the UK, but the data are not consistently available across countries.

The previous figures suggest important patterns by education and could be taken as reflecting causal relationships from education to health. However, it can also be documented that education is partly determined by health by looking at height. Height is generally thought of as an excellent indicator of early childhood environment, as much of the variation in adult heights is determined by the age of 3 years. Thus, the coefficients of education on height from eqn [1] most likely

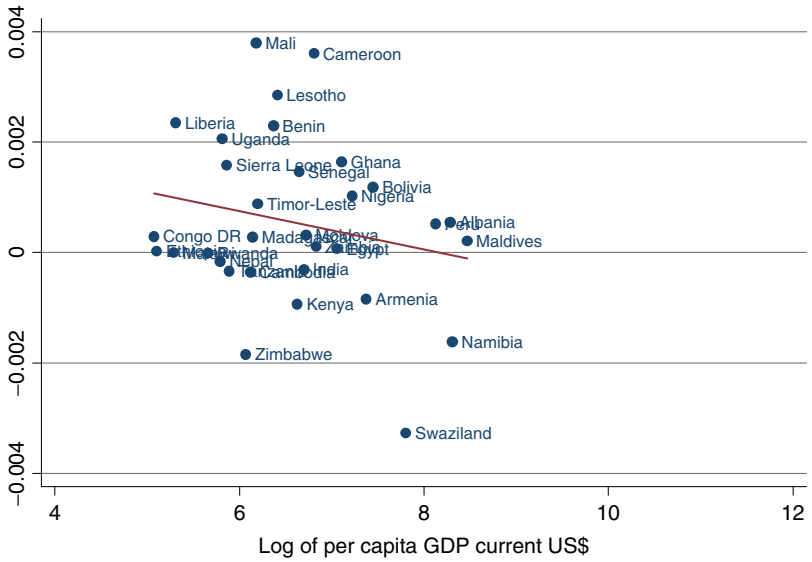


Figure 6 Coefficient of education on STIs by GDP.

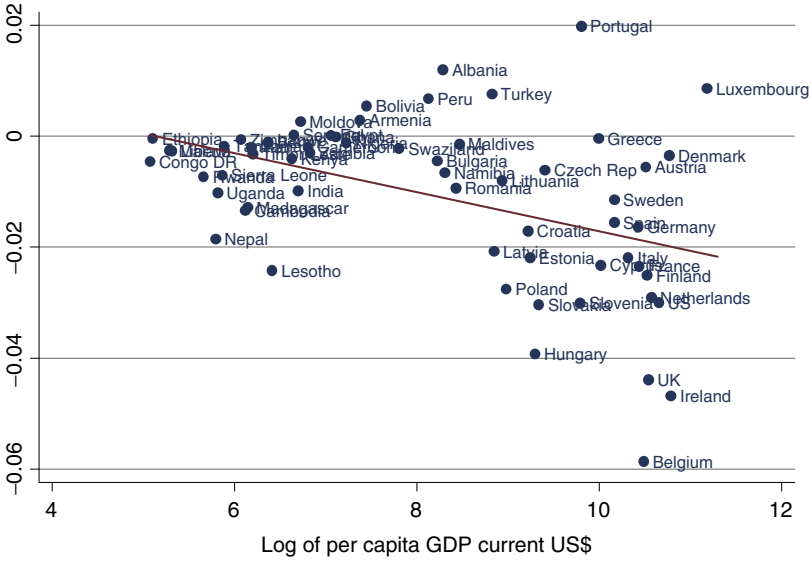


Figure 7 Coefficient of education on smoking by income.

reflect the effect of health on the quantity and quality of education individuals obtain, rather than the effect of education on final height. Before looking at this, two caveats are in order. First, there is a critical growth period in adolescence where the remaining differences in final adult height are determined. Thus, it is possible that some of the relationship between height and education is due to an effect of schooling on height. Second, height itself may be a function of parental education, which may independently affect child education. Nevertheless, most researchers treat the relationship between height and education as mostly reflecting the impact of exogenous health on education.

Figure 9 shows the results for height. For almost all the countries examined, more educated women are taller and the relationship is generally statistically significant. And although the effect falls a bit with GDP, education is still very strongly associated with height, even in very rich countries (with a couple of interesting exceptions among the richest countries).

Summary

All said, the international data on health and education show several stylized facts. The clearest relationship is between income and the education gradient in nutritional intake. Poorer countries are characterized by a mix of undernutrition and overnutrition. Many people are undernourished or anemic in poorer countries, and these outcomes are strongly negatively related to education. Less educated individuals are more likely to be underweight and anemic; better educated people are

Table 1 The evolution of knowledge and smoking gradients in education in the US 1949–69

Year of survey:	*1949*	*1954*	*1957*	*1969*
Panel A: Effect of education on knowledge				
Dependent variable:	"Do you think cigarette smoking is harmful or not?"	What is your opinion – do you think cigarette smoking is one of the causes of lung cancer, or not?		
Less than high school	0.057*	−0.054*	−0.065**	−0.041
Some college	0.012	0.032	0.116**	0.045
College +	0.021	0.067	0.172**	0.111**
Panel B: Effect of education on smoking				
Dependent variable:		Current Smoker?		
Less than high school	−0.056*	−0.016	0.024	0.054*
Some college	0.019	−0.026	−0.008	0.011
College +	−0.045	−0.061	−0.003	−00.076*

All regressions are adjusted for age, sex, and race. Individuals with a high-school degree only are the reference group.
Note: *, significant at the 10%; **, at the 5%.

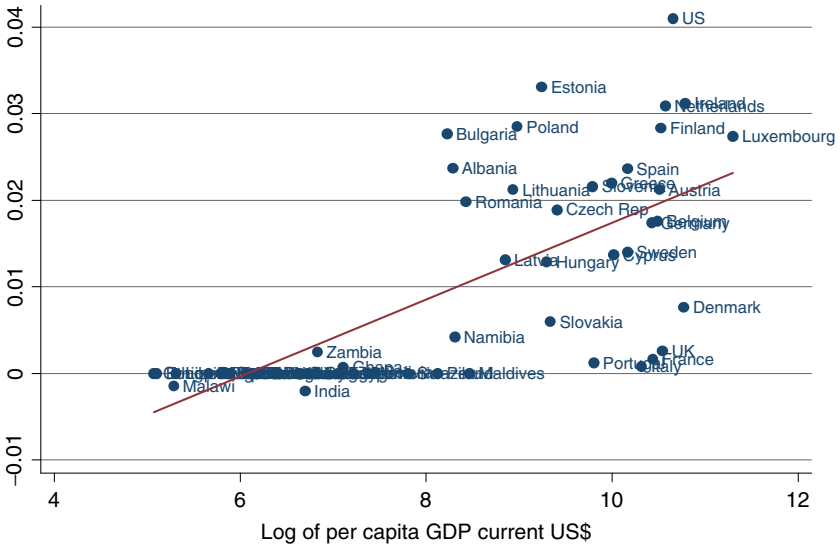


Figure 8 Coefficient of education on drinking by income.

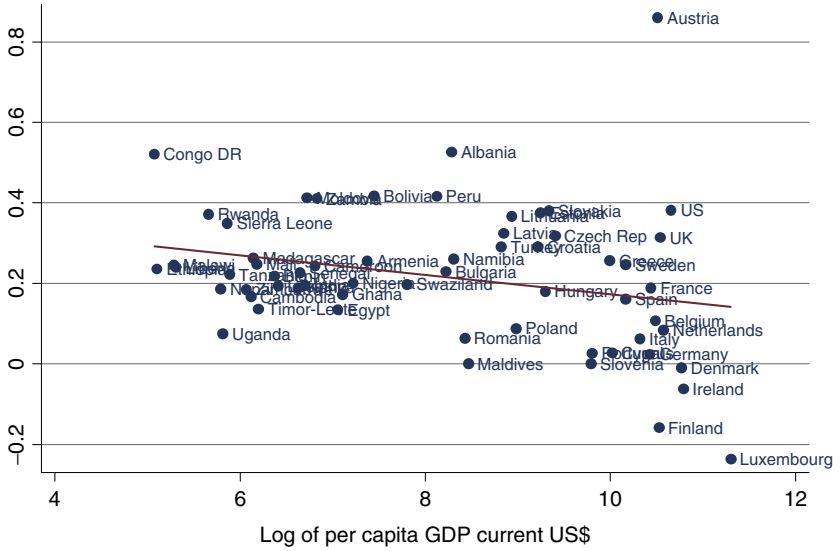


Figure 9 Coefficient of education on height by income.

more likely to be overweight or obese. In richer countries, where undernutrition is not very prevalent, there is no education gradient in undernutrition. In contrast, in these countries the prevalence of obesity is large and there is a large positive education gradient in obesity. This suggests that education is protective for the outcomes that are known to be bad for health.

The link between education and height is also clear. In all countries – even the richest – better educated people are taller than less educated people. The magnitude of the relationship is large throughout the world.

The link between education and other measures of health is much less clear. The correlation between education and smoking is nonlinear in income; the relationship between education and height or STIs is unrelated to income. These patterns demand a different explanation than a simple rich–poor dichotomy. They also suggest that the effect of education on health varies depending on the level of development, and holding GDP constant, on the specific health problems the country faces.

Understanding the Relationship between Education and Health

Education and health may be related for three reasons: poor health in early life may lead to less educational attainment; lower educational attainment may adversely affect subsequent health; or some third factor such as differences in discount rates may affect education and health-seeking behavior. Each of these pathways are briefly discussed.

The next section starts by describing the most important commonly unobserved determinants of both education and health. The first is parental resources: parents with more resources (broadly construed to include wealth, social networks, knowledge, etc.) will devote part of them to improving the survival of their children (by investing in their health) and also to improving their future outcomes, which, in turn, means they will invest perhaps more on their children's education. Second, there are some important individual characteristics that theoretically are expected to increase both education and health. *Ceteris paribus*, more patient individuals are more likely to invest more in both education and health. Also, smarter individuals might be more likely to obtain more schooling and also have better health.

Effect of Early-Life Health on Education

As the previous results indicate, there is a very strong correlation between early life indicators of health (such as height) and educational attainment – and this is true across all countries of the world. These correlations have been documented many times before, particularly in developing countries. As education is largely determined at young ages, this suggests that at least part of the correlation between education and health among adults is due to the fact that unhealthy children obtain few years of schooling and become unhealthy adults.

Recent studies show that the relationship observed – shorter (and sicker) children obtain less education – is a causal one. Two types of studies investigate the causal effect of health shocks on human capital accumulation: some take advantage of the so-called 'natural experiments,' whereas others use randomized controlled trials to investigate the question.

Most studies that investigate this causal chain find support for it: there are several examples of how disease and nutrition affect human capital formation. For instance, individuals affected *in utero* by the 1918 influenza pandemic obtained fewer years of schooling than those not affected. Individuals born during the Great Famine in China had lower educational achievement than those not born during the Great Famine. Malaria eradication in the US, various Latin American countries, and Sri Lanka resulted in greater education, although malaria eradication in India did not. Deworming campaigns had substantial effects on schooling in early-twentieth century American South and in Kenya today.

A related literature explores the consequences of birth weight on adult outcomes and finds similar results suggesting that those born with lower birthweights have lower levels of education, income, and health as adults. Although these studies do not directly look at nutrition, but rather at extreme events that influence birth weight, in many cases nutrition and disease are the most likely intervening mechanisms.

Direct evidence on the effect of nutrition and disease on education is available from several randomized experiments. Nutritional supplements, iron supplementation, and iodine supplementation trials *in utero* or during early childhood have resulted in higher educational attainment and increased cognitive ability.

Whether early life health affects education through morbidity at younger ages or expectation of life extension at older ages is unknown. Indeed, there is scant evidence on the extent to which expectations of longer life affects schooling but some evidence suggest that this channel also matters: when maternal mortality fell in Sri Lanka, girls' education increased (but not that of boys). But it is not clear as to what extent the education–life expectancy relationship is accounted by this channel.

Overall, the evidence is consistent in showing that nutrition and disease shocks early in life are quite detrimental for human capital formation. Interestingly, the reduction in educational attainment associated with early-life health insults is not the only theoretical possibility. Sickness increases the cost of going to school in terms of effort and might also lower the returns to school if it lowers life expectancy. Thus, parents of sick children might optimally choose lower levels of schooling for those children. However, illness also increases the cost of work and might increase the returns to school (in terms of avoiding more physically demanding jobs). Thus, it could be that the return to schooling increases as people become less healthy. However, there is no empirical evidence of this alternative, although perhaps it explains why education and height are negatively related in two very rich countries: Finland and Luxembourg (**Figure 9**). This discussion also underscores the fact that the observed relationship between height and education reflects not only the physical effects of disease in childhood but also the behavioral responses of

parents which might attenuate or exacerbate the effects of the health shock itself.

Given that health is an important determinant of schooling and the fact that education and health could simply be determined by common factors such as parental resources, it is extremely challenging to document whether in addition to these well-documented relationships, education itself affects health – this question is considered next.

The Effect of Education on Health: Theory

Theoretical foundations for a causal effect of education on health were first provided by the seminal work of Grossman (1972), based on the human capital model of Becker (1964). One key insight of Grossman's model of health capital is that individuals derive utility from health directly (they do not like being sick) and indirectly by affecting labor market outcomes (sick individuals work less and earn less). The other essential feature of the model is the recognition that there is a 'health production function' – that there are known factors that individuals (or institutions) can manipulate in order to affect health in predictable ways. These two features give rise to a behavioral model in which individuals demand medical care, food, and other goods and services because they are aware these factors will improve their health and ultimately increase their utility. (See Strauss and Thomas (2008) for an excellent exposition of the theoretical production of health over the life course and its determinants.)

In this type of model, education can affect health in a variety of ways. Most obviously, education affects the type of jobs that individuals get and the income they earn. A year of education raises income by at least 7%, and this is true in both developed and developing countries. Higher incomes increase the demand for better health, but they affect health in other ways as well. Richer people can afford gyms and healthier foods; they can also afford more cigarettes. Furthermore, when an individual's wage increases, it raises the opportunity cost of time: because many health inputs require time (such as exercise or doctor visits or cooking), in the short run, wage increases might reduce health. Thus, the income associated with higher education may or may not improve health.

Higher educated individuals are also more likely to take jobs that provide health insurance and other benefits such as retirement accounts. Although one expects these benefits to have a positive effect on health, it is theoretically possible that they do not. For example, individuals with insurance could be less likely to care for themselves because they face lower financial costs in the event of a disease. However, because the uncompensated costs of disease are large (morbidity and premature mortality), it is not expected that these indirect effects would dominate the access associated with better insurance.

Finally, more educated people work in different industries and occupations than less educated people. To the extent that job characteristics affect health, sorting into jobs may affect health as well. At the dawn of the industrial era, this relationship was undoubtedly positive. Early in the twentieth century, the more educated were more likely to work in white collar occupations, which were substantially safer than working in agriculture or manufacturing (fewer accidents, exposure to chemicals, physical strain, etc.). Today, most individuals work in the service sector and the better educated may have jobs that are worse for their health – for example, they spend more time sitting in front of computers, which could turn out to be bad: sitting (independently of exercise) has been recently shown to detrimental to health.

Thus, the effect of education on health, through its effect on the labor market, is ambiguous. Moreover, a positive association between education and disease can arise through the conscious choices of individuals: individuals may well know that exercise is needed to remain in good physical shape, but they may optimally trade off some of their health for increased incomes when wages are high. At the extreme, when individuals have no other resources than their bodies to earn a living, they will optimally 'use up' their bodies to earn a living: trading off higher lifetime earnings for shorter, sicker lives. The theory of compensating differentials predicts just that: individuals can be 'paid off' to accept risky occupations.

The second mechanism explored by Grossman is that education can affect the production function of health directly, acting as a 'technology' parameter. This is the so-called 'productive efficiency' mechanism, in contrast to the 'allocative efficiency' mechanism which has already been described (the more educated optimally chose different levels of health inputs because they face different prices and budget constraints). In its simplest formulation, productive efficiency posits that the better educated will have better health outcomes, even conditional on access to the same health inputs at the same prices. Better use of information is the classic example. More educated individuals might be better at following doctor's instructions (because they may have better self control for instance) or they might be more likely to believe the information produced by the scientific establishment and follow its recommendations perhaps (because they took science courses in school or know scientists directly).

Car safety knowledge provides another interesting case. Both more and less educated people strongly agree that one should wear a seatbelt while driving a car. But when the survey question is asked a different way, the pattern changes: the less educated are much more likely to agree with the statement that seatbelts are just as likely to harm as help you in an accident. It may be that better educated people have a deeper understanding of the risks of not wearing a seatbelt and the probabilities that go into a calculation of optimal seatbelt use. Another example concerns how successful individuals are at using certain health technologies such as devices to help quit smoking. Conditional on making an attempt to quit smoking, the better educated are more likely to be successful quitters.

There is a third theoretical reason why education could be related to health: education could change the 'taste' for a longer, healthier life. For example, education may lower individuals' discount rates, making them more 'patient.' There are two reasons for this. First, attending school per se is an exercise in delaying gratification, and school may teach patience; this may carry over into other aspects of life. Second, to the extent that individuals can 'choose' or learn what to like (in other words if discount rates can be chosen), then those with more education have a greater incentive to choose patience, because they face steeper income profiles over their lifetimes. The same argument might hold for risk aversion.

Finally, education affects the peers that individuals spend time with, and different peer sets may encourage different health behaviors. This is particularly important in the context of health, given that many health behaviors have an important social component. For example, individuals generally drink together and often smoke together. More generally, peers are thought to be essential in determining risky behaviors. Also, peers and social networks are an important source of information, and of financial, physical, and emotional support and hence can affect whether individuals get sick and how well individuals fare when they do. If on average more educated individuals have more educated peers, they will have access to a greater set of resources. If more educated individuals are more likely to be better informed (because they learned so in school or because they remain better informed later), then peers will help individuals reinforce their knowledge, in a 'multiplier' setting.

Note, however, that peers can influence behavior in a positive or negative manner. A peer group that focuses on sedentary lifestyles and lack long-term investment may encourage that same behavior among all members of the peer group, but one that focuses on exercise and fitness would promote the opposite.

Beyond the Grossman model, there are other theories that predict associations between education and health. The most prominent is that education predicts rank in society, and those with higher rank are in better health than those with lower rank. In small hierarchical groups such as apes and (perhaps) humans, those at the top will have access to more resources and greater control over their lives in general, whereas those at the bottom will have both fewer resources and control. As a consequence, those at the bottom will suffer more 'stress' and this, in turn, lowers immune responses and increases the likelihood of short-term illness and long-term chronic disease.

This theory has been shown to be accurate among mammals and other species (Sapolsky, 2004) and has been tested experimentally with animals to rule out genetic factors as the main explanation (e.g., the top of one hierarchy will suffer in health if they are transferred into a different group where they have a lower place in the hierarchy). Although it is not entirely clear whether and how this theory applies to humans in large modern societies – where reference groups are multiple and they are chosen endogenously – it provides another rationale by which education may affect health. It is to be noted that this theory has an interesting prediction: if all that matters is relative rank in society, a society with higher average levels of education may have no better outcomes than a society with lower average levels of education.

Education may also affect health because the things that kids do while in school are different than what they do outside of school. Although this is a trivial observation, this so-called 'incarceration effect' is extremely important to consider. For example, children in school may have less exposure to criminal activity or poor role models.

Finally, there are other possibilities. The more educated could inadvertently be better or worse off because of biological processes that are not well understood. For example, more educated women have higher mortality rates of cancers of the reproductive system. It has been hypothesized that this 'wrong' gradient emerges because more educated women have fewer children, and having children turns out to be protective from certain cancers. Overall, education appears to lower mortality even after all health behaviors are accounted for, which suggests that some of these nonbehavioral mechanisms might be important – although it is not obvious that all important health behaviors can be observed.

Certainly it is very likely that many of these mechanisms are at play at any one time and place and in combination they will yield complex patterns. The complex relationship between education and HIV in Africa is an interesting case in point – de Walque reports that "education predicts protective behaviors like condom use, use of counseling and testing, discussion among spouses, and knowledge, but it also predicts a higher level of infidelity and a lower level of abstinence." In this example it would appear as if the educated not only seek out information at higher rates, know more, and use their information and resources to purchase protection but also have some higher risky behaviors, perhaps because of their higher incomes or lower risk (they can 'afford' it).

Evidence on the Causal Effect of Education

A large number of early studies found supporting evidence for the Grossman model using largely descriptive tools. The usual prediction tested was that education and health were positively correlated. Clearly they are; the literature struggled with instruments for education to determine causality. However, these studies were not entirely convincing about whether education had a causal effect on health, because descriptive methods and imperfect instruments are not well suited to establishing causality.

A second generation of studies attempted to provide clearer evidence of a causal link between education and health again using 'natural experiments.' Many of these studies make use of compulsory schooling as a source of plausibly exogenous variation of education to investigate whether more school improves adult health. The intuition for this approach is simple: some individuals are forced to attend school longer because of compulsory school legislation, and researchers can examine whether the health of those who are forced to obtain more schooling improves compared with the health of those who are not required to stay in school. Studies in the US, Denmark, Sweden, the UK, and Germany using changes in compulsory schooling find that indeed these laws ultimately improved the health of the affected populations. However, recent work finds no effect of the same compulsory schooling laws on health in England and Sweden, and a study focusing on France also finds no effect of education on mortality.

The literature that has estimated the effect of education on health behaviors using natural experiments is also mixed. For example, some find that schooling lowers smoking rates but other studies find no evidence that schooling affects smoking behavior.

It is difficult to interpret this conflicting evidence. All of the papers that find positive effects of education on health use natural experiments to construct instrumental variables (IV) estimates of the impact of education. They tend to find effects that are larger than OLS. Although this has generally been interpreted as reflecting heterogeneity of treatment effects

(those that are affected by the legislation have larger returns), the alternative interpretation is that the 'natural experiment' did not in fact work well as an experiment, and there is still substantial bias in the education estimate. For example, the results for using compulsory schooling reforms in the US are not robust to the inclusion of state-specific trends. However, there is very little variation left once these controls are added, so it is not clear whether the effects are truly overestimated or whether the variation in the laws is not sufficient to estimate an effect of education. This discussion underscores the limitations of IV studies in general. From a methodological point of view, the regression discontinuity studies make the fewest assumptions, and they find no effects of education on health.

Also interesting to note is that available studies report impacts along different margins, not only because of the obvious reason that they study different times and places but also because the 'experiments' themselves are different. In the UK, the changes in compulsory schooling were strictly followed and an entire cohort of individuals was forced to obtain almost 1 more year of schooling as a result. In contrast, in the US, the laws that are typically studied increased educational attainment by 0.05 of a year – that is only 1 in 20 individuals obtained one more year of schooling. There are two important differences here. First, the affected population in the US is a small sample among those that were potentially affected – it is indeed possible that returns are different for this subset. Second, in the US, only a few individuals in a given cohort and place were affected, but entire cohorts were affected in the UK. If, for example, education matters because it affects a person's rank in society, then in the US, those who stayed in school had their rank increased relative to the counterfactual of no compulsory schooling law. This would not necessarily have been the case in the UK: an entire cohort increased their education by approximately 1 year, so an individual's rank within their cohort was unaffected by the policy.

It is also theoretically possible that the effect of education varies over time and place, and that the results from the previous studies correctly document this variation. Indeed, the international evidence suggests that the returns to education do vary across countries. It is notable that the two studies that find no effects of education in the UK and France, study cohorts during and shortly after World War II (WWII), a time when the income returns to education were falling and generally low.

The fact that the effect of education on labor market earnings itself is causal also suggests a positive effect of schooling: if schooling is rewarded in the labor market because it raises productivity, how does it do so? Whatever general human capital is learnt in school and rewarded in the labor market might also be useful in the production of health, because it is useful in the production of goods. If education makes workers better by making them better decision makers or better able to deal with complexity or uncertainty, then these abilities can be used in other domains, in particular for health.

One central conclusion of this discussion is that investigating the specific mechanisms by which education affects health would improve the understanding of education–health link substantially. The following paragraphs discuss what is known about this next, after describing the latest attempts to infer causality in the literature.

In addition to natural experiments described at the beginning of this section, there are a variety of experimental interventions that have been carried out, mostly in developing countries, that can be used to infer the effect of education on health. In Kenya, random distribution of school uniforms – a significant cost associated with school – among upper primary-level students increased levels of schooling for both genders by a substantial amount (the dropout rate fell by 18%). Seven years later, treated girls had significantly lower rate of marriage and pregnancy, but the treatment had no effect on sexually transmitted diseases. However, random provision of HIV information to the curriculum of some students had no effect on sexually transmitted diseases, but the rate of unwed teenage pregnancies fell.

Many countries have implemented conditional cash transfers programs to help the poor. Conditional cash transfers are transfer programs where the receipt of income is conditional on certain behaviors, generally related to health or schooling. Unconditional cash transfers do not have any strings attached. Studies find that the conditional cash transfer programs have resulted in lower levels of sexual activity, teen pregnancy, and marriage rates among young girls in the short term, in addition to increasing schooling in Africa.

Although curriculum information on HIV in Africa had little effect on schooling, other information campaigns have worked. For example, a small intervention in the Dominican Republic informed 14-year-old boys about the labor market returns to school. The intervention successfully increased schooling by 0.2 years, and significantly decreased work in the formal labor market. As a consequence of this, treated boys delayed debut of heavy drinking and were less likely to smoke than untreated boys.

These studies suggest that education affects specific health behaviors, but not all behaviors. However, even here, it is not clear that one can infer that education is the ultimate cause of the changes in the observed health behavior. The gold standard for establishing causality would call for randomly assigning individuals to various levels of education. Clearly, this approach is unethical and unfeasible. Instead, these studies look at an 'intent-to-treat' intervention, where individuals are randomly 'incentivized' to obtain different levels of education. With this design, it is possible to estimate the effect of education on health, if (1) the intervention successfully raises education levels and (2) the random incentives that are provided to increase schooling affect health only through education (the exclusion restriction assumption).

In this light, consider whether randomized interventions that potentially raise schooling can be used to estimate the causal effect of schooling. Typically, interventions are designed so that reasonably sized effects on education can be detected with the chosen sample. But even if this requirement is met and the intervention increases education levels, the intervention must induce students to attend school but not directly or indirectly impact any other determinant of health. It is difficult to design an intervention that meets this assumption. Providing scholarships to those that are credit constrained is equivalent to increasing income in the short run, which directly or indirectly is likely to affect health. Providing uniforms

is not quite like providing income, but it increases incomes indirectly by substituting for household spending. The more constrained individuals are in their consumption, and the higher the effect of the intervention on schooling is, the more likely it is that the income effects of the intervention are large. Finally, informing misinformed students of the returns to school affects the present discounted value of earnings of all participants, regardless of whether they are induced to attend school or not. Because health (and its determinants) is likely to depend on permanent rather than temporary (current) income, this intervention also fails the exclusion restriction assumption.

Another important limitation of randomized interventions is that in the short run schooling is not expected to affect health because the young are generally in excellent health and because health is a stock – instead it is expected that the health effects emerge slowly and cumulate. But it is difficult and expensive to follow individuals for many years; the interventions above follow individuals for several years but on average the participants are still quite young at the last follow-up (e.g., in the Dominican Republic study, the intervention takes place when boys are 14 years old and they are 18 years old when they are last interviewed). The interventions then look at health behaviors, but it is not clear how these effects will eventually translate into, for example, mortality.

There are only two studies of randomized education interventions that follow individuals over a long period of time. One looks at the participants Perry Preschool School program (PPP) 37 years later and the other looks at the participants of the Carolina Abecederian (ABC) Project at the age of 21 years. Both of these interventions occurred early in childhood, and they have been shown to have had persistent effects on wages and other outcomes. The results from these two studies are again in conflict: the treated students in the ABC program had significantly better health than the controls, but that was not true in the PPP program, although in both cases the treated appear to have better health behaviors. These results are to be taken with caution as in both cases the number of observations consists of only approximately 100 individuals.

Thus, simple randomized trials cannot conclusively answer the question of whether education affects health. But it is possible to make progress on this question by investigating mechanisms through which interventions affect education or designing more complex randomized interventions. The authors discuss the evidence on mechanisms next and conclude with a series of observations on what questions could be explored in future research.

Evidence on the Mechanisms Linking Education and Health

To be convincing, studies of the effect of education on health will need to understand the pathways that link the two. Because there are a large number of potential mechanisms, this is a difficult task. In addition, the evidence on mechanisms is somewhat weaker than the evidence on causality, because often assumptions about what constitutes a mechanism have to be made.

Some studies have attempted to look at why education matters for health. Consider the evidence on the effect of education on sexual behaviors and fertility. An important reason why education improves outcomes for girls is that it delays marriage and fertility, because the common practice is for girls to marry soon after finishing school. This, in turn, means girls will have fewer years of 'exposure' to get pregnant, and thus fewer children over their lifetime. Also girls in school have children later, which is beneficial because reproduction during the early teenage years is riskier for the health of the mother and the infant compared with reproduction in prime adult years.

The results from the randomized trial in the Dominican Republic also seem to be driven in part by the incarceration effect: most boys who are not in school start working or are idle – the set of people whom they interact with when they are not in school is different from their peers in school, and 'treated boys' (those given the message about the value of education) report that their peers are significantly less likely to drink and smoke. Note further that early exposure to a different set of peers could have important long-term consequences, as smoking and drinking are addictive behaviors that affect youth's physical and mental development.

Consider now the natural experiments that use compulsory schooling as an instrument for education. In the US in the 1910s, children who were not in school were either idle or working. The main occupation for children of ages 10–15 years at the time was agriculture. Agricultural work is substantially more hazardous to health than school work. Thus, it is possible that the health effects of forcing children to stay in school during this period are driven by the difference in health hazards across environments. However, by the 1940s the types of jobs adolescents engaged in when they were not in school were substantially different, and perhaps not as hazardous. This may explain why the returns to post-WWII compulsory education in the UK were smaller.

However, the evidence suggests that the effect of education is not limited to this incarceration effect alone. Uniform provision in Kenya delayed marriage well beyond the increase in years of schooling generated by the intervention, so at least in this case, incarceration alone cannot explain the observed effects.

Another possibility is that education matters (sometimes) for health because schools directly provide information on how to improve health, and it is the health information itself, rather than being in school that affects behaviors. More educated individuals are indeed better informed about health risks in developed countries. And when information first becomes available, it seems to first become known to the more educated, who, in turn, seem to be the first to respond. Educated mothers stopped smoking at higher rates after the 1964 Surgeon General Report first widely publicized the harms of smoking, and their babies' health increased more as a result. Smoking rates started declining for the best educated in the 1950s, before the Surgeon General's report, as the dangers of smoking were increasingly discovered. Similarly in Uganda in 1990, there was no relationship between education and HIV, but one emerged by 2000 after a decade of information campaigns on prevention. In the UK, when information was first (incorrectly) reported about possible autism risks

associated with the mumps, measles, and rubella (MMR) vaccines, vaccination rates fell more in areas with more educated individuals. In fact, in some studies it appears as if all of the effect of education is explained by information, for example, studies find that most of the effect of maternal education on child height can be explained by differences in information.

But information cannot be the whole explanation; differences are observed in health behaviors by education even when there are no differences in information by education. For example, in the experiment in the Dominican Republic that informed children on the returns to school, there were no differences in the extent to which smoking and drinking were perceived as harmful by the treated and the control boys, and yet the treated boys stayed in school longer, smoked less, and drank less. Similarly, in developed countries today, knowledge of the harms of smoking is nearly universal, and although there are some small differences by education in knowledge, these differences are very small compared with the differences in smoking rates by education. Curriculum interventions alone had little impact on behavior in the Kenyan intervention. Finally, observational studies suggest that a small portion of the effect of education on behaviors is due to differences in knowledge. It appears that when knowledge first becomes available on how to improve health, it substantially increases education disparities. But in the long run, information diffuses and other factors are more important in explaining the associations between education and health.

In this sense, information may be like other innovations in health. For example, more educated individuals are more likely to use recently approved drugs than the less educated, and this appears to be driven by those with chronic conditions who use drugs repeatedly, suggesting that learning is an important component of the education effect. Similarly, in developing countries, more educated individuals are generally more likely to adopt new innovations. Whether the initial advantage of the educated fades away or gets stronger with time, might, in turn, depend on the type of health technology. For example, some medication regimes are difficult to adhere to, and the educated might have a permanent advantage at using them – this is the case for diabetes type 1. Other innovations instead are 'deskilling,' such as the birth control pill, in which case eventually the less educated catch up. The results from malaria interventions provide some interesting evidence on this point: when access to malaria treatment improves, the gap in access between the educated and the uneducated falls. However the educated still behave quite differently from the uneducated in their treatment-seeking behaviors: they appear to be more likely to know the likelihood that they have malaria and they are more likely to visit a health-care center and less likely to use other treatments when their symptoms are worse. This is not true among illiterate individuals.

The evidence from randomized interventions suggests that some mechanisms are important, whereas others are not – but certainly as this paper discussion suggests the extent to which any findings are generalizable is not clear. Some of the effects of schooling might operate through the incarceration effect as already discussed. Another important mechanism is income, as the Malawi conditional cash transfer intervention suggests. Finally peers are also important. In the Dominican Republic intervention discount rates, risk aversion and health information were not affected by the intervention, even when schooling increased. However, treated boys had lower incomes and reported that their peers drank and smoked less – these two channels most likely explain the observed decreases in smoking and drinking among the treated.

Interestingly, this evidence is consistent with the exploratory and descriptive studies. Rough calculations from these suggest that observed factors can account for approximately 70% of the effect of education (in a statistical sense), through resources (30%), family and friends (10%), and information (10%) and cognition (20%). However, risk aversion, discounting, stress, and other personality traits did not appear to mediate the relationship between education and behaviors – although the noise in these measures gives one some pause.

Summary

On balance, the literature reviewed highlights a wealth of interactions between education and health. Education appears to be causally related to health in many settings, but not always, and the reverse is true as well.

Equally important, this review highlights some unanswered issues. The most important issue is to understand in more detail when and how education translates into health. To what extent is education associated with specific knowledge, with cognitive ability in general, or with different social settings, either during school or after? Some evidence on this may come from looking at the quality of education individuals receive. Most of the literature has looked at the impact of additional years of schooling. Yet many of the theories say that the quality of the years should matter as well. This has not been explored in any great detail.

Simple experimental designs that randomly encourage individuals to obtain schooling can be useful in providing further evidence of causality on health and health behaviors, but they cannot conclusively answer the question of whether education alone is responsible for the observed effects because in general it is difficult to satisfy the exclusion restriction that is needed to reach such conclusions. However, more sophisticated designs could be implemented to help identify mechanisms and causality both. For example, one could design an experiment with three treated groups, where individuals are given unconditional cash transfers (cash-only group), conditional cash transfers if they attend school (attendance group), and conditional cash transfers for both going to school and obtaining good grades (performance group). Under the assumption that all treatments induce changes in education, income, and grades, the separate effects of education, income, and health can be learned. By comparing the controls with the cash-only group one can estimate the effects of income on health and health behaviors. By comparing the outcomes of the cash-only group and the attendance group one can obtain an estimate of the effects of attendance. Finally, by comparing the performance group and the attendance group one can learn about the effects of education content.

Furthermore, it is vitally important to understand the translation from intention into action. In developed countries,

everyone knows the behaviors that are good for health and (as suspected) many would like to improve their health. Yet people systematically fail at this task, that is, they struggle to change their behaviors. How are these failures understood, and what types of interventions would reduce them? In a way, this is asking for a benchmark by which to compare education. Improving health by inducing more education is costly; many people do not enjoy schooling, and forcing additional years of schooling comes at a price. If the impact of education on health can be replicated using other methods, this would be very attractive.

In sum, the burgeoning literature on education and health is just the beginning. A review written a decade from now will ideally have many more specific conclusions to draw.

Data Appendix

DHS Surveys

The authors selected 31 countries with either a DHS-IV or a DHS-V survey that includes data on a woman's anthropometry (height and weight), education level, and her drinking or smoking habits. All surveys contain nationally representative samples of ever-married women between the ages of 15 and 49 years.

Height is the respondent's height in centimeters. BMI is computed as weight (in kilos) divided by height (in meters) squared. Underweight is equal to 1 if the person's BMI $\leq$ 18.5; obese is equal to 1 if the person's BMI $\geq$ 30. Anemia is coded 1 if the person is anemic at all, irrespective of the level of anemia (slight, moderate, and severe). Hemoglobin is the individual's hemoglobin level in g/dl adjusted for altitude. Anemia and hemoglobin were considered unknown if hemoglobin levels were less than 5 or greater than 50. If the adjusted hemoglobin level was not available, the unadjusted level was used. Smoke is coded 1 if the individual has currently smoked, 0 if not. STI is equal to 1 if the individual had a STI in the past 12 months. Drink is a binary variable if the individual has ever or recently consumed alcohol (this varies by country).

Regressions control for age, age2, education, married, religion dummies, and ethnicity dummies. Age and education are measured in years. Religion and ethnicity dummies are country specific. Marital status is 1 if the woman is married or living with a partner as if married, and 0 otherwise. All means and regression coefficients were computed taking survey design into account, unless strata or sample weights were not provided by the survey.

Eurobarometer Data

Our European data are drawn from two waves of the Standard Eurobarometer. Women's anthropometry (height, weight, BMI, and probability of being underweight or obese) are drawn from Eurobarometer 64.3, which was collected in November–December 2005. All other outcome variables of interest (alcohol consumption, smoking, physical activity and sport, and fruit consumption) are drawn from Eurobarometer 72.3, which was collected in October 2009. Both surveys contain nationally representative samples of women between the ages of 15 and 49 years in 29 European countries.

Height is the respondent's height in centimeters. BMI is computed as weight (in kilograms) divided by height (in meters) squared. Underweight is equal to 1 if the respondent's BMI $<$ 18.5; obese is equal to 1 if the respondent's BMI $\geq$ 30. Currently, smokes is equal to 1 if the respondent currently smokes, and is 0 otherwise; consumed alcohol in past year is equal to 1 if the respondent has consumed any alcoholic beverages in the past 12 months.

Regressions control for age, age2, education level, and marital status. Age is measured in years. Marital status is 1 if the woman is married or living with a partner, and 0 otherwise. Education level is the age at which the respondent left school, in years. All means and regression coefficients were computed using the poststratification weights provided with the surveys.

Behavioral and Risk Factors Survey for the United States

For the US, the authors use the 2005 wave of the Behavioral and Risk Factor survey, which contains height, weight, drinking, and smoking. Only women of ages 15–49 years are included.

Height is the respondent's height in centimeters. BMI is computed as weight (in kilograms) divided by height (in meters) squared. Underweight is equal to 1 if the respondent's BMI $<$ 18.5; obese is equal to 1 if the respondent's BMI $\geq$ 30. Currently, smokes is equal to 1 if the respondent currently smokes, and is 0 otherwise. A person is said to drink if they drank any alcohol in the past 30 days.

Regressions control for age, age2, education level, and marital status. Age is measured in years. Marital status is 1 if the woman is married or living with a partner, and 0 otherwise. Education level is measured in years of school. Race and ethnicity dummies are included. All means and regression coefficients were computed using the poststratification weights provided with the surveys.

GDP Data

The GDP per capita data come from the World Bank, using the GDP per capita (current US$) indicator. When the data set comes from a survey taken over multiple years, the GDP per capita figure is the mean during that period.

Acknowledgment

The authors are grateful to Pascaline Dupas and John Strauss for comments, to John Min and Tisa Sherry for excellent research support, and to the National Institutes on Aging for research funding.

See also: Alcohol. Education and Health in Developing Economies. Peer Effects in Health Behaviors. Smoking, Economics of

References

Becker, G. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. Chicago: University of Chicago Press.

Grossman, M. (1972). *The Demand for health – A theoretical and empirical investigation*. New York: National Bureau of Economic Research.

Sapolsky, R. M. (2004). *Why zebras don't get ulcers. An updated guide to stress, stress-related diseases, and coping*, 3rd ed. New York: Freeman.

Strauss, J. and Thomas, D. (2008) Health over the life course. In Schultz, T. P. and Strauss, J. (eds.), *Handbook of development economics*, vol. 4, pp 3375–3474. Amsterdam: North Holland Press.

Further Reading

Cutler, D. M. and Lleras-Muney, A. (2008). Education and health: Evaluating theories and evidence. In Schoeni, R. F., House, J. S., Kaplan, G. A. and Pollack, H. (eds.) *Making Americans Healthier: Social and Economic Policy as Health Policy*, pp 37. New York: Russell Sage Foundation.

Cutler, D. M. and Lleras-Muney, A. (2010a). Understanding differences in health behaviors by education. *Journal of Health Economics* **29**(1), 1–28.

Grossman, M. (2000). The human capital model. In Culyer, A. and Newhouse, J. (eds.) *Handbook of health economics*, vol. 1A, pp 347–408. Amsterdam: North Holland.

Kitagawa, E. M. and Hauser, P. M. (1973). *Differential mortality in the United States: A study in socioeconomic epidemiology*. Cambridge, MA: Harvard University Press.

Strauss, J. and Thomas, D. (1998). Health, nutrition, and economic development. *Journal of Economic Literature* **36**(2), 766–817.

Strauss, J. and Thomas, D. (1995). Human resources: empirical modeling of household and family decisions. In Behrman, J. R. and Srinivasan, T. N. (eds.) *Handbook of development economics*, vol. 3A, pp 1883–2023. Amsterdam: North Holland Press.

Education and Health in Developing Economies

TS Vogl, Princeton University, Princeton, NJ, USA, and The National Bureau of Economic Research, Cambridge, MA, USA

Glossary

Fetal origins hypothesis A theory that posits an effect of *in utero* conditions on later life health and socioeconomic outcomes.

Human capital The stock of productive knowledge and skills embedded in an individual, most commonly measured as educational attainment.

Intergenerational link Effect running between generations within a family, for instance from parents to their children.

Intragenerational link Effect operating within an individual, either instantaneously or over time.

Longitudinal data Data following individuals over multiple time periods.

Natural experiment An observational study design in which individuals (or groups of individuals) are assigned to a treatment by a mechanism that mimics randomization but is outside the researcher's control.

Randomized controlled trial An experimental design in which researchers randomly assign individuals (or groups of individuals) to receive a treatment.

Introduction

In the course of development, few processes are as intertwined with economic growth as human capital accumulation. Schooling makes workers more productive, speeds the development of new technologies, and better equips parents to raise skilled children, all of which promote economic growth. Growth, in turn, incentivizes investment in human capital. Causal links point in every direction, traversing phases of the lifecycle as well as generations.

However, the entangled role of human capital is not limited to aggregate income growth. Education exhibits complex dynamic relationships with several components of wellbeing, including health. For example, education affects health in adulthood; life expectancy affects educational investment in childhood; and the health and education of parents – particularly mothers – affect both outcomes in their children. Just as with income, these relationships are likely to be especially important in developing countries, where levels of both schooling and health are low but have risen rapidly over the past half-century.

This article gives an overview of the current knowledge on the relationships linking health and education in developing countries. To emphasize the dynamic aspects of these relationships, the article will trace them out first within a generation, between childhood and adulthood, and then across generations, from parents to children. It will focus on reduced-form evidence of these effects rather than efforts to precisely pin down mechanisms, for two reasons. First, the existing literature focuses on reduced-form evidence. Mechanisms have received some attention, but the evidence comes mainly from rich nations; even that evidence remains sparse.

Second, the reduced-form evidence on dynamic links casts in stark relief the potential joint role of education and health in accounting for the intergenerational persistence of disadvantage. That is to say, the children of unhealthy and uneducated parents grow up to be unhealthy and uneducated parents themselves. Others have proposed similar arguments about the intergenerational dynamics of the relationship between health and socioeconomic status, more broadly construed. But the links between education and health, which typically lie at the crux of these arguments, can by themselves account for the dynamics. Given the current extent of inequalities in income, human capital, and health in developing countries, the links between education and health may prove important in shaping long-term trends in the levels and distributions of both variables.

Associations between health and education are not new, but with such tangled causal pathways, these associations sometimes prove to be uninformative. The recent literature in economics has made its main contribution in causal inference. Analyses of natural experiments and prospective trials have shed new light on long-standing hypotheses. They have also improved our ability to interpret careful associational studies, which are in many cases more generalizable than experimental studies but less internally valid. These advances have been key to identifying both the direction and the timing of effects in the causal system linking education and health. With this better understanding of what matters and when, policymakers will be better equipped to identify opportunities for well-targeted policies.

Mapping the Relationship between Education and Health

With its numerous pathways, the causal system linking education and health may seem convoluted. However, one can represent it in a simple but informative diagram. Figure 1 traces out the links between education and health, first over the lifecycle and then across generations. Each arrow represents a causal link that has empirical support in the literature. The blue lines signify intragenerational links – in other words, causal links that operate within a single person – whereas the red lines correspond to links that work across generations within a family.

 doi:10.1016/B978-0-12-375678-7.00109-7

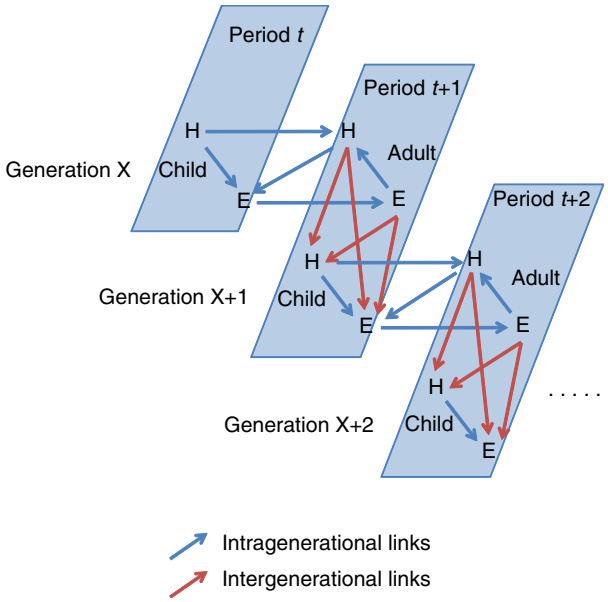


Figure 1 Causal links between health and education.

The system lays out a roadmap for the rest of the article. In childhood, good health improves educational outcomes. Additionally, the expectation of good adult health increases schooling investments in childhood. Both health and education persist from childhood to adulthood, at which point education boosts health. But adults are also parents, so their circumstance in middle age spills over onto the next generation. Healthier mothers have healthier and more educated children. Conversely, parental education promotes both the health and the education of the next generation. At this stage, the causal system repeats in the next generation. In the remainder of the article, the focus will be on the subset of the arrows in Figure 1 that connect health and education.

Intragenerational Links

Effects of Childhood Health on Educational Outcomes

Educational outcomes in childhood

The author begins with childhood, where abundant evidence suggests that health affects school enrollment and academic achievement. Health enables children to travel to school, concentrate, and think clearly, all of which may improve educational outcomes. Until recently, the evidence has primarily taken the form of cross-sectional associations between children's health and their educational outcomes. Many have critiqued these studies for inadequately addressing issues of causality and omitted variables.

Starting in the mid-1990s, a few analyses have made some headway on these issues by focusing on within-family variation. One early study in this literature analyzes data from Ghana and finds, in models with family fixed effects, that shorter siblings start school later than their taller brothers and sisters. A more recent article analyzes twin pairs and sibling sets in Chile, showing that twins or siblings born at higher birth weight perform better on exams. Within-family comparisons of this type eliminate concerns about family-level omitted variables, although they leave some concern about how parents allocate scarce resources among children with observably different health.

Analyses of natural experiments in disease eradication, micronutrient supplementation, and health care provision have also made progress on causal identification. One innovative study investigates the eradication of hookworm from the US South in the early twentieth century, finding that areas with higher initial hookworm burdens, and thus likely experienced larger declines in worm prevalence, saw larger increases in school enrollment. Another uses contemporary data from Tanzania, focusing on a maternal iodine supplementation program in Tanzania. Drawing on policy variation across time and space, as well as on sibling differences in program exposure, the study finds that *in utero* exposure to the program increased school participation. In addition to these effects on school participation and enrollment, early-life health boosts test scores. Using data from Chile (and Norway), a recent study takes advantage of the fact that infants born just below the threshold for very low birth weight (VLBW) receive much more care than those born just above. The study documents discontinuities around the VLBW threshold in both infant mortality rates and subsequent test scores, such that infants born below the threshold do better.

In addition to these innovative ways to glean causal effects from observational data, the past decade has seen a series of randomized controlled trials testing the effect of child health on schooling outcomes. Perhaps the best known is a deworming experiment in Busia district, Kenya. Intestinal worms cause anemia and other ailments, which may make children too weak or lethargic to study. After researchers experimentally varied access to deworming medications across 75 primary schools in the district, pupils in treatment schools exhibited significantly lower rates of worm infection, anemia, and school absence, although not test scores. Experimental data on other programs, including one that distributed iron supplements and deworming medication to Indian children and one that distributed protein supplements to Kenyan children, provide corroborating evidence.

Educational outcomes in adulthood

The fact that education is relatively fixed by adulthood facilitates the study of its relationship with health. Coupled with retrospective measures of child health, data on adult educational attainment can shed light on the effect of health on education in childhood. For example, just as height and schooling outcomes are associated in children, so too are they related in adults. Adult height positively predicts educational attainment in nationally representative data from Mexico, as well as in data on urban populations in Barbados, Mexico, Cuba, Uruguay, Chile, and Brazil.

In adulthood, too, the results of natural experiments and randomized controlled trials suggest that the associations partly represent an effect of health on education. One noteworthy finding comes from long-term follow-up of the deworming experiment in Kenya. When observed in young adulthood, individuals in the treatment group had stayed enrolled in school longer and performed better on a battery of tests than their counterparts in the control group. However,

long-term follow-up of hookworm eradication in the South US gives different results. If one compares birth cohorts born too early to be exposed to eradication to those born later, across areas with differing baseline worm infection prevalence, the results imply significantly positive effects on literacy but not years of schooling.

Several articles have used a similar strategy to estimate the long-term effects of malaria eradication on human capital, with mixed but on net positive results. One study draws on data from the South US, Brazil, Colombia, and Mexico. Here again, significant effects emerge for literacy but not years of schooling, which the author interprets as evidence that eradication made children more productive as students and as child laborers. Separate analyses have applied the same research design to men and women in India, as well as women in Paraguay and Sri Lanka. Although the Indian data show no evidence of positive effects on either literacy or years of schooling, the Paraguayan and Sri Lankan data show the opposite, with large gains in both outcomes.

Effect of Life Expectancy on Investment in Education

Unlike the effect of child health on education, which is rooted in the technology of skill formation, the effect of life expectancy on human capital investment is, at its core, about optimizing choices by households and individuals. According to the standard reasoning, if an individual expects a longer time horizon to reap the returns to human capital, then that individual will invest more. Analyses of macroeconomic data offer limited support for this hypothesis. Although adult mortality is negatively associated with secondary school enrollment, the relationship is not robust to the inclusion of covariates. However, given the paucity of high-quality data on adult mortality in most countries and the difficulty of assessing causality from cross-country associations, the macroeconomic patterns are suggestive.

Indeed, two microeconomic analyses have yielded convincing evidence that reductions in adult mortality risk increase human capital investment. One novel study uses a period of rapid decline in maternal mortality in Sri Lanka as a natural experiment in adult mortality. Parts of the country with higher baseline maternal mortality rates (and therefore larger subsequent declines in maternal mortality) saw larger increases in female educational attainment. A second study, analyzing the human immunodeficiency virus (HIV)/acquired immune deficiency syndrome (AIDS) epidemic in Africa, shows that the subnational regions that were hardest hit by the epidemic have also experienced the largest declines in education.

Effect of Education on Health in Adulthood

A long-standing literature reports positive associations between education and health in adults in wealthy countries, although the mechanisms linking the two variables are not fully known. To the extent that the association reflects an effect of education on health, important mediators of this effect may include income, working conditions, health-related knowledge, cognitive ability, patience, attitudes toward risk, and cultural capital (especially in interactions with health providers). Similar associations are evident in data from developing countries, although studies are rarer.

Both natural experiments and prospective trials suggest that although education can affect health, such effects may depend on characteristics of the population and the material being taught in school. Several studies use compulsory schooling laws in the US and Europe as instruments for education, with mixed but mildly positive results; some indicate positive effects on health and longevity, whereas others indicate no effect. Unfortunately, no similar studies exist on developing countries.

However, longitudinal follow-up of the recent spate of education-related randomized controlled trials in developing countries has begun to yield useful results on health behavior in young adulthood. One such study analyzes a program in the Dominican Republic that gave teenage boys information about the return to schooling. The information led the boys to stay in school longer, to delay the onset of heavy drinking, and to reduce smoking at the age of 18 years. Across the Atlantic in Africa, another study estimates the effects of a program that sought to provide adolescent girls with both vocational training and information about risky health behaviors. HIV-related knowledge and condom use both increased. However, less promising results have emerged from a Kenyan study on the medium-run impacts of a school subsidy program. Although the program increased schooling for both boys and girls, follow-up data show at best weak impacts on sexual behavior and sexually transmitted disease infection. Together, these studies suggest that keeping boys 'off the streets' and equipping girls with health information may be key to any effect of education on health in young adulthood.

Intergenerational Links

Effect of Parental Education on Child Health

In the context of poor countries, by far the most widely studied education-health association is that between maternal education and child health. Following a canonical study of child mortality in Nigeria in 1979, a large literature has emerged on this topic. The literature bares widespread correlations between maternal education and child health, measured by illness, anthropometry, or death.

Several studies question the extent to which the correlation reflects a causal effect running from maternal education to child health, as opposed to omitted variables. The relationship is not always robust to the inclusion of socioeconomic and community-level covariates, or to the inclusion of a fixed effect for the mother's sibship or for a multifamily household. However, one could interpret many of the socioeconomic and community-level covariates in the literature as mediators rather than confounders, and the inclusion of fixed effects exacerbates problems related to measurement error. The results of the revisionist literature are therefore inconclusive.

Analyses of natural experiments support a causal interpretation. The most compelling evidence comes from the US, where local college openings improve birth weight and gestational age. But some results are also available for

developing countries. Among Indonesian women, for example, exposure to a school construction program in childhood reduced mortality rates among their children.

Effect of Parental Health on Child Education

Parental health also affects children's schooling outcomes. Two mechanisms stand out in the literature. The first is indirect: Healthier mothers have healthier children, who in turn become better-educated adults. For instance, *in utero* exposure to the 1918 influenza epidemic decreased educational attainment for the cohort born in 1919 in the US, Brazil, and Taiwan. This effect supports the 'fetal origins hypothesis,' which posits that *in utero* conditions are crucial for the later health and skill development of her child.

The literature also highlights a second mechanism through which parental health affects child education: parental death. Good evidence comes from the HIV/AIDS epidemic, which has orphaned more than 15 million children, some 90% of them in Africa. Across Africa, orphans have lower school enrollment rates than the biological children of their caretakers. Furthermore, in South Africa and Kenya, the timing of parental death is associated with the timing of school dropout. The same is true in Indonesia, where parental deaths typically have little to do to HIV/AIDS. One can thus view the African results as representing a more general effect of losing a parent. Nevertheless, given the scope of the continent's orphan crisis, the results are most relevant there.

Open Questions

The existing literature fills in many of the links sketched in Figure 1, but open questions remain. For one, the distinction between aggregate and individual educational attainment has received little consideration but is almost certainly relevant for health systems in developing countries. How important is a country's education system in producing health professionals to support its health system? Additionally, the potential for the backwards intergenerational transmission of health information – from children to parents – remains underexplored. Such information transmission could prove useful in combating the rise of smoking and obesity in poor countries. Concerning intergenerational dynamics in the other direction, from parents to children, the literature would benefit from more focus on how parental behavior reinforces or compensates for exogenous changes in the health environment or educational opportunity. This last line of inquiry would put behavior back in the center of economic research on health and education.

See also: Education and Health. Education and Health: Disentangling Causal Relationships from Associations. Health Care Demand, Empirical Determinants of. Health Status in the Developing World, Determinants of. Intergenerational Effects on Health – *In Utero* and Early Life. Nutrition, Health, and Economic Performance

Further Reading

Alderman, H., Behrman, J. R., Lavy, V. and Menon, R. (2001). Child health and school enrollment: A longitudinal analysis. *Journal of Human Resources* **36**(1), 185–205.

Almond, D. and Currie, J. (2011). Human capital development before age five. In Ashenfelter, O. and Card, D. (eds.) *Handbook of labor economics*, vol. 4A, pp 1315–1486. Amsterdam: Elsevier – North Holland.

Bharadwaj, P., Løken, K. V. and Neilson, C. (2013). Early life health interventions and academic achievement. *American Economic Review* **103**(5), 1862–1891.

Bleakley, H. (2007). Disease and development: Evidence from hookworm eradication in the American South. *Quarterly Journal of Economics* **122**(1), 73–117.

Bleakley, H. (2010). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics* **2**(2), 1–45.

Caldwell, J. C. (1979). Education as a factor in mortality decline an examination of Nigerian data. *Population Studies* **33**(3), 395–413.

Cleland, J. G. and Van Ginneken, J. K. (1988). Maternal education and child survival in developing countries: The search for pathways of influence. *Social Science and Medicine* **27**(12), 1357–1368.

Cutler, D. M., Fung, W., Kremer, M., Singhal, M. and Vogl, T. (2010). Early life malaria exposure and adult outcomes: Evidence from malaria eradication in India. *American Economic Journal: Applied Economics* **2**(2), 196–202.

Cutler, D. M. and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics* **29**(1), 1–28.

Desai, S. and Alva, S. (1998). Maternal education and child health: Is there a strong causal relationship? *Demography* **35**(1), 71–81.

Field, E., Robles, O. and Torero, M. (2009). Iodine deficiency and schooling attainment in Tanzania. *American Economic Journal: Applied Economics* **1**(4), 140–169.

Fortson, J. G. (2011). Mortality risk and human capital investment: The impact of HIV/AIDS in sub-Saharan Africa. *Review of Economics and Statistics* **93**(1), 1–15.

Jayachandran, S. and Lleras-Muney, A. (2009). Life expectancy and human capital investments: Evidence from maternal mortality declines. *Quarterly Journal of Economics* **124**(1), 349–397.

Lucas, A. M. (2010). Malaria eradication and educational attainment: Evidence from Paraguay and Sri Lanka. *American Economic Journal: Applied Economics* **2**(2), 46–71.

Miguel, E. and Glewwe, P. (2008). The impact of child health and nutrition on education in less developed countries. In Schultz, T. P. and Strauss, J. A. (eds.) *Handbook of development economics*, vol. 4, pp 3561–3606. Amsterdam: Elsevier – North Holland.

Miguel, E. and Kremer, M. (2003). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**(1), 159–217.

Nelson, R. E. (2010). Testing the fetal origins hypothesis in a developing country: Evidence from the 1918 influenza pandemic. *Health Economics* **19**(10), 1181–1192.

Thomas, D., Strauss, J. and Henriques, M. H. (1991). How does mother's education affect child height? *Journal of Human Resources* **26**(2), 183–211.

Education and Health: Disentangling Causal Relationships from Associations

P Chatterji, University at Albany and NBER, Albany, NY, USA

Glossary

Endogeneity An economic variable is said to be endogenous if it is a function of other parameters or variables in a model.

Fixed effects models A statistical way of controlling for omitted variable bias when using panel data. The method is so-called on account of the fact that it holds constant ('fixes') the average differences between the determinants of a variable by using dummy variables.

Omitted variable bias In econometrics, the difference between the value of an estimated parameter and its true value due to failure to control for a relevant explanatory (confounding) variable or variables.

Production function A technical relationship between inputs and the maximum outputs or outcomes of any procedure or process. Also sometimes referred to as the 'technology matrix'. Thus a production function may relate the maximum number of patients that can be treated in a hospital over a period of time to a variety of input flows like doctor- and nurse-hours, and beds.

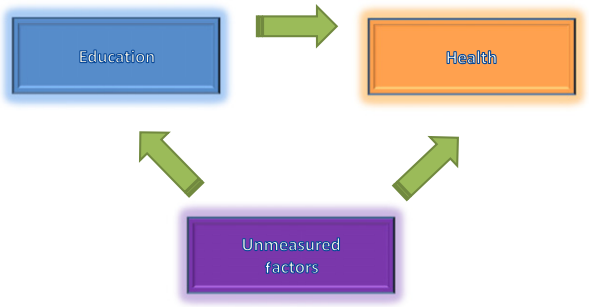
Utility Variously defined in the history of economics. Two dominant interpretations are hedonistic utility, which equates utility with pleasure, desire-fulfilment, or satisfaction; and preference-based utility, which defines utility as a real-valued function that represents a person's preference ordering.

Utility function A technical relationship that relates utility to the rate of consumption of various goods and services, or in some sophisticated cases, to the characteristics of consumer goods and services. Such determinants as health and educational attainment are postulated to yield utility directly as well as indirectly through an enhanced enjoyment of goods and services.

Introduction

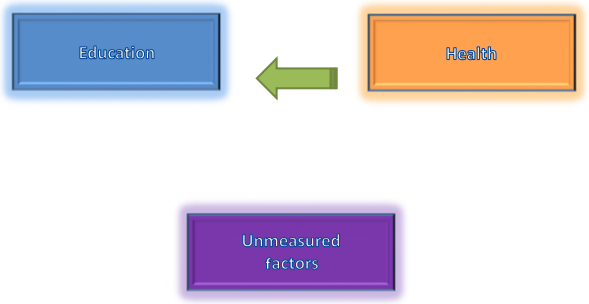
Most people would not be surprised to learn that education is positively associated with health. This seems intuitive, and consistent with what is observed in society. However, many would be surprised by the strength and pervasiveness of the link between education and health across different contexts and different indicators of health. More educated people live longer than those who are less educated, and the importance of education as a determinant of mortality is only growing over time. Chronic diseases, such as asthma and diabetes, are more prevalent among lower educated groups compared to higher educated groups. Even among those with chronic disease, education is positively associated with timely disease diagnosis, effective self-management, and better disease outcomes. Education is positively correlated with healthy behaviors such as exercise and use of preventive care and it is negatively associated with virtually all the risky health behaviors such as poor eating habits, lack of exercise, problem drinking, illegal drug use, and smoking. Maternal education plays a similar role as a determinant of children's health. Maternal education is positively associated with a broad range of children's health and developmental outcomes, ranging from children's preventive health care to mental health outcomes.

Some people argue that is it not education per se, but rather factors correlated with education, such as income, that lead to better health. It may be observed that educated people, for example, exercise more than the less educated. But this may be the case not because of education but rather because educated people earn higher incomes and can afford, say, gym memberships. To some extent, this is true – factors correlated with education, especially income and ability, do account for some portion of the association between education and health. But, in general, the strong and pervasive association between health and education persists and remains policy-significant in magnitude even when researchers take into account a broad range of other factors that are correlated with both education and health, such as income, family background, and demographic characteristics. Does this mean that education truly improves health, or are there factors that cannot be measured well that underlie this relationship? If education does indeed cause better health, what makes education so crucial to health? These questions have intrigued economists for the past four decades.

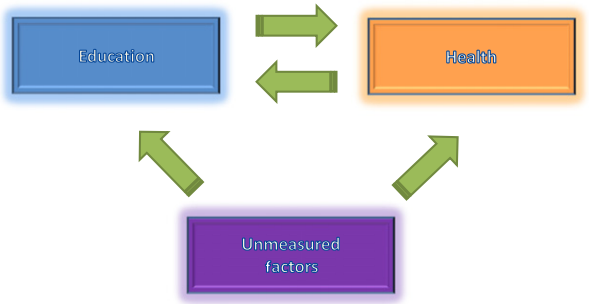


The existence of a robust, positive association between education and health does not necessarily mean that more education causes better health. The reverse causal pathway is also plausible. Better health early in life may lead individuals to complete more schooling, because longer life expectancy increases the benefits of educational investments, and/or because better health improves school attendance and helps students to learn better. There is a growing body of evidence suggesting that early health – even health in utero – can have profound implications for future, adult health, and well-being. Thus, an observed association between education and health among adults may result not from education casually affecting health but rather from early health affecting both health and education in adulthood.

 doi:10.1016/B978-0-12-375678-7.00813-0



Also possible is a noncausal explanation for the correlation between education and health. The correlation may come from unmeasured variables that are associated with both health and education, such as ability, genetics, or family socioeconomic status (SES). Some have suggested that individuals with strong preference for present versus future outcomes – that is, individuals with high discount rates – will not make long-term investments in health or education. If this trait is hard to measure in data, it may appear that education is positively associated with health, but in reality individuals' time preference, which is unmeasured, determines both health and education. In this case, a strong, positive association between education and health may exist, but it does not reflect a causal relationship. It is also possible that more schooling causes individuals to be more future-oriented. In this case, education may affect health causally through its effect on the rate of time preference.



In recent literature, economists employ innovative econometric methods to determine whether the association between education and health is causal. Although most studies are based on data from the US and the UK, increasingly data from around the world are being used to examine this relationship. The economics literature on education and health is very large, and it is expanding rapidly. This short review does not cover all health economics literature in this area. Instead, the focus is on empirical research on education and health in developed countries published in the past 10 years in economics journals. The goal is to highlight some provocative papers, synthesize results, and draw conclusions from recent studies that have attempted to distinguish causal relationships from associations between health and education.

The Grossman Model

Most empirical literature on education and health is motivated by the Grossman Model (Grossman, 1972a,b; 2000). The Grossman Model is a model of the demand for the commodity 'good health,' which is treated as a durable good, or a type of capital. Health has both direct consumption value as well as investment value in this model. Health has consumption value because individuals derive utility from being in good health. Health has investment value because it determines the total amount of time that is available to work in the market and nonmarket sectors. Briefly, in the Grossman Model, individuals maximize a utility function which includes health and other commodities with respect to investments in health, given budget and time constraints. Optimal gross investment in health determines the optimal amount of health, because the depreciation rate (e.g., wear and tear on health capital) and initial health are given.

Grossman analyzes the effect of education on health in his pure investment model. In this version of the model, the consumption value of health is not considered. Education is viewed as the technology of the health production function. More education makes individuals better producers of health. In other words, an increase in education would allow an individual to obtain more health from a given set of inputs, decreasing the marginal cost of an investment in health. The decrease in the marginal cost of investment increases the returns on health capital, and the optimal level of health is higher than before. Thus, according to the Grossman Model, more education leads people to choose higher levels of health because education increases individuals' productive efficiency in producing health (Grossman, 1972a,b; 2000).

The mechanism through which education increases productive efficiency is hard to pinpoint. One can argue, in fact, that it is more likely that education causes better health by improving individuals' allocative efficiency in producing health (Grossman, 2008, 1972a). For example, more education may cause individuals to understand better how to combine inputs to produce health; thus, individuals may make more efficient choices about how much to exercise, what to eat, how to adhere to medical treatments, and what health behaviors to avoid. The distinction between the productive and allocative efficiency arguments can be important from an empirical perspective. If education increases health primarily through improvements in allocative efficiency, if one is estimating a health production function, then there should not be an association between education and health if all inputs are included in the model as well (Grossman, 2008). This is not the case if education improves productive efficiency, because more education leads to individuals directly obtaining more health from a given set of inputs.

Grossman (1972a) emphasizes that one's stock of health is an endogenous choice variable. Current health depends on initial health, depreciation of the health stock in all previous periods, and gross investment (and thus inputs used to produce investments) in all previous periods (Grossman, 1972a). Therefore, when researchers estimate the effect of early health on subsequent education outcomes, it is important to include controls for factors that may affect education directly and also may affect early health through prior health investments, such as family background. However, it is possible that the controls included do not completely account for prior health investments, and that these factors remain in the error term of the equation. Thus, endogeneity resulting from omitted variable bias is a concern to researchers when estimating the effects of early health on later education outcomes.

When estimating effects of education on health, omitted variables bias is still a possibility, because unmeasured factors such as ability may exist that are correlated with both education and health. But in this case structural endogeneity is potentially a problem as well because in a full model of education and health, education and health may be determined simultaneously. Moreover, when estimating the effects of education on health, a reverse causal pathway, with current health affecting current education, is plausible. Thus, when estimating the effects of education on health, health is considered to be endogenous in a structural as well as in a statistical sense.

Econometric Methods Used to Test for Causal Effects

In recent literature, two main empirical approaches have been used to distinguish causal relationships between education and health from associations. The first approach is to rely on a natural experiment. Some examples of natural experiments that have been used to identify effects of early health on later outcomes are famines (Chen and Zhou, 2007), periods of religious fasting (Almond and Mazumder, 2011), outbreaks of illness (Bleakley, 2010; Almond, 2006), rainfall (Maccini and Yang, 2009), nuclear accidents (Almond *et al.*, 2009), and crop infestation (Banerjee *et al.*, 2010). These events are treated as exogenous shocks to early health. One drawback of examining the effects of these events is that the results sometimes may not be readily generalizable to other settings.

In studies of the effects of education on health, researchers have taken advantage of the natural experiments induced by variation in educational policies across time and place. Some examples of natural experiments that have been used to identify effects of education on health are variation in policies that affect school entry (Braakmann, 2011), access to secondary education (Arendt, 2005), and variation in county-level access to college (Currie and Moretti, 2003). Frequently, researchers have drawn on these natural experiments to implement instrumental variables (IV) methods (Eide and Showalter, 2011). The primary advantage of using IV methods in this context is that this approach addresses both the statistical and structural endogeneity. A drawback of this approach, however, is the possible low predictive power of the instruments and its associated problems (Staiger and Stock, 1997). Also, IV findings cannot be generalized to individuals whose educational decisions are not 'at the margin' or, in other words, individuals whose educational decisions are not influenced by the policy that is being used as an instrument.

The second approach used to test for causal relationships in this literature are sibling/twin fixed effects models. This method involves estimating the correlation between within-twin (or within-sibling) differences in birth outcomes and within-twin (or within-sibling) differences in later educational outcomes. This approach essentially 'differences out' family-specific fixed characteristics that may confound an observed association between early health and later education outcomes. Sibling/twin fixed effects models address a specific form of statistical endogeneity – confounding by unmeasured, fixed family-specific characteristics.

There are some advantages in using twins rather than siblings to implement these models. In studies on the educational consequences of birth weight, within-sibling birth weight can vary because of differences in intrauterine growth retardation (IUGR) and/or differences in gestational length, whereas between twins, variation must come from a single source, IUGR (Oreopoulos *et al.*, 2008; Almond *et al.*, 2005). Also, sibling fixed effects models do not address time-varying family characteristics. Maternal health behaviors may vary by birth order, or SES could change between births. These changes may be unmeasured (Oreopoulos *et al.*, 2008; Royer, 2009) and confound an observed relationship between early health and later education outcomes. This issue does not arise in the case of twins, who are born at the same time. Also, unobserved individual heterogeneity, such as genetic differences, may exist within siblings and within fraternal twins (Almond *et al.*, 2005).

In these ways, sibling fixed effects models implemented using data on twins, particularly monozygotic twins, are subject to fewer biases. However, an important advantage of using siblings instead of twins is that the results are more easily generalized to the population. Moreover, even analyses based on within-twin differences can suffer from problems related to measurement error, unstable estimates (Royer, 2009), and selection problems caused by mortality at birth within-twin pairs (Black *et al.*, 2007; Royer, 2009).

Effect of Health on Education

Health at Birth

Malnutrition and poor health in utero or early in childhood is a predictor of later health outcomes, including infant mortality, height, cognitive function, chronic disease, and disability (Barker *et al.*, 1989; Banerjee *et al.*, 2010; Case and Paxson, 2009; Van Ewijk, 2011; Delaney *et al.*, 2011; Chay and Greenstone, 2003). Economic conditions measured at birth have been found to be related to adult mortality (Van Den Berg *et al.*, 2006). These findings, which demonstrate the importance of the early health environment for later health, imply that poor health environment early in life may affect economic outcomes as well. In estimating long-term effects of health at birth, the challenge is in determining whether poor early health is the cause of later problems, or whether it is instead a correlate of such problems (Oreopoulos *et al.*, 2008; Black *et al.*, 2007). There is a burgeoning health economics literature in this area, focusing on education as an outcome, with many innovative identification strategies being used. Numerous studies focus on estimating the long-term effects of birth weight, a single aspect of early health. However, increasingly other measures of early health are being considered. In fact, in many studies, researchers estimate reduced-form models in which the health environment early in life is linked directly to later educational outcomes. In these papers, the mechanism through which early health detracts from later education is not always well-understood.

In a landmark study, Almond *et al.* (2005) examined the long-term consequences of low birth weight (LBW) using data on twins born in the US between 1983 and 2000. They examined the correlation between within-twin differences in birth weight and within-twin differences in (1) hospital

charges, (2) other measures of health at birth, and (3) infant mortality rates. The authors also estimate the effect of prenatal smoking on a variety of infant health outcomes using singleton births, controlling for sociodemographic variables available on birth certificates. In these analyses, they attribute the entire effect of smoking on infant health to the effect of smoking on birth weight, which is probably an overstatement. The authors cannot fully control for unobserved heterogeneity using this approach – but they can gauge whether the magnitudes of the effects generated using the sample of twins are reasonable.

The cross-sectional estimates suggest that a 1 standard deviation increase in birth weight leads to reduction in hospital costs, reduction in infant mortality, increase in Apgar score, and reduction of assisted ventilator use after birth of .51, .41, .51 and .25 standard deviations, respectively. Based on the twins analysis, however, these magnitudes fall to .08, .03, .03, and .01. The smoking analysis shows that smoking affects birth weight appreciably, but smoking is not related to most infant health outcomes – as a result, cost savings of smoking cessation during pregnancy are modest. Either the true effect of birth weight on infant health has been overstated in prior work; and/or each analysis isolates a different set of determinants of birth weight.

Using a similar approach to that of Almond *et al.* (2005), there have been several studies based on samples of twins which examine the effects of birth weight on long-term educational outcomes. All these studies support the idea that birth weight has long-term consequences for adult education and health outcomes. Black *et al.* (2007), for example, draw on administrative data on twins born between 1967 and 1981 in Norway and study the consequences of birth weight. They build on Almond *et al.* (2005) in that they are able to examine the effects of birth weight not just on short-run health outcomes (infant mortality and 5 min Apgar score) but also on long-run outcomes including adult height, intelligence quotient (IQ), employment, earnings, education, and birth weight of the first child. Like Almond *et al.* (2005), Black *et al.* (2007) found that within-twin differences in birth weight are associated with smaller effects on short-run outcomes compared to cross-sectional, ordinary least squares (OLS) estimates. However, Black *et al.* (2007) report that there are long-term effects of birth weight on adult height, body mass index, IQ, education, earnings, and birth weight of the first born child. For these outcomes, OLS and within-twin estimates are similar in magnitude.

Royer (2009) studies the effects of birth weight on educational attainment, later pregnancy complications, and birth weight among offspring using data on same-sex, female twins born in California between 1960 and 1982. Among these twins, long-term outcomes can be studied for those who survive to adulthood, remain in California, and give birth to infants between 1989 and 2002. Consistent with other research, Royer finds that cross-sectional estimates of the effect of birth weight on short-run health are overstated. The estimated within-twin effect of birth weight on 1-year mortality is similar to that of Almond *et al.* (2005) in magnitude. Royer finds small, long-term effects of birth weight on women's educational attainment. It is interesting and unexpected that Royer finds that the positive effect of birth weight on education is largest for infants who are of normal birth weight (>2500 g). Royer also finds that within-twin differences in birth weight are correlated with women's later pregnancy complications and birth weight of their own children. Currie and Moretti (2003), also using data from California, report a similar finding. They find that birth weight differences within pairs of sisters are correlated with within-sister variation in subsequent birth of an LBW infant. This effect is stronger for women living in low SES neighborhoods.

Behrman and Rosenzweig (2004) also estimate twin-fixed effects models to study the association between birth weight and adult outcomes, including educational attainment. They use a sample of monozygotic twins born in Minnesota between 1936 and 1955. The findings show that fetal growth (weight divided by length squared) is positively associated with both height and educational attainment in adulthood.

Other researchers have examined effects of health at birth using data that include siblings as well as twins. Oreopoulos *et al.* (2008), for example, test whether within-sibling differences in health at birth are correlated with within-sibling differences in later outcomes. The sample includes more than 96% of all children born in Manitoba, Canada between 1978–82 and 1984–85. They examine the effects of infant health not just on infant mortality, but also on long-term educational and employment outcomes, including childhood mortality, language scores in grade 12, physician services utilization during adolescence, reaching grade 12 by age 17, and social assistance receipt. Notably, they use multiple measures of infant health including birth weight, Apgar score, and gestational length. The findings from this paper based on twins are consistent with those from Almond *et al.* (2005) – the effect of poor infant health on mortality rates diminishes when twin differences are examined. However, infant health – especially birth weight and Apgar score – are associated with educational attainment at age 17 and public assistance receipt, suggesting that there are long-run effects of infant health on human capital accumulation.

Johnson and Schoeni (2010) also find long-term effects of LBW and early economic disadvantage on educational attainment, labor market, and health outcomes measured in adulthood. They use data from the Panel Study of Income Dynamics (PSID) and sibling fixed effects models. Similarly, Fletcher (2011), using data from the National Longitudinal Study of Adolescent Health (Add Health), estimates the effects of LBW on education outcomes using siblings fixed effects models. He finds that LBW is associated with early grade repetition, special education placement, and diagnosis of learning disability. However, unlike Oreopoulos *et al.* (2008) and Johnson and Schoeni (2011) does not find effects of LBW on longer term educational outcomes such as educational attainment.

In addition to examining effects of health at birth, there are many papers examining the effects of prenatal shocks to health, including inter-uterine exposure to famines, religious fasting, illness, adverse economic conditions, and toxins. Chen and Zhou (2007), for example, test for causal effects between exposure to the 1959–61 famine in China and health and labor market outcomes in adulthood among those who survived. They find that children born in 1962 (who were in

utero during the famine) became shorter adults than they would have been had they not been exposed to the famine. Among those exposed during early childhood, famine exposure is associated with reduced labor supply and earnings in adulthood. Almond *et al.* (2009) study effects of prenatal exposure to radiation stemming from the 1986 Chernobyl nuclear accident in the Ukraine. These authors study effects on academic outcomes among children in Sweden who were exposed 8–25 weeks post-conception to varying degrees of fallout from the accident. The findings show that low levels of prenatal exposure to radiation has no discernible effects on children's health, but it is associated with worse academic performance in high school. The effects are stronger for children from more disadvantaged backgrounds.

Almond (2006) use US data to test for long-term effects of prenatal exposure to the 1918 influenza pandemic on economic outcomes including education. They find that such exposure is associated with about a 15% reduction in the likelihood of graduation from high school and a 5–9% fall in men's wages, as well as with increases in physical disability and receipt of public assistance. Maccini and Yang (2009) estimate reduced-form models to examine the effect of rainfall around the time of birth on Indonesian adults' socioeconomic and health outcomes. They find that rainfall in utero does not affect adult outcomes. However, rainfall in the first year of life is positively associated with health and educational attainment among women, presumably because higher rainfall increases agricultural yields and household resources. Almond and Mazumder (2011) study long-term effects of prenatal exposure to Ramadan, a period of religious fasting. Using data from Michigan, they find that prenatal exposure is associated with lower birth weight. Using data from Uganda and Iraq, these authors report that exposure to Ramadan in utero is associated with large increases in the likelihood of adult disability. Case and Paxson (2009) use data from the Health and Retirement Study and find region-level infant mortality and disease rates in the first 2 years of life are associated with cognitive function in old age (Case and Paxson, 2009). In sum, there is a convincing body of evidence that prenatal health conditions and health at birth have long-term effects on later educational attainment and other adult outcomes. In some cases, the causal mechanism appears to be adult health, but in other cases, mechanisms linking early health to later outcomes are not clear.

Health during Youth

There also is a small but growing literature on the effects of health during childhood on educational outcomes in developed countries. Case *et al.* (2005), for example, examine this relationship using the 1958 National Child Development Study. This survey includes data collected from birth until age 42 on all children born in the UK during the week of 3 March 1958. The results show that chronic health conditions in childhood, as well as LBW, are associated with reductions in educational attainment, employment, social status, and adult health. Although this study draws on unusually rich data which should minimize problems of unobserved heterogeneity, the methods do not directly address the problem of disentangling causality from correlation.

Some researchers, however, have used sibling fixed effects models to difference out family-specific factors that may drive both children's health and educational outcomes. These studies generally support the idea that health during childhood affects educational attainment. Some studies have used self-rated overall health rankings to measure child health. Smith (2009), for example, estimate sibling fixed effects models using data from the PSID to examine the effect of child health on adult labor market outcomes. Child health is measured using a retrospective self-report of overall health before age 17. The sibling fixed effects model findings do not show a statistically significant relationship between health in childhood and educational attainment. However, there are positive effects of child health on family income, household wealth, individual earnings, and labor supply.

Chay *et al.* (2009) focus on how access to and quality of health care early in life affects later educational outcomes. They examine the effects of desegregation and forced integration of hospitals in the US during the 1960s and 1970s on racial disparities in test scores in the 1980s. They find that access to better health care in early childhood reduced African-American/white disparities in achievement test scores later in life.

Other studies estimate effects of specific chronic health conditions during childhood on later educational and labor market outcomes. Fletcher *et al.* (2010), for example, use data from the National Longitudinal Study of Adolescent Health (Add Health) to examine the effect of childhood asthma on missed days from school and work, obesity, and adult health. They use sibling fixed effects models and find large, detrimental effects of childhood asthma on absenteeism. Rees and Sabia (2011), also using Add Health, find that migraine headaches detract from educational outcomes. Sabia (2007), using data from Add Health, finds a negative association between body weight and grades for white females, but not for other sociodemographic groups. Grossman and Kaestner (2009), however, using data from the NLSY79, do not find any statistically significant association between body weight and children's achievement test scores.

There is also evidence that exposure to tropical disease in childhood affects later educational outcomes. Bleakley (2007) studies the effect of hookworm on long-term educational outcomes in the US, taking advantage of a natural experiment in which a public health campaign was instituted in the early 1900s to eradicate the disease. Bleakley finds that childhood hookworm has very large effects on adult wages, mostly through reducing the returns to schooling. In another paper, Bleakley (2010) finds that childhood malaria reduces income in adulthood. In this study, to identify effects of malaria on outcomes, he takes advantage of malaria eradication campaigns instituted in the US and in Latin America.

Results from several studies highlight the importance of mental health for educational and other human capital outcomes. Currie *et al.* (2010) draw on administrative data from Manitoba, Canada, and examine whether childhood health problems are associated with adult educational attainment, test scores, and social assistance receipt. The primary estimation strategy is sibling fixed effects models. The results show that childhood health problems, especially mental health problems, detract from adult educational attainment and other outcomes. These findings are consistent with those of Currie and Stabile

(2006). They employ sibling fixed effects models and use national survey data from the US and Canada and find that hyperactivity symptoms during childhood are associated with worse educational outcomes, such as grade repetition and special education placement. Fletcher and Wolfe (2008) are able to replicate these findings of the effects of hyperactivity on short-run educational outcomes using a different data source (Add Health). However, Fletcher and Wolfe find that hyperactivity does not affect longer term educational outcomes, such as educational attainment.

In addition to these studies that focus on hyperactivity, other economics studies show that depressive symptoms during youth are associated with lower grades and lower educational attainment (Eisenberg *et al.*, 2009; Fletcher, 2010). In addition, a few new studies using data from the US show that having genetic markers for depression and attention deficit hyperactivity disorder are associated with adverse educational outcomes (Ding *et al.*, 2009; Fletcher and Lehrer, 2009). However, Contoyannis and Dooley (2010), using data from the Ontario Child Health Study, examined the association between child health (measured by conduct or emotional disorder, and by chronic condition or functional limitation) on a range of educational attainment and labor market outcomes measured in adulthood. They find that child health is negatively associated with educational attainment and labor market outcomes, but these findings do not persist when sibling fixed effects are included in the models.

Effect of Education on Health

Maternal Education and Child Health

Maternal education is a powerful correlate of children's health outcomes, but whether this relationship is causal remains an open question. Several recent papers focus on testing whether a causal relationship exists between maternal education and child health. Currie and Moretti (2003) make important progress in this area by examining the effect of maternal education on infant health at birth using data from US individual birth certificates from 1970 to 1999. They hypothesize four potential causal pathways linking maternal education to infants' health: (1) effects of maternal education on prenatal care; (2) effects of maternal education on spousal earnings; (3) effects of maternal education on health behaviors (prenatal smoking); and (4) effects of maternal education on fertility (quality/quantity tradeoff). They use an IV method with availability of colleges at the county level as an instrument for maternal education. Currie and Moretti find that higher maternal education improves children's birth weight and gestational age at birth. This is a large effect – an additional year of college is estimated to reduce the incidence of LBW by 10%. Their results show that maternal education increases the probability of marriage, increases husband's education, reduces parity, increases use of prenatal care, and reduces smoking. These pathways, therefore, may be mechanisms through which maternal education affects infants' health.

McCrary and Royer (2011), however, use US birth certificate data and come to different conclusions. They test whether maternal education affects fertility and infant health (birth weight, prematurity, infant mortality) using large samples of birth records from Texas and California which include the exact date of birth. They rely on school entry cutoffs, which allow them to compare birth outcomes of women born just before and just after their states' school entry cutoffs. Although women born just after the school entry date do complete less education than women born just before, their infants are as healthy as those of women born just before the school cutoff. These findings, then, suggest that for women whose educational decisions are affected by school cutoff policies, maternal education does not appear to play a causal role in infant health.

Carneiro *et al.* (2011) examine the effects of maternal education on children's cognitive test scores, behavior problems, and the home environment using data from the National Longitudinal Survey of Youth 1979 (NLSY79). They instrument for maternal education using local labor market conditions, college tuition, and the existence of a 4-year college in the county where the mother lived at age 14. The findings show that maternal education is positively associated with test scores and negatively associated with behavioral problems among children.

Chou *et al.* (2010) estimate the effect of maternal and paternal education on LBW and infant mortality using birth certificate data on infants born in Taiwan between 1978 and 1999. They take advantage of a natural experiment related to educational attainment. In 1968, Taiwan extended compulsory schooling from 6 to 9 years and opened 150 new junior high schools. Before 1968, junior high enrollment was restricted by a difficult exam. The findings show that maternal education and paternal education both affect infant health, but maternal education appears to be more important.

Finally, Chen and Li (2009) use Chinese data to examine whether maternal education affects the health of adopted versus biological children. They find that maternal education is associated with better child health for both adopted and biological children. This finding does not definitely establish a causal relationship, but it is revealing that maternal education is strongly associated with child health, even when genetic explanations are eliminated.

Education and Health

There is a large literature on the effects of education on one's own health. In this literature, economists have studied the effects of education on mortality, chronic health conditions, and a wide range of health behaviors. In an influential paper, Lleras-Muney (2005) uses a quasinatural experiment to determine whether the association between education and mortality represents a causal relationship. The natural experiment consists of states changing their compulsory schooling and child labor laws between 1915 and 1939, inducing some individuals to obtain more schooling than they would have otherwise. Data come from the US Censuses from 1960, 1970, and 1980. Her sample includes whites born in 48 states who were 14 years old between 1914 and 1939, with available data on education. She creates synthetic cohorts by aggregating Census data into groups by gender, cohort, and state-of-birth, calculates mortality rates for these groups, and examines direct effect of changes in compulsory schooling on mortality rates

by comparing mortality rates of cohorts immediately before and after there was a change in legislation. This regression discontinuity approach offers only suggestive evidence of an effect of education on mortality. She then uses the compulsory education laws as instruments, and finds statistically significant negative effects of education on mortality. The effect is large in magnitude – a 10% increase in education lowers mortality by 11%.

Albouy and Lequien (2009) examine the effect of education on mortality in France and come to different conclusions. They rely on changes in compulsory schooling laws as a natural experiment and use regression discontinuity and IV methods, as was done by Lleras-Muney (2005). However, their findings show that while changes in schooling laws affected education, there was no effect on mortality.

Numerous studies examine the effect of education on health and health behaviors using variation in school policies to instrument for education. These studies have yielded mixed findings. Arendt (2005), for example, examines the relationship between education and health (measured by self-reported overall health, body mass index, and smoking) in Denmark. He instruments for education using school reforms intended to expand access to secondary school education. The findings suggest that better education is associated with better health, but the instruments do not perform well empirically in this study, making it hard to draw conclusions from the IV results. Kemptner *et al.* (2011) explore the relationship between education and health using German data, instrumenting for education using changes in compulsory schooling laws. They find evidence of causal effects of education on having a long-term illness for men, for results for other health outcomes are less consistent. Braakmann (2011) studies the effect of education and a range of health and health behaviors using data from the UK. He instruments for education using month of birth, because in the UK, school policies interact with the month of birth such that children born after 30 January are forced to attend school longer than those born before 30 January. The IV results show no effects of education on health. Other studies using compulsory schooling laws for identification show that additional schooling improves self-reported health (Oreopoulos, 2007; Silles, 2009), and may decrease the likelihood of having hypertension (although these findings are mixed) (Powdthavee, 2010).

In addition to school policies, researchers have drawn on other natural experiments to isolate the causal relationship between education and health. de Walque (2007), for example, tests whether the correlation between post-high school education and smoking behaviors (measured by the likelihood of current smoking and the likelihood of having quit smoking) is causal, using the risk of induction into Vietnam War as an instrument for education. He uses data from the smoking supplements of the NHIS between 1983 and 1995. The sample includes persons born between 1937 and 1956 with the age of 25 years and above at the time of the survey. The findings indicate that college education is causally related to a reduction in the likelihood of smoking. However, it can only be concluded that this effect occurs among individuals induced to attend college because of Vietnam draft. Grimard and Parent (2007) address the same question using a similar identification strategy, but different data. They find similar, but less consistent, evidence that education is causally related to smoking.

Siblings/twins fixed effects models also have been used to study the effect of education on health. Webbink *et al.* (2010), for example, use fixed effects models and data on identical twins from the Australian Twin Register to examine the causal effect of education on body weight. They find a strong association between education and overweight status, but this association only persists within twins for males (not for females). Fletcher and Frisvold (2009) estimate the association between college attendance and investments in preventive care using longitudinal data on a sample of individuals who graduated from high school in 1957 in Wisconsin. These individuals are followed for approximately 50 years. The findings show strong associations between college attendance and preventive care usage. These results persist when sibling fixed effects are included in the models. These findings are consistent with those of Lange (2011). Using data from the National Health Interview Survey (NHIS), he finds that more educated people are more likely to respond to individual risk factors for cancer by investing in preventive care than less educated people. This study suggests the mechanism through which education affects use of preventive care may be individuals' understanding and processing of health information.

There is growing interest not just in the effect of the quantity of education on health, but also on the effects of school quality on health. Frisvold and Golberstein (2011) use data from the 1984 to 2007 NHIS, linking respondents to race-specific state-year of birth measures of school quality (such as pupil teacher ratios). A range of health outcomes are examined, including overall self-rated health, mortality, and obesity. Their findings show that higher quality schooling magnifies the effect of education on health. Similarly, Johnson (2010) uses data from the PSID and shows that within siblings, long-term childhood exposure to desegregated schools is associated with adult health, suggesting that school quality has long-run effects on health. Similarly, Kenkel *et al.* (2006) find that high school completion is associated with lower rates of smoking and higher rates of quitting smoking, but there are lower health returns to the GED versus the traditional high school diploma. These results also suggest there is some interaction between schooling quality and the effects of schooling on health.

Drawing Conclusions from Health Economics Research

From a policy perspective, it is critical to disentangle causal relationships between education and health from associations. If more and/or better education causes better health, then public policies that expand access to and/or improve the quality of education will also be effective in improving health. Similarly, if better health causes individuals to obtain more education, health policies can be used to increase education. If causal relationships do indeed exist, health policy and education policy are intertwined.

Economists have made important contributions in this area. There is now a convincing body of economics research supporting the idea that early health is causally related to

long-term education and other economic outcomes. Health measured in utero, at birth, and during childhood and adolescence, affect outcomes such as educational attainment, labor supply, and wages in adulthood. There is also some evidence to support the idea that education causes better health, but these findings are inconsistent and vary by the health outcomes studied and the data used.

For research on education and health to be useful in shaping health and education policies, it is important not just to test for causality but also to identify causal mechanisms. Cutler and Lleras-Muney (2010) take an important step in this direction by examining the education gradient in health behaviors using data from a range of national data sets from the US and the UK. Their approach is to estimate a model in which education affects health behaviors and then include increasingly richer sets of controls in the model to see how inclusion of additional covariates affects the estimated coefficient on education. Overall, the authors conclude that material resources account for approximately 20% of the effect of education on health behaviors. Ability also accounts for a portion of the effect of education on health behaviors. This paper is an important addition to the literature because mechanisms through which education may affect health can now be understood.

Moreover, it is important to understand whether the effects of education on health, and the effects of health on education, are heterogeneous in the population. For example, some research suggests that the effects of education on health vary by individuals' sociodemographic characteristics (Cutler and Lleras-Muney, 2006). Other studies support the idea that education causes better health, but the results are relevant to only subpopulations (often the lowest part of the education distribution), and, based on existing research, cannot be generalized to the entire population (Lleras-Muney, 2005). It is essential to know which groups are most likely to respond to changes in education, or to changes in health. Economics research has the potential to answer these questions about mechanisms and heterogeneity of effects, and thus help in shaping the development of effective health and education policies.

See also: Education and Health. Education and Health in Developing Economies

References

Albouy, V. and Lequien, L. (2009). Does compulsory education lower mortality? *Journal of Health Economics* **28**, 155–168.

Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of *in utero* influenza exposure in the post-1940 U.S. population. *Journal of Political Economy* **114**, 672–712.

Almond, D., Chay, K. and Lee, D. (2005). The costs of low birthweight. *Quarterly Journal of Economics* **120**(3), 1031–1083.

Almond, D., Edlund, L. and Palme, M. (2009). Chernobyl's subclinical legacy: Prenatal exposure to radioactive fallout and school outcomes in Sweden. *The Quarterly Journal of Economics* 1729–1772.

Almond, D. and Mazumder, B. (2011). Health capital and the prenatal environment: The effect of Ramadan observance during pregnancy. *American Economic Journal: Applied Economics* **3**, 56–85.

Arendt, J. N. (2005). Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education Review* **24**, 149–160.

Banerjee, A., Duflo, E., Postel-Vinay, G. and Watts, T. (2010). Long run health impacts of income shocks: Wine and phylloxera in nineteenth-century France. *The Review of Economics and Statistics* **92**, 714–728.

Barker, D. J. P., Winter, P. D., Osmond, C., Margetts, B. and Simmonds, S. J. (1989). Weight in infancy and death from ischaemic heart disease. *Lancet* **334**, 577–580.

Behrman, J. R. and Rosenzweig, M. R. (2004). Returns to birthweight. *The Review of Economics and Statistics* **86**, 586–601.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. *The Quarterly Journal of Economics* **122**(1), 409–439.

Bleakley, H. (2007). Disease and development: Evidence from hookworm eradication in the south. *The Quarterly Journal of Economics* **122**, 73–117.

Bleakley, H. (2010). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics* **2**, 1–45.

Braakmann, N. (2011). The casual relationship between education, health and health related behaviour: Evidence from a natural experiment in England. *Journal of Health Economics* **30**, 753–763.

Carneiro, P., Meghir, C. and Parey, M. (2011). Maternal education, home environments and the development of children and adolescents. *Institute for Fiscal Studies Working Paper*. Cambridge, MA: National Bureau of Economic Research.

Case, A., Fertig, A. and Paxson, C. (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics* **24**, 365–389.

Case, A. and Paxson, C. (2009). Early life health and cognitive function in old age. *American Economic Review: Papers & Proceedings* **99**, 104–109.

Chay, K. and Greenstone, M. (2003). The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession. *The Quarterly Journal of Economics* **118**, 1121–1167.

Chay, K. Y., Guryan, J. and Mazumder, B. (2009). Birth cohort and the black-white achievement gap: The roles of access and health soon after birth. *Working Paper 15078*, National Bureau of Economics Research. Available at: http://www.nber.org/papers/w15078

Chen, Y. and Li, H. (2009). Mother's education and child health: Is there a nurturing effect? *Journal of Health Economics* **28**, 413–426.

Chen, Y. and Zhou, L. (2007). The long-term health and economic consequences of the 1959–1961 famine in China. *Journal of Health Economics* **26**, 659–681.

Chou, S., Liu, J., Grossman, M. and Joyce, T. J. (2010). Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan. *American Economic Journal: Applied Economics* **2**, 33–61.

Contoyannis, P. and Dooley, M. (2010). The role of child health and economic status in educational, health, and labour market outcomes in young adulthood. *Canadian Journal of Economics* **43**, 323–346.

Currie, J. and Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics* **118**, 1495–1532.

Currie, J. and Stabile, M. (2006). Child mental health and human capital accumulation: The case of ADHD. *Journal of Health Economics* **25**, 1094–1118.

Currie, J., Stabile, M., Manivong, P. and Roos, L. L. (2010). Child health and young adult outcomes. *The Journal of Human Resources* **45**, 518–548.

Cutler, D. M. and Lleras-Muney, A. (2006). *Education and health: Evaluating theories and evidence*. Working Paper 12352. Cambridge, MA: National Bureau of Economic Research.

Cutler, D. M. and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics* **29**, 1–28.

Cutler, D. M., Meara, E. R. and Richards, S. (2008). The gap gets bigger: Changes in mortality and life expectancy by education, 1981–2000. *Health Affairs* **27**, 350–360.

Delaney, L., McGovern, M. and Smith, J. P. (2011). From Angela's ashes to the Celtic tiger: Early life conditions and adult health in Ireland. *Journal of Health Economics* **30**, 1–10.

Ding, W., Leher, S. F., Rosenquist, J. N. and Audrain-McGovern, J. (2009). The impact of poor health on academic performance: New evidence using genetic markers. *Journal of Health Economics* **28**, 578–597.

Eide, E. R. and Showalter, M. H. (2011). Estimating the relation between health and education: What do we know and what do we need to know? *Economics of Education Review* **30**, 778–791.

Eisenberg, D., Golberstein, E. and Hunt, J. B. (2009). Mental health and academic success in college. *The B.E. Journal of Economic Analysis and Policy* **9**, 1–35.

Fletcher, J. M. (2010). Adolescent depression and educational attainment: Results using sibling fixed effects. *Health Economics* **19**, 855–871.
Fletcher, J. M. (2011). The medium term schooling and health effects of low birth weight: Evidence from siblings. *Economics of Education Review* **30**, 517–527.
Fletcher, J. M. and Frisvold, D. E. (2009). Higher education and health investments: Does more schooling affect preventive health care use? *Journal of Human Capital* **3**, 144–176.
Fletcher, J. M., Green, J. C. and Neidell, M. J. (2010). Long term effects of childhood asthma on adult health. *Journal of Health Economics* **29**, 377–387.
Fletcher, J. M. and Lehrer, S. F. (2009). The effects of adolescent health on educational outcomes: Casual evidence using genetic lotteries between siblings. *Forum for Health Economics & Policy* **12**, 1–31.
Fletcher, J. M. and Wolfe, B. (2008). Child mental health and human capital accumulation: The case of ADHD revisited. *Journal of Health Economics* **27**, 794–800.
Frisvold, D. and Golberstein, E. (2011). School quality and the education–health relationship: Evidence from Blacks in segregated schools. *Journal of Health Economics* **30**, 1232–1245.
Grimard, F. and Parent, D. (2007). Education and smoking: Were Vietnam war draft avoiders also more likely to avoid smoking? *Journal of Health Economics* **26**, 896–926.
Grossman, M. (1972a). On the concept of health capital and the demand for health. *Journal of Political Economy* **80**, 233–255.
Grossman, M. (1972b). *The Demand for Health: A Theoretical and Empirical Investigation*. New York: Columbia University Press, for the National Bureau of Economic Research.
Grossman, M. (2000). The human capital model. In Culyer, A. and Newhouse, J. (eds.) *Handbook of Health Economics*, vol. 1A, pp. 348–408. Amsterdam: North Holland.
Grossman, M. (2008). The relationship between health and schooling. *Eastern Economic Journal* **34**, 281–292.
Grossman, M. and Kaestner, R. (2009). Effect of weight on children's educational achievement. *Economics of Education Review* **28**, 651–661.
Johnson, R. C. (2010). The health returns of education policies from preschool to high school and beyond. *American Economic Review: Papers and Proceedings* **100**, 188–194.
Johnson, R. C. and Schoeni, R. F. (2010). The influence of early-life events on human capital, health status, and labor market outcomes over the life course. *The B.E. Journal of Economic Analysis and Policy* **2**, 188–194.
Johnson, R. C. and Schoeni, R. (2011). The influence of early-Life events on human capital, health status, and labor market outcomes over the life course. *The B.E. Journal of Economic Analysis & Policy: Advances* 11,(3), article 3.
Kemptner, D., Jurges, H. and Reinhold, S. (2011). Changes in compulsory schooling and the casual effect of education on health: Evidence from Germany. *Journal of Health Economics* **30**, 340–354.
Kenkel, D., Lillard, D. and Mathios, A. (2006). The roles of high school completion and GED receipt in smoking and obesity. *Journal of Labor Economics* **24**, 635–660.
Lange, F. (2011). The role of education in complex health decisions: Evidence from cancer screening. *Journal of Health Economics* **30**, 43–54.
Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *Review of Economic Studies* **72**, 189–221.
Maccini, S. and Yang, D. (2009). Under the weather: Health, schooling, and economic consequences of early-life rainfall. *American Economic Review* **99**, 1006–1026.
McCrary, J. and Royer, H. (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review* **101**, 158–195.
Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics* **91**, 2213–2229.
Oreopoulos, P., Stabile, M., Walld, R. and Roos, L. (2008). Short, medium, and long-term consequences of poor infant health: An analysis using siblings and twins. *Journal of Human Resources* **43**.
Powdthavee, N. (2010). Does education reduce the risk of hypertension? Estimating the biomarker effect of compulsory schooling in England. *Journal of Human Capital* **4**, 173–202.
Rees, D. I. and Sabia, J. J. (2011). The effect of migraine headache on educational attainment. *The Journal of Human Resources* **46**, 317–332.
Royer, H. (2009). Separated at girth: US twin estimates of the effects of birth weight. *American Economic Journal: Applied Economics* **1**, 49–85.
Sabia, J. J. (2007). The effect of body weight on adolescent academic performance. *Southern Economic Journal* **73**, 871–900.
Silles, M. A. (2009). The casual effect of education on health: Evidence from the United Kingdom. *Economics of Education Review* **28**, 122–128.
Smith, J. P. (2009). The impact of childhood health on adult labor market outcomes. *The Review of Economics and Statistics* **91**, 478–489.
Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586.
Van Ewijk, R. (2011). Long-term health effects on the next generation of Ramadan fasting during pregnancy. *Journal of Health Economics* **30**, 1246–1260.
de Walque, D. (2007). Does education affect smoking behaviors? Evidence using the Vietnam draft as an instrument for college education. *Journal of Health Economics* **26**, 877–895.
Webbink, D., Martin, N. G. and Visscher, P. M. (2010). Does education reduce the probability of being overweight? *Journal of Health Economics* **29**, 29–38.

Further Reading

Barker Theory (2010). Available at: http://www.thebarkertheory.org (accessed February 2012).
Chatterji, P., Joo, H. and Lahiri K. (2012). Racial/ethnic and education-related disparities in control of risk factors for cardiovascular disease among diabetics. *Diabetes Care* **35**, 305–312.
Chay, K. and Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of Political Economy* **113**, 376–424.
Conti, G., Heckman, J. and Urzua, S. (2010). The education–health gradient. *American Economic Review Papers and Proceedings* **100**, 234–238.
Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature* **47**, 87–122.
Currie, J. (2011). Inequality at birth: Some causes and consequences. *American Economic Review: Papers & Proceedings* **101**, 1–22.
Cutler, D. M., Lange, F., Meara, E., Richards-Shubik, S. and Ruhm, C. J. (2011). Rising educational gradients in mortality: The role of behavioral risk factors. *Journal of Health Economics* **30**, 1174–1187.
Cutler, D. M. and Lleras-Muney, A. (2012). *Education and health: Insights from international comparisons*. Working Paper 17738. Cambridge, MA: National Bureau of Economic Research.
Fuchs, V. R. (1982). Time Preference and Health: An Exploratory Study. In Fuchs, V. R. (ed.) *Economic Aspects of Health*. Chicago: University of Chicago Press.
Goldman, D. R. and Smith, J. P. (2002). Can patient self-management help explain the SES health gradient? *Proceedings of the National Academy of Sciences* **99**, 10929–10934.
Grossman, M. (2006). Education and non-market outcomes. In Eric, H. and Finis, W. (eds.) *Handbook of the Economics of Education*, vol. 1, pp. 577–633. Amsterdam: Elsevier.
Kaestner, R. and Grossman, M. (2009). Effects of weight on children's educational achievement. *Economics of Education Review* **28**, 651–661.
Mazumder, B. (2008). Does education improve health? A reexamination of the evidence from compulsory schooling laws. *Federal Reserve Bank of Chicago Economic Perspectives* **32**, 2–16.
Meara, E. (2001). *Why is health related to socio-economic status? The case of pregnancy and low birth weight*. Working Paper 8231. Cambridge, MA: National Bureau of Economic Research.
Van Den Berg, G. J., Lindeboom, M. and Portrait, F. (2006). Economic conditions early in life and individual mortality. *American Economic Review* **96**, 290–302.
Van Der Pol, M. (2011). Health, education and time preference. *Health Economics* **20**, 917–929.

Efficiency and Equity in Health: Philosophical Considerations

JP Kelleher, University of Wisconsin–Madison, Madison, WI, USA

Concepts of Efficiency

The everyday concept of efficiency is fairly straightforward. It connotes an optimizing relation of gains to losses, as well as the avoidance of wastage. Within economics, more technical notions of efficiency include Pareto efficiency and potential Pareto efficiency. These are central notions for cost-benefit analysis (CBA), which seeks to identify efficiencies across multiple policy domains. CBA converts each policy domain's benefits into monetary equivalents and assumes that maximizing overall monetized benefits is a worthy goal (even if not the only worthy goal). By contrast, domain-specific analyses seek locally efficient policies and often employ the notion of cost-efficiency or cost-effectiveness. Within health policy, for example, cost-effectiveness analysis (CEA) seeks to identify policies that would maximize certain health-related outcomes given a fixed budget. Here, there is no need to convert relevant outcomes to monetary equivalents because there is no need to express both health and nonhealth benefits in terms of some common unit.

This article's discussion on efficiency will focus on the notion of cost-effectiveness as it is employed within health policy. Two issues in particular are addressed: first, because the idea of cost-effectiveness suggests the importance of maximizing something, some specific health-related benefit(s) must be identified as the maximand at the individual level; and second, after an individual-level maximand is determined, many philosophical and ethical considerations bear on the selection and interpretation of the social maximand that will ultimately inform policymaking at the population level. These two issues are explored in the Sections Individual-Level Maximands and Social Maximands and the Ethics of Maximization, respectively.

Individual-Level Maximands

Health

It is natural to think that efficiency in health policy should be construed as maximizing health itself. However, two related reasons have been put forward against that proposal. First, it can be difficult to make the assessments of overall health that it requires. Second, asking if someone is in 'good health' is often a way of asking if their health adversely affects their life. If health's impact depends on the way it interacts with other features of one's situation, then it may be misguided to focus on health itself rather than on the ways health, together with other factors, affects people's lives.

Some have replied to these and similar worries about focusing on health itself by noting that it is clearly possible to make at least some relevant comparisons, such as when it is said that someone with a mild sore throat is healthier (assuming all else is equal) than someone who cannot walk. This judgment is a plausible assessment of health itself, not a judgment about health states' impact on the goodness or badness of a life. But is it possible to build a rigorous assessment of population health around specific health-focused judgments? Doing so would require a large number of health-state assessments, but many such judgments are not as clear-cut as the example just offered. To illustrate the difficulty, Daniel Hausman presents the example of a person with a mild learning disability and someone with quadriplegia. Although the first person is presumably in better health than the second, Hausman doubts that there is an objectively defensible framework for comparing units of mobility with units of cognitive functioning. This, he argues, highlights the difference between saying the first person literally has more health (a descriptive judgment) and saying that it is better to be in his health state than to suffer from quadriplegia (an evaluative judgment). Given the conceptual difficulties with measuring a population's literal health (especially when health is multidimensional), and given that health policy's main interest is in how good a population's health is or can be, it is reasonable to conclude that the maximand at the individual level should be evaluative, rather than descriptive. This in effect would bypass the need to measure health itself, but it also raises new questions about how health should be valued.

Well-Being

If one seeks to evaluate the goodness or badness of a health state, a natural proposal is to focus on the impact the state has on individual well-being. Of course, much would turn on the nature of well-being, and philosophers have identified important problems for several accounts of it.

One central candidate is subjective well-being, i.e., the sense of satisfaction with one's life and prospects. A central worry with this approach is that it could ignore significant health improvements that accrue to those who already enjoy high subjective well-being. Ronald Dworkin used the example of Dickens' Tiny Tim to make a similar point. If the magnitudes of relevant health benefits are tied to improvements in subjective well-being, then an intervention that restores Tiny Tim's mobility may bring very little health benefit, given Tim's already cheerful disposition. A similar problem concerns adaptive and even malformed preferences: intuitively significant improvements in health will be downplayed if they would go to individuals who are already subjectively satisfied with very little because of exposure to aspiration-numbing deprivation or injustice.

Preference Satisfaction

Economists often equate well-being with the satisfaction of preferences, and many assessments of health policy draw on valuations derived from data about respondents' preferences over health states. 'Satisfaction' can be a misleading term here,

because what is relevant is getting what one wants, not a subjective feeling that may (or may not) come from getting what one wants.

There is strong reason to keep individual well-being and preference satisfaction separate, and to avoid tying the importance of health improvements to individual preferences. For example, one may prefer a policy in part for altruistic reasons, i.e., because of its impact on third parties. In such cases, satisfying one's preferences could actually come at a cost to one's own well-being. Second, it is possible for individuals self-interestedly to want and prefer things that are not in fact good for them, that do not in fact promote their well-being. This can be due both to false empirical beliefs and to misguided prudential outlooks. Prudential preferences are hardly ever brute 'gut' preferences. As TM Scanlon put it, "My preferences are not the source of reasons but reflect conclusions based on reasons of other kinds" (Scanlon, 2003, p. 177). This opens the possibility that individuals' preferences may be insensitive to objective reasons for thinking that a given health state is better or worse for them.

Opportunities and Capabilities

Many take examples like the one involving Tiny Tim to justify focusing on more objective consequences of deficits in health. Regardless of its impact on his subjective welfare, Tiny Tim's impairment reduces the opportunities that are available to him in significant ways. Amartya Sen has long advocated for a metric of policy evaluation that focuses on people's objective capabilities. Such a framework would divorce the public importance of health-related capabilities from any given agent's personal preferences about them: one person in a given health state may be made miserable by it, whereas another in a similar state may have adjusted fully and now lives a flourishing life. From a perspective of opportunity and capability, these individual viewpoints (and their aggregation) may not matter as much as the disinterested assessment of whether the health state generally impedes or closes off life opportunities that society deems it morally important for citizens to have access to. A view of this sort will therefore not base the valuation of population health on individual preferences or subjective well-being, because these capture the state's importance along the wrong evaluative dimension. The main questions raised by opportunity-based frameworks concern which health-related capabilities should be the focus of health policy, and how they can be measured in a scientifically respectable way. Hausman (2010) has offered the most detailed current proposal, which suggests using deliberative groups to evaluate health states "with respect to the relation 'is a more serious limitation on the range of objectives and good lives available to members of the population'" (p. 280).

These are the most prominent individual-level maximands on offer, and it is important for health economists to be able to distinguish between them and to keep in mind the reasons for and against them. Consider again economists' most common maximand, preference satisfaction. If the value of a health state is determined by individuals' (aggregated) preferences about it, then questions arise about whose preferences should count. One natural thought is that relevant preferences should be adequately informed, and this leads to the suggestion that the preferences of those who are most familiar with the health state should count for more. But here the issue of adaptation arises, because it is possible to live an excellent life after adapting to a given health state. If adaptive preferences inform society's ultimate appraisal of health states, then the importance of having a range of opportunities open to one will be downplayed: At a certain stage in life, what matters from the first-person perspective is that one is able to lead the kind of life one has decided on for oneself; and once one has decided on living a certain kind of life, it is less important that one be able to choose from among options one has already ruled out. Further, from the perspective of a healthy person who views health states P and Q as equally terrible because of how they conflict with his current life plans, an imagined change from P to Q may not seem all that meaningful, even if in objective terms the change would significantly enhance the range of life opportunities open to the average person in P.

Much of the practical relevance of these debates lies in their bearing on how CEA should be carried out. CEA typically uses quality-adjusted life-years (QALYs) as the individual-level maximand. QALYs are designed to integrate longevity considerations with quality of life considerations in a way that enables comparisons between health interventions targeting very different dimensions of health. Because they are built by aggregating individual preferences over health states, QALYs should not be viewed as a measure of literal health; they are rather a measure of the value of changes in health, where the value of a change is interpreted as the difference in the values assigned to the two relevant health states. However, just as it is possible to carry out a CEA using a decidedly descriptive maximand (e.g., number of surgical complications averted), it should also be possible to employ CEA's techniques in the context of different evaluative individual-level maximands. Whatever individual-level maximand is chosen, it remains to be determined how interpersonally comparable benefits and losses to individuals should be combined and valued at the aggregate level in the service of shaping and guiding public policy. Notwithstanding the ethical issues already raised for preference-based evaluative metrics, the following discussion of aggregate-level 'social maximands' will, purely for ease of illustration, be conducted using QALYs as the illustrative individual-level benefit.

Social Maximands and the Ethics of Maximization

The term 'social maximand' seems to suggest that health policy should aim, at least in part, to maximize something. And many philosophers criticize CEA precisely because they believe it embodies a single-minded focus on maximizing QALYs (or on whichever individual-level maximand is ultimately chosen). But this appraisal is too quick. For CEA can be put forward as an assessment of efficiency only, rather than a complete decision-making framework. And even if CEA is proposed as a complete decision-making framework, it is possible within CEA to employ a social maximand that ranks policies on the basis of their interpersonally comparable effects on individuals but which also places differential evaluative weight on otherwise similarly sized benefits depending

on who receives them. To use the language of welfare economics, different CEAs can thus operate with different social welfare functions as the social maximand, thereby operating with different adjustments to efficiency. This even opens up the possibility of what might be called an equity-sensitive social maximand. One problem with this approach, however, is that efficiency and equity are usefully viewed as distinct concepts, and an equity-sensitive social maximand blurs the distinction between them. Thus, to keep these dimensions of evaluation distinct, this section begins with ethical concerns that arise when CEAs employ an equity-insensitive social maximand – that is, when CEAs recommend the single-minded pursuit of efficiency, and when efficiency is construed simply as QALY-maximization. The section will close by noting a difficult issue that arises if one seeks to incorporate a certain equity consideration into the social maximand.

Few would claim that QALY-maximization is an irrelevant goal. The question is whether and when it should be constrained by other ethical factors. Philosophers have identified four main factors that are neglected by what shall here be called 'pure' CEAs, i.e., CEAs that recommend straightforward QALY-maximization.

Aggregation

Pure CEA permits small benefits to lots of people to be summed up to outweigh large benefits to a smaller number of people. For example, a government-sponsored commission in Oregon (US) in 1990 released a draft priority list of health care services that prioritized some oral and dental treatments over life-saving procedures like appendectomy and surgery for ectopic pregnancy. Dollar for scarce dollar, providing appendectomies was not as cost-effective as those nonlife-saving services.

Discrimination against the Disabled

Suppose that subpopulation A is disabled whereas subpopulation B is not; each subpopulation is the same size and all individuals are otherwise equally healthy. Now suppose an epidemic afflicts both populations and leaves all individuals with a life-threatening illness. Assume also that logistical limitations allow for life-saving treatment to be administered to just one subpopulation; all members of the treated subpopulation will be restored to their preillness condition and if saved each would live the same number of additional years. Pure CEA recommends against choosing the disabled population, because this generates fewer QALYs. Many find this a troubling form of discrimination.

Priority to the Worse Off

Pure CEA cannot explain why one should give priority to the worse off when this intuitively seems required. Suppose the individuals in Group A generate 0.3 QALYs per year and could be brought to produce 0.5 instead. And suppose that equally numerous individuals comprising Group B generate 0.8 QALYs per year and can be brought to full health (1.0). Once again suppose that scarcity or logistics require choosing just one group to assist. Pure CEA recommends flipping a coin, because from the standpoint of the maximizer, adding 0.2 QALYs per year to a person's life has the same importance regardless of that person's initial condition. Many find this counterintuitive and believe there is a moral presumption in favor of treating the worse off.

Fair Chances versus Best Outcomes

Suppose the members of two equally numerous groups, A and B, each currently generate 0.5 QALYs per year. Now suppose that either A can be helped or B can, but not both: members of A can be brought to generate 0.8 QALYs per year, or members of B can be brought to generate 0.95. Pure CEA favors helping B and neglecting A, but many find this problematic. As Frances Kamm puts it, although the members of B can be helped a bit more, it is true both that members of A are capable of gaining the major part of what members of B can gain, and that this major part is what each cares most about – namely, a substantial improvement in health. This way of describing the situation leads some people to support giving equal or perhaps proportional chances to A and B, rather than choosing to only help B.

Each of these stylized scenarios raises equity concerns, but there is no consensus on how to incorporate equity considerations into health-economic analysis. Consider, for example, the problem of aggregation. Employing different variations of Oregon's methodology and personal valuations of health states from respondents, Ubel *et al*. found that pure CEA can equate the successful treatment of 10 cases of appendicitis with the successful treatment of between 111 and 1000 cases of mild hand pain. Yet when the same respondents were asked directly how many cases of mild hand pain would be equivalent to 10 cases of appendicitis, 17 of 42 respondents said it would take an infinite number of cases. This finding comports with a common response to Oregon's draft proposal: Many believe that, morally speaking, no number of capped teeth could equal or outweigh saving a life with an appendectomy. But this raises a puzzle, as virtually no one claims that it is always wrong to give priority to less serious but more numerous needs over more serious but fewer needs. Suppose, for example, one could either prevent 10 000 people from developing paraplegia or one could save one person's life, but not both. It seems clear that the relative numbers tip the ethical scales toward the 10 000. But note that there are no people among the 10 000 who, if not helped, could reasonably complain that they were left without mobility while someone else's life was saved. In that respect, this case parallels the case involving dental services and appendectomy: There are not people among candidates for tooth capping who could reasonably complain that their tooth will be left uncapped if the legislature pays for appendectomies instead. But then if it can still be permissible to favor large numbers in the case involving paraplegia, why not also in the case involving tooth capping? The difficult question, therefore, is not whether aggregation can be morally permissible, but rather when and on what basis aggregation is permissible.

Partially in response to the equity concerns connected to the problem of aggregation, health economists have explored ways to build respondents' direct rationing preferences into an 'impure,' equity-sensitive CEA framework. Such preferences

can be elicited using the so-called 'personal trade-off' (PTO) exercises of the sort Ubel *et al*. used to uncover the discrepancy between pure CEA and respondents' direct rationing judgments. One notorious problem with the PTO methodology is the problem of multiplicative intransitivity. The problem is nicely described by Ubel (2000), pp. 168–169:

> Imagine a person who thinks that curing one person of condition A is equally beneficial as curing ten people of condition B, and that curing one person of condition B is equally beneficial as curing ten of condition C. To be consistent, this person ought to think that curing 1 person of condition A is equally beneficial as curing 100 people of condition C. However, when we conducted PTO measurements for three such conditions and multiplied the PTO values of the two "nearer comparisons" (such as A vs B and B vs C), we calculated a different value for the relative importance of the "far comparisons" (such as programs A and C) than people told us when they were directly asked to compare these programs [i.e. A and C].

Because no survey can ask respondents directly to compare every possible pair of competing health interventions, health economists seek a solution to the problem of multiplicative transitivity that could license inferences from discrete preferences about 'nearer comparisons' (A vs. B, B vs. C, . . . , Y vs. Z) to preferences about 'far comparisons' (A vs. Z). One problem not mentioned in the economics literature is that success in this endeavor would conflict with some of the equity concerns that raised the problem aggregation in the first place. Suppose that a very long chain of near comparisons begins by comparing an appendectomy that saves one person's life with an intervention that cures some number of cases of paraplegia. Suppose the next comparison on the chain compares the curing of one case of paraplegia with the curing of some number of cases of one paralyzed arm. Now suppose the chain continues down the line until one gets to the near comparison between curing one case of mild tendonitis with curing some number of cases of individuals who suffer very mild headaches once per week. The worry now is that any solution to the problem of multiplicative transitivity would entail that there is some noninfinite number of mild headaches that would be granted priority over curing a case of appendicitis. There is a deep divide in the philosophical literature as to whether a result like this is tolerable or whether it should be avoided at all costs.

In light of these ongoing and potentially intractable philosophical issues, it may be advisable for health economists simply to rank policies with respect to QALY-maximization only and then to explicitly leave it to policy-makers to decide for themselves whether and when to depart from maximization for equity-related reasons.

The Concept of Health Equity

The most commonly cited definition of health equity is Margaret Whitehead's (1991, p. 219):

> The term 'inequity' has a moral and ethical dimension. It refers to differences which are unnecessary and avoidable but, in addition, are also considered unfair and unjust.

This definition leaves open the possibility that some differences in health are neither unfair nor unjust. This seems to be a virtue. It is not clear, however, that a health inequality must be avoidable before it can be counted an inequity. Here is what Whitehead says about this aspect of equity (1991, p. 219):

> We will never be able to achieve a situation where everyone in the population has the same type and degree of illness and dies after exactly the same life span. This is not an achievable goal, nor even a desirable one. Thus, that portion of the health differential attributable to natural biological variation can be considered inevitable rather than inequitable.

There are two ideas at work here. First, there is the idea of the desirability of equality: everyone being the same in some respect or respects. But, second, Whitehead also refers to the impossibility of equality, and it is this that seems to motivate the condition that an inequity in health must be an avoidable inequality.

There is a problem with Whitehead's avoidability condition. To see this, suppose a subset of the population is afflicted by a health impairment that cannot be avoided or resolved medically – perhaps an unalterable genetic defect makes amputation below both knees a necessity for this group. Suppose also that the legislature is considering whether to pay for wheelchairs for those afflicted by the disorder. On Whitehead's definition, considerations of equity might say nothing about whether the state should provide these assistive devices. This is because wheelchairs arguably cannot eliminate the differences in health caused by the disorder. Whitehead's definition therefore seems flawed, because it definitionally entails that the provision of assistive devices is not a demand of equity (Wilson, 2011).

If the concept of health equity should not prejudge substantive issues that a theory of health equity is intended to address, it is better to start from a much more modest version of Whitehead's definition. Thus, health inequities are simply health differences that are unjust, all things considered. The 'all things considered' qualification means that if a difference is an inequity, then there exists a moral requirement on the part of (certain) agents or institutions to do something about them. It clearly follows from this definition that some view of justice is required before a health difference can be counted a health inequity. But at least this new definition does not rule out the possibility that unavoidable health differences raise issues of equity, because an unavoidable health difference could still be unjust if it is not compensated for in the right way.

Unfairness and Equality

Whitehead's 'necessary and avoidable' condition is therefore problematic. Recall, however, that Whitehead's definition included another condition, viz. that an inequity is an inequality that is unfair. It might seem that this unfairness condition adds nothing to the definition, because whatever is unfair is unjust. But whether unfair inequalities are also unjust depends on what unfairness is, how it is related to justice and moral obligation, and whether other considerations can outweigh or displace fairness in the final determination of what is, all things considered, just and unjust.

How does Whitehead's definition of health equity connect up with the moral value of equality? In the quotation above, she argues that it is neither achievable nor desirable to have everyone in exactly the same health. Setting aside the question of achievability, why would equality not be a desirable goal? Imagine that medical progress has left us with just one disease – heart disease, say – that sets in at the age of 100 years and leaves us dead at 105 years. Would this not be desirable? Surely it would. Imagine a slightly different scenario in which heart disease sets in at the age of 100 years for both men and women, but men tend to die at 105 years whereas women die at 110 years. If one then had to choose between giving males an extra 5 years of life expectancy and giving females an extra 6 years, would not there be something to be said in favor of closing the gap rather than widening it with the more efficient female-focused policy? And might not the value of equality explain why it would be unfair to help the women before helping the men?

These considerations might suggest that equality is indeed intrinsically desirable, so long as its place is known. Having human beings be equal in each and every respect would surely be undesirable, and this may be all Whitehead is saying. But this does not entail that it would be undesirable to promote greater equality of health prospects. In some contexts, equality may be very important, and in others it may simply be less important than some other moral considerations.

Equality of Outcomes versus Process Equity

This last point is sometimes invoked in the context of sex differences in longevity. In 1994, the World Health Organization's Global Burden of Disease team used high-income populations in low-mortality countries to peg the biologically-determined sex-based inequality in longevity at 2–3 years. It might therefore be suggested that if one is committed to equity in health, health care systems should tilt in favor of treating men, as a way to achieve equality of health. However, Amartya Sen and Angus Deaton distinguish between equality of outcomes and process equity (Sen, 2002, pp. 660–661; Deaton, 2002, p. 24). Process equity is the idea that procedural fairness – for example, in health care access and delivery – is of independent moral importance. In Sen's and Deaton's view, process equity can sometimes be more important than equality of outcomes. This line of argument would enable one to give some value to equality of health outcomes without letting it dictate health policies that seem intuitively unjust for other reasons.

There is, however, a response that can be made by someone skeptical of process equity. Indeed, it is a response that Deaton himself has made. He first concedes that the inequality in life expectancy between men and women may justify tilting medical research toward understanding the factors that disproportionately affect men (Deaton, 2011). This is the sort of bias that seems defensible in cases where diseases disproportionately afflict racial minorities. It is, therefore, not clear that it should be ruled out in the context of sex differences in longevity. But if a bias in state-funded research and development can be justified, then why not a bias in health care delivery?

Here Deaton provides an answer that invokes the importance of equality of outcomes, not process equity. He notes that although women have lower prevalence of conditions with high mortality, they have a higher prevalence of conditions with high morbidity. Thus, in some contexts, providing equality of access to health care could actually be one way of equalizing overall health between men and women, because women's advantage in life expectancy might offset the morbidity disadvantages they face. Indeed, there are surely many health and nonhealth disadvantages faced by women that a few extra years might help (partially) to offset. Thus, perhaps process equity seems to conflict with equality of outcomes only when one is focused on the wrong outcome. For example, if we instead focus on guarantying that a certain range of life opportunities is open to all, there may be no reason at all to eliminate women's current advantage on the single dimension of longevity.

Questioning the Value of Equality

So far no reason has been identified to reject a form of egalitarianism that is prominent in the philosophical literature and that nicely explains the connection between Whitehead's reference to fairness and the close linguistic relation between equity and equality. The egalitarian philosopher Larry Temkin puts it thus (Temkin, 2003, p. 775):

> The essence of the egalitarian's view is that comparative unfairness is bad, and that if we could do something about life's unfairness, we have some reason to. Such reasons may be outweighed by other reasons, but they are not…entirely without force.

Temkin maintains that unfairness exists when some are worse off than others through no fault of their own. Temkin identifies two objections that might be used to rebut his view that undeserved inequality is intrinsically bad. These are the so-called Raising-Up and Leveling-Down objections (**Figure 1**).

Consider first the choice between scenarios A and B. There are two social groups in each of A and B. The width of the bars reflects the size of the group's population, and the height reflects how well-off each individual within a group is. Height may here capture years of life lived, quality-adjusted years of life enjoyed, life expectancy, etc. Taking A as the status quo, one is asked to consider whether an otherwise benign policy should be implemented that would lead to scenario B. The Raising-Up objection to a Temkin-style egalitarianism simply points out that, insofar as one is an egalitarian, one must condemn the move from A to B. The antiegalitarian who

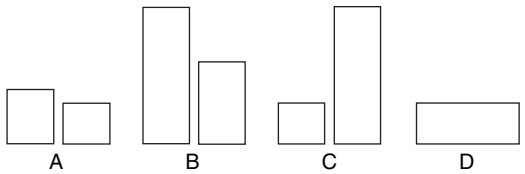


Figure 1 Depicting the Raising-Up and Leveling-Down objections. Adapted with permission from Figure on p 247 in Temkin, L. (1993). *Inequality*. Oxford: Oxford University Press.

makes this objection emphasizes that the move to B makes everybody better off. How, she will ask, could there be any reason not to improve the lives of everybody? (The assumption here is that the improvement is welcomed and not forced on anyone who does not want it.)

Temkin's response to this objection underscores a point made above, namely, that if equality has value, it does so only in the context of other important values. To use an example of Joseph Raz's, it is not important that everyone be equal with respect to the number of hairs on their shirts. That sort of egalitarianism is precisely the sort that Whitehead would be right to call undesirable. So where it makes sense to talk about the value of eliminating undeserved differences, there will always be other genuine values that are also relevant. But then if equality is not the only value, it is possible that equality can be outweighed by the other values whose presence makes equality relevant. This is Temkin's response. He agrees that a move from A to B may be the right choice once all values and reasons are given their due. He simply notes that one consideration, equality, counts against the move. To some, this is a fine response in defense of egalitarianism. True, it may seem strange to deflate equality's relative importance this much, but that seems necessary if one is attracted to Temkin's brand of egalitarianism.

The second objection to Temkin-style egalitarianism seems much more damaging. Imagine that scenario C is the status quo and one is deciding whether to support a move to scenario D (which would bring everyone in C down to the level of C's worst-off group). Plainly, D is superior with respect to equality. But there is also no one for whom D would be better than C. And yet the Temkin-style egalitarian is forced to say that there is something to be said in favor of moving from C to D. Here again Temkin insists that despite being easily outweighed by other considerations, equality still has some value even in this case.

Again, this rebuttal is clearly open to Temkin. But here the antiegalitarian's reply seems even stronger. She will highlight how bizarre it is to say there could be any reason to move from C to D, especially because no one is benefited and many people are significantly harmed. That is the Leveling-Down objection.

From Equality to Priority

The Raising-Up and Leveling-Down objections lead many to give up entirely their belief in the intrinsic value of equality. But others, like Temkin, remain steadfast. Consider the following diagram, which replicates a diagram first drawn by Michael Marmot and discussed in his book The Status Syndrome (Marmot, 2004, p. 246) (**Figure 2**).

The diagram graphs the mortality effects of a policy change on four social groups arranged from left to right in descending order of social advantage. The top line (call it Diamond) depicts the current situation and the top line (call it Square) depicts what the situation would be after implementing the proposed policy. Thus, the policy widens inequalities in mortality. But Square also offers Pareto improvements over Diamond, because each social group in Square has lower mortality than the corresponding social group in Diamond. Marmot drew the graph during a conversation with Deaton. Deaton wanted to know if

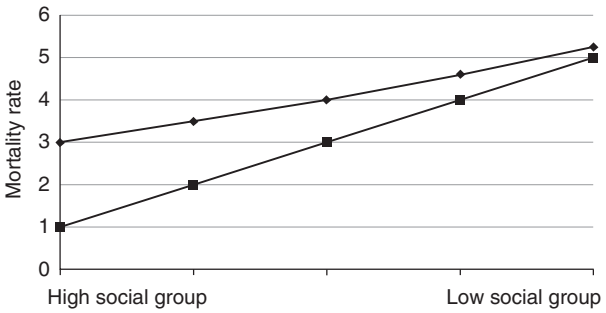


Figure 2 Social position and mortality rate: Two versions. Adapted from Marmot, M. (2004). *The status syndrome: How social standing affects our health and longevity*, p 246. New York: Henry Holt, with permission from Sir Michael Marmot.

Marmot cared more about reducing inequalities than he did about reducing sickness and death. Marmot writes:

> I demurred. [Deaton] was in no doubt that all economists would choose the bottom graph because everyone is better off…[He] suspected that I went for the one with less inequality where everyone suffered more…It is my view that we should reject both alternatives and aim for a society where health for everyone has improved and inequality is less (Marmot, 2004, pp. 245–246).

The economists Deaton referred to will likely be motivated by a commitment to Pareto improvements. In contrast, many philosophers who agree with Deaton's choice of Square over Diamond will be driven by a belief in prioritarianism. There are a number of versions of prioritarianism, but its general thrust is that, morally speaking, benefiting people matters more the worse off they are. Although prioritarians will agree that there is often reason to promote greater equality, they do not think equality is intrinsically important. Rather, a system that tilts in favor of helping the worse off will often end up more equal merely as a side effect of the prioritarian focus on improving the disadvantaged. But if improving the lot of the worse off should require or entail increases in inequality, prioritarians (like many economists) will not care.

Once prioritarianism is introduced, an intrinsic concern with equality can seem like an esthetic preference rather than a moral conviction. Where the egalitarian claims that things have gotten more unfair even though everyone is doing much better and even if the worst off are as well-off as possible, the prioritarian demands to know who (other than the egalitarian!) is complaining about unfairness. It cannot be the best off, because they are doing better than anyone and so have no right to complain. And it is unlikely to be the worst off, because they surely would not demand to be worse off than they already are.

The Value of Equality Revisited

Without concluding that Temkin-style egalitarianism is false, consider further the alternative of 'opportunity prioritarianism,' i.e., the view that social policy should tilt in favor of promoting the substantive life opportunities of those worse off (at least to the extent consistent with respecting individual

choice and personal responsibility). Such a view sees nothing intrinsically valuable in distributive equality.

Consider now an objection to opportunity prioritarianism. When one looks outside the narrow sphere of personal prospects for pursuing worthwhile life opportunities, one encounters other spheres of life within which equality seems to have intrinsic importance. Consider the spheres of political liberty and social mobility. Many believe there is a presumption in favor of equality of access to political influence and equality of opportunity (whereby no child is systematically disadvantaged in their life prospects because of their parents' socioeconomic status). If it seems appropriate to stress the intrinsic importance of equality in these political and socioeconomic domains, should this be interpreted as support by analogy for the intrinsic importance of equality within the narrower domain of personal life prospects?

The first thing to note is that the spheres of political influence and social mobility are zero sum. So even if one is a consistent prioritarian across the three domains of personal life prospects, political influence, and social mobility, equality will be the only distribution available for the last two domains: it is simply impossible to boost one social group's share of political influence or social mobility without making another worse off. This does not of course prove that equality in these realms has no intrinsic value. But it might explain why one would remain attracted to distributive equality in some spheres even if one's most fundamental ethical framework was prioritarian.

Further, if inequalities in life prospects led to unequal political influence, unequal social mobility, and to significant improvements in the range of worthwhile life plans open to those in lower- and middle-income groups, then the trade-off might be worth it on prioritarian grounds.

Of these two considerations – (1) that prioritarian inequalities are not possible in the local spheres of political liberty and social mobility and (2) that there may be prioritarian reasons to tolerate inequalities in more local spheres – neither proves conclusively that equality in these spheres is of no intrinsic importance. Indeed, there is one more way of conceiving of equality and its importance that differs from Temkin's approach and that raises the possibility that egalitarian and prioritarian concerns are both valid and in fact closely related.

The Possibility of an Egalitarian Prioritarianism

It was suggested near the end of the Section From Equality to Priority that once prioritarianism is introduced, a commitment to distributive equality can begin to resemble an esthetic preference for uniformity rather than a commitment to the real needs of individuals. However, many who hold egalitarian views about equal political influence and equal social mobility are not primarily motivated by a general desire to eliminate undeserved disadvantages between individuals. Rather, they are often moved by the independent values of nondomination, reciprocity, and equal social status. According to an increasing number of philosophers, these are the values that should ground egalitarian political convictions, as they are specially relevant for societies that care about treating all persons as moral equals. To say that each person is the moral equal of all is not yet to say that goods should be distributed in a particular way. So the ideal of moral equality is not, at bottom, a distributive ideal, although many claim that it has implications for the distribution of specific sorts of goods and for life opportunities generally. For example, distributive implications may flow from considerations about the demands of reciprocity and benevolent concern that are warranted when moral equals stand in particular social and political relationships with one another.

Some philosophers suggest that when prioritarian distributions are demanded by justice, this demand is ultimately grounded in these nondistributional premises about moral equality and ethically mandated concern (Miller, 2010). A stylized example of Thomas Nagel's provides a useful illustration. In an essay that in many ways sparked the contemporary philosophical debate between distributive egalitarians and prioritarians, Nagel describes a fictional scenario in which he has one healthy child and one suffering from a painful disability. He imagines that he must make a choice between moving to a city where the second child could receive medical treatment but which would be unpleasant for the first child, or moving to a semirural suburb where the first child alone would benefit. He stipulates that "the gain to the first child of moving to the suburb is substantially greater than the gain to the second child of moving to the city." Nagel then claims that, "If one chose to move to the city, it would be an egalitarian decision. It is more urgent to benefit the second child, even though the benefit we can give him is less than the benefit we can give the first child" (Nagel, 1991, p. 124). In response, Derek Parfit claims that Nagel has misdescribed his own moral commitments. Nagel says that the duty to attend to the disabled child's needs is an egalitarian duty. Parfit insists that Nagel is not concerned with distributive equality between the two children at all, and that Nagel instead appears motivated by prioritarian concern for the worse off child. One might reply on Nagel's behalf by claiming that Parfit works with a false dichotomy. In insisting that Nagel must be a prioritarian, Parfit ignores the brand of egalitarianism that stresses the moral demands of distinctive interpersonal relationships, including relationships that call for the display of equal and robust concern for those to whom one is specially related. When multiple individuals compete for that concern (as they may when they are our children or – plausibly but more controversially – our fellow citizens) it is reasonable to conclude that treating all of them as equals requires a prioritarian response to their diverse needs.

This last brand of egalitarianism – call it egalitarianism of concern – may hold great promise to unify and explain many intuitions about the demands of equity across multiple policy domains, including the domain of health policy. If, for example, it can be shown that compatriots or indeed 'global citizens' owe robust duties of equal concern for one another, then the distribution of medical care and other resources bearing on health should arguably follow whatever pattern is required to address the relevant needs of the worst off.

See also: Cost–Value Analysis. Disability-Adjusted Life Years. Efficiency in Health Care, Concepts of. Health and Health Care, Need

for. Health and Its Value: Overview. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Measuring Equality and Equity in Health and Health Care. Measuring Vertical Inequity in the Delivery of Healthcare. Quality-Adjusted Life-Years. Resource Allocation Funding Formulae, Efficiency of. Valuing Health States, Techniques for

References

Deaton, A. (2002). Policy implications of the gradient of health and wealth. *Health Affairs* **21**, 13–30.

Deaton, A. (2011). What does the empirical evidence tell us about the injustice of health inequalities? In Eyal, N., Norheim, O. F., Hurst, S. A. and Wikler, D. (eds.) *Inequalities in health: Concepts, measures, and ethics*. New York: Oxford University Press.

Hausman, D. (2010). Valuing health: A new proposal. *Health Economics* **19**, 280–296.

Marmot, M. (2004). *The status syndrome: How social standing affects our health and longevity*. New York: Henry Holt.

Miller, R. W. (2010). *Globalizing justice: The ethics of poverty and power*. Oxford: Oxford University Press.

Nagel, T. (1991). Equality. In Nagel, T. (ed.) *Mortal questions*, pp 106–127. Cambridge: Cambridge University Press.

Scanlon, T. M. (2003). Value, desire, and the quality of life. In Scanlon, T. M. (ed.) *The difficulty of tolerance: Essays in political philosophy*, pp 169–186. Cambridge: Cambridge University Press.

Sen, A. (2002). Why health equity? *Health Economics* **11**, 659–666.

Temkin, L. (2003). Egalitarianism defended. *Ethics* **113**, 764–782.

Ubel, P. (2000). *Pricing life*. Cambridge, MA: The MIT Press.

Whitehead, M. (1991). The concepts and principles of equity and health. *Health Promotion International* **6**, 217–228.

Wilson, J. (2011). Health inequities. In Dawson, A. (ed.) *Public health ethics: Key concepts in policy and practice*, pp 211–230. Cambridge: Cambridge University Press.

Further Reading

Daniels, N. (1994). Four unsolved rationing problems. *Hastings Center Report* **24**(4), 27–29.

Dworkin, R. (1981). What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs* **10**(3), 185–246.

Hausman, D. (2006). Valuing health. *Philosophy and Public Affairs* **34**, 246–274.

Kamm, F. M. (2009). Aggregation, allocating scarce resources, and the disabled. *Social Philosophy and Policy* **26**(01), 148–197.

Murray, C. J. L., Salomon, J. A., Mathers, C. D. and Lopez, A. D. (2002). Summary measures of population health: Conclusions and recommendations. In Murray, C. J. L., Salomon, J. A., Mathers, C. D. and Lopez, A. D. (eds.) *Summary measures of population health*, pp 731–756. Geneva: World Health Organization.

Parfit, D. (1997). Equality and priority. *Ratio* **10**, 202–221.

Sen, A. (1980). Equality of what? In McMurrin, S. (ed.) *Tanner lecture on human values*, vol. 1, pp 195–220. Cambridge, UK: Cambridge University Press.

Ubel, P., Loewenstein, G., Scanlon, D. and Kamlet, M. (1996). Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Medical Decision Making* **16**, 108–116.

Efficiency in Health Care, Concepts of

D Gyrd-Hansen, University of Southern Denmark, Odense, Denmark

Glossary

Allocative efficiency A situation in which resources are allocated to production processes and the outputs of those processes to consumers or clients so as to maximize the net benefit to society. The net benefit may be some weighted measure of 'health'.

Asymmetry of information A situation in which the parties to a transaction have different amounts or kinds of information as when, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances, or people seeking insurance have more reliable expectations of their risk exposure than insurance companies.

Contingent valuation A survey method of eliciting valuations of goods or services by which individuals are asked to state their maximum willingness to pay or the minimum willingness to accept going without contingent on a specific hypothetical scenario, like descriptions of health states, and a description of options available.

Cost–benefit analysis A form of economic evaluation by comparing the costs and the (money-valued) benefits of alternative courses of action.

Cost–effectiveness analysis A method of comparing the opportunity costs of various alternative health or social care interventions having the same benefit or in terms of a common unit of output, outcome, or other measure of accomplishment.

Incremental cost–effectiveness ratio The ratio of the difference between the costs of two alternatives and the difference between their effectiveness or outcomes.

Kaldor–Hicks criterion A test for judging whether a proposed change (say, the introduction of a new drug or the demolition of a new hospital) is welfare-enhancing. It is named for Nicholas (Lord) Kaldor (1908–86) and Sir John Hicks (1904–89). The Kaldor criterion says that if the minimum the gainers from the change are willing to pay is more than enough to compensate the losers fully, then the project is welfare-enhancing. The Hicks criterion says that if the maximum amount the losers are prepared to offer to the gainers in order to prevent the change is less than the minimum amount the gainers are prepared to accept as a bribe to forgo the exchange, then the project is welfare-enhancing. The Kaldor compensation test takes the gainers' point of view; the Hicks compensation test is made from the losers' point of view. If both conditions are satisfied, both gainers and losers will agree that the proposed activity will be welfare enhancing.

Opportunity cost The value of a resource in its most highly valued alternative use. In a world of competitive markets, in which all goods are traded and where there are no market imperfections, opportunity cost is revealed by the prices of resources: The alternative uses forgone cannot be valued higher than these prices or the resources would have gone to such uses. Within a health service with fixed budget, opportunity cost has to be judged in terms of the alternative outputs (like health) forgone when expenditure on some activity increases.

Production efficiency A given output is produced using the least-cost technically efficient combination of inputs, or conversely, output is maximized for a given level of cost.

Revealed preference A person's willingness to pay for a good or service as revealed by market transactions or a controlled experiment.

Technical efficiency A given output is produced using no more inputs than are technically necessary – there will normally be a wide variety of different combinations arising out of their substitutability.

Willingness to pay The maximum sum an individual (or a government) is willing to pay to acquire some good or service, or the maximum sum an individual (or government) is willing to pay to avoid a prospective loss. It is usually elicited from controlled experiments.

Introduction

Being efficient means 'doing something well without wasting time or energy.' To economists, efficiency is a relationship between ends and means. What is important to note is that economists refer to the relationship between the value of the ends and means, not physical quantities. In economic terms, the value of using resources is equivalent to the maximum value that the resources could have generated in alternative use, and is often referred to as the opportunity cost. The acknowledgment that all actions are associated with various degrees of opportunity costs is at the core of economics, the goal being to generate the maximum benefit with available resources. This goal requires two conditions to be fulfilled: (1) that benefits are generated at the lowest minimum cost, so that overall benefits can be maximized and (2) that the right goods or services are produced in order to generate the maximum benefits. Basically it is a question of what should be produced and how is it best produced.

How and what to produce are questions that are answered differently depending on perspective. The 'how' mainly relates to how the production of a given health care service is organized. A leader of a health care firm may want to minimize production costs to his/her firm, thus keeping focus on

minimizing costs relating to his/her own part of the production line, without keeping an eye on overall societal costs. Cost shifting may take place, and an efficient production of health care services from a hospital manager's perspective may not necessarily mean that the production is efficient from the perspective of society as a whole. The 'what' should be produced is also a matter of perspective. Which services generate the most benefit can be defined in terms of a consumer's or patient's willingness to pay (WTP) for the health care service. Alternatively, it can be defined in terms of health gains or other goals that are thought to be beneficial to the recipients of health care services or society. When reading health economic analyses that seek to portray efficiency issues, one should be wary of which budgetary perspective is being applied and on whether one believes that there is focus on the relevant utility generating components of the specific health care production.

Two concepts are important for ensuring overall efficiency: production and allocative efficiency. Production efficiency addresses the issue of using optimal combinations of resources to maximize health output. It is about choosing different combinations of resources to achieve the maximum output for a given cost. Allocative efficiency involves ensuring the right allocation of resources across programs such that the overall good is maximized. 'Utility' is an economic term, which measures the value/good of a produced outcome as perceived by the recipient. Utility-generating outcomes include factors beyond health outcomes, such as process utility or disutility and the value of information and choice. Alternatively, if allocative efficiency is defined more narrowly, it is about achieving the right mixture of healthcare programs in order to maximize the health of society.

Production Efficiency: Minimizing Cost of Production

Production efficiency corresponds to accomplishing a job with minimum expenditure of time and effort. In the production of health care services, this can be translated into having an optimal combination of operating theaters and staff. If the hospital is understaffed, the operating theaters will not be utilized efficiently, and if there are too many staff members some will at times be redundant. In ensuring production efficiency, focus may be on improving staff ratios, shortening length of stay in hospitals, or eliminating unnecessary diagnostic procedures. An array of combinations of minimum input factors that can produce the same level of output are identified, and production efficiency is obtained by considering unit costs in order to determine which of the possible combinations of input factors minimizes overall costs. In the case that unit costs differ across regional health care authorities due to variations in the scarcity of specific resources (and thus opportunity costs), different combinations of input factors may represent production efficiency across regions. Some people will distinguish between technical efficiency (which focuses on the minimum amount of factors required for a specific level of output) and production efficiency (which in addition considers unit costs). For ease of presentation, no distinction is made between these concepts in the text that follows.

Allocative Efficiency: Determining What Should be Produced

Allocative efficiency is about allocating resources such that the maximum utility is generated in terms of either health outcomes or a broader definition of utility-generating outcomes. An allocative efficient distribution may be Pareto efficient: A given distribution of resources that is not Pareto efficient implies that a certain change in allocation of goods may result in some individuals being made 'better off' with no individual being made worse off. A reallocation of resources can, therefore, improve overall welfare and a Pareto improvement is feasible. A less restrictive criterion for allocative efficiency is the Kaldor–Hicks efficiency, where an outcome is considered more efficient if those individuals that are made better off could in theory compensate those that are made worse off despite compensation not actually taking place.

Why Measuring Efficiency is Pertinent in the Context of Health Care

In theory the market for goods will automatically reach production and allocative efficiency if certain criteria are fulfilled. On the demand side, buyers in the market must be facing the full price of the good at the point of purchase and they must be able to make rational choices based on perfect and full information of the good. On the supply side, suppliers must be profit maximizers, there should be many competing suppliers, and there should not be factors deterring suppliers from moving easily in and out of the market.

In the market for health care services, these criteria are not fulfilled. First, there is a high degree of asymmetry of information, and those demanding health care services are not necessarily fully aware of which services they need, nor are they always able to judge the effectiveness of the services. Moreover, there is uncertainty regarding when the services are needed and how much they will cost. The economic uncertainty creates a market for health insurance, which means that the condition of the buyer facing the full price of the good is often not fulfilled. On the supply side, suppliers have been restricted from freely accessing the market in order to protect the less than perfectly informed patient/consumer. For example, doctors and other health care personnel have to be certified. Further, there has been a push for establishing nonprofit health care organizations on the market, again in order to protect the patient from profit-seeking suppliers.

Hence, on the supply side there are factors, which undermine a competitive market and thus the mechanisms, which will ensure that health care services are produced at minimum cost. This means that production efficiency is not guaranteed. At the same time, consumers/patients are often not equipped to judge which health care services they require and are unlikely to face the full price at the time of purchase. This means that there is insufficient basis for ensuring allocative efficiency. Consequently, production efficiency and allocative efficiency are not guaranteed by market forces, and ensuring efficiency on the market for health care services is, therefore, an important issue for health care planners, politicians, and health economists.

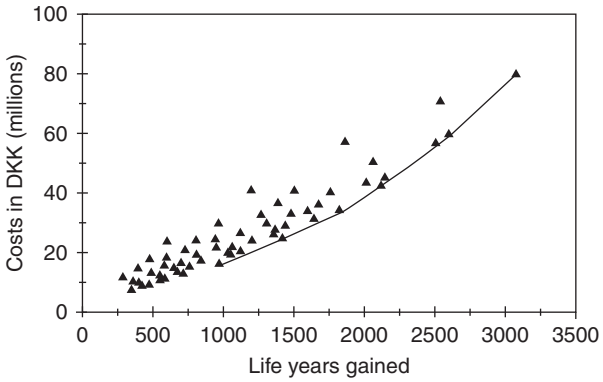


Figure 1 Alternative screening programs for colorectal cancer plotted according to costs and effects incurred over a period of 36 years. Costs and effects are discounted by 5%. A curve is drawn connecting the efficient programs. Reproduced with permission from John Wiley and Sons.

Methods of Measuring Efficiency in Health Care

Production and allocative efficiency are not independent concepts. Clearly, the unit of production that has to be maximized at minimum cost when focusing on production efficiency should be produced at the levels of quantity and quality that ensure allocative efficiency. In other words, one may be able to produce an inferior health care service very efficiently, but if there is no demand for the service, it is not worth bothering. Moreover, in ensuring a high level of production efficiency, one may be compromising allocative efficiency if the quality of the service is undermined when costs of production are reduced.

In the following are presented various methods of measuring production and allocative efficiency along with comments on the strengths and weaknesses of different methods.

Measuring Production Efficiency

Production efficiency entails producing the maximum output at a given level of employed resources. To measure and monitor production efficiency, it is essential to define the output produced as well as the production process that is under scrutiny. Outputs are typically measured in terms of services (hospital discharges, episodes of care, or covered lives) or in terms of health outcomes (postprocedure mortality rates, life expectancy, infant mortality rates, etc). There are two tools, which are typically applied in order to measure production efficiency in the context of production of health care services: productivity analysis and cost-effectiveness analysis (CEA). These will be described in turn in the paragraphs below.

Productivity analyses typically focus on an organizational unit's ability to produce maximum output at minimum costs. The output measured is often the most obvious, i.e., number of treated patients or number of consultations. The cost is most often the cost to the organization (i.e., hospital costs). Productivity analyses are often used to benchmark hospitals or hospital departments in order to identify hospitals or hospital departments which demonstrate inefficiency in production.

The level of production efficiency of a particular hospital is characterized by the relationship between observed production and some ideal or potential production. The measurement of efficiency is based on deviations of observed output from the best production or efficient production frontier. If a hospital lies below the frontier, then it is inefficient, with the ratio of the actual to potential production defining the level of efficiency of the individual firm. There are two distinct methods for estimating production efficiency: parametric and nonparametric methods. Some general concerns and challenges in applying such methods should be mentioned. The cost of production is generally limited to that of the hospital or the hospital department and may, therefore, ignore other costs involved in the production process if these lie outside the organization which is analyzed. An observed improvement in production efficiency from this narrow perspective may, therefore, not necessarily reflect cost savings from a wider (societal) perspective. Moreover, the measure of output in productivity analyses is often reliant on available output measure such as number of patients discharged or number of hospital bed days. To the extent that these intermediate measures of output do not adequately reflect utility-generating outcomes, allocative efficiency may be compromised. This is especially the case if there are strong incentives to ensure cost savings, although the quality of services produced remains unmonitored. Recently, there is an increasing focus on refining productivity analyses by incorporating dimensions of quality in output (such as mortality and wound infections) in addition to number of hospital discharges.

As in productivity analyses, CEA focus on comparing predefined outputs and comparing these with costs of production (where a perspective is chosen which may be more or less restrictive with respect to what cost items are included). If a CEA focuses on intermediate outcomes (such as numbers of cancers detected or reduction in blood pressure), the analysis is as restricted as a productivity analysis in the conclusions that can be drawn. Comparisons can only be made across interventions producing the narrowly defined unit of production, and only if an intervention is less effective and more costly or as costly as another intervention, can it be concluded that the former is inefficient. Note that for this type of CEA as well as for productivity analyses no conclusions can be drawn with respect to the relative merits of the efficient interventions (i.e., those interventions that lie on the production possibility frontier (PPF)). Figure 1 gives an example of such a frontier, where each triangle denotes a potential colorectal cancer screening strategy (target group and screening interval is varied) and the line represents the PPF (Gyrd-Hansen and colleagues, 1998). Program options that lie within the PPF are inefficient as they are dominated by at least one other program, which is either cheaper and/or more effective. Those programs that lie on the PPF represent programs that are technically efficient. However, which (if any) of these programs that fulfill the criteria for allocative efficiency is undetermined.

CEA can be applied as a tool for guiding resource allocations across the health care sector as a whole. In this case, the production unit is either defined in terms of the health care sector or society as a whole and the output of interest is quality

of life years (QALYs) gained. The broader definition of output ensures that CEA can guide the allocation of resources across various types of health care interventions aimed at different patient groups. The key parameter in this case is the cost per QALY, also referred to as the incremental cost-effectiveness ratio (ICER). Many health economists would define CEA applied in this way as a tool for ensuring production efficiency within the health care sector, i.e., ensuring that the maximum amount of output (QALYs) is produced at a given level of cost (given by the health care budget constraint). Other health economists perceive that we are in essence dealing with issues of allocative efficiency (within the bounds of the health care sector), where the aim is to ensure the optimal allocation of resources across services in the health care sector, and the maximization of QALYs is equivalent to maximizing benefits. Clearly, any disagreement on how the role of CEA is best defined is a matter of whether one defines allocative efficiency as necessarily meaning the allocation of resources across society as a whole or whether one can accept QALYs as a sufficient measure of benefits.

A more pertinent question is how CEA and ICERs are used in practice to inform decision making. In the ideal and very unrealistic scenario where all candidate health care interventions are subjected to economic evaluation and only those which are most cost effective (i.e., those with lowest ICERs) are included in the health insurance package subject to the given budget constraint, the CEA can fulfill the role of ensuring efficiency. In the more realistic scenario where resources are currently being used to run existing health care services, and there is only information on the ICER of a new intervention, the usefulness of the cost-effectiveness information is likely to be reduced. If the new health care intervention is cost saving or cost neutral, but more effective than the present intervention, the decision is straightforward. The intervention should clearly replace current practice. And vice versa if the intervention is cost neutral or cost generating and less effective. However, in many cases, new interventions are cost generating as well as more effective. In this case, it is not easy to draw any conclusions as to the welfare implications of introducing the new intervention. Introduction of the new intervention will necessarily incur opportunity costs in terms of health foregone, as there will be fewer resources available for other activities. It cannot be determined whether the health benefits foregone are larger or smaller than the acquired health benefits. Only if the health care services that may be deferred can be identified and evaluated, can an informed decision be made.

To improve the usefulness of ICERs as a tool for decision making, researchers have sought to identify a cost-per-QALY threshold as an indicator of whether an intervention is sufficiently cost-effective to warrant implementation. However, such a threshold is of little use as long as the true opportunity costs remain unidentified, which is likely to be the case if decisions are made under a predetermined budget constraint. Thresholds, as produced by way of a citizen's WTP (out of own pocket) per QALY, are only useful instruments so long as the introduction of new interventions that pass the threshold requirements are facilitated through expansion of the health care budget, thus incurring opportunity losses from reduced private consumption.

QALY league tables rank (candidate) health care interventions according to their cost-effectiveness (cost per QALY). Such tables can be useful to identify whether efficiency could be improved if some interventions take the place of others, but this necessitates the inclusion of both existing and new interventions. The more exhaustive a QALY league table is, the more useful it can be as a means of improving overall allocation of resources. However, in presenting ICERs in QALY league tables, or elsewhere, it is important that the ICERs presented are those that most precisely reflect the cost-effectiveness of the given policy relevant choice. In many cases, it is not only a question of whether or not to implement a health care service but also of how to implement it and to whom it should be offered. Interventions such as neonatal care, screening programs for cancers, prophylactic treatment of high blood pressure, etc. can be designed in many ways. Offering a health care intervention to all may appear reasonably cost-effective on the basis of the average cost per QALY. The average value may, however, hide some very expensive QALYs if a specific group of recipients experience little health gain at a high cost. It is important to choose the right comparator, and the corresponding ICER, in order to appropriately inform on efficiency implications.

Measuring Allocative Efficiency

Measuring allocative efficiency is about determining which aspects of health care services are of value to citizens, and to determine the relative importance of health care services. Measuring allocative efficiency must, therefore, in principle involve consumer preferences. In CEA, allocative efficiency (more frequently labeled production efficiency) within the bounds of the health sector is obtained by measuring benefits in terms of QALYs. Although quality adjustments are to some extent based on consumer preferences, it has been argued that this measure of benefit may in some cases be too restrictive because it does not include other utility-generating aspects of health care services such as process disutility or the value of information. Although such factors are not present in all contexts, ignoring these may result in some degree of inoptimal resource allocation.

A guide to efficient resource allocation is cost–benefit analysis. Cost–benefit analysis is based on the Kaldor–Hicks criterion, where an outcome is more efficient if those that are made better off could in theory compensate those that are made worse off. In the case of a publicly funded health care service, the losers would be the taxpayers who are financing the service and the winners are those citizens who can expect to receive the service, should they need it. Assuming that individuals are rational and fully informed about the quality of a good, consumers will be willing to pay equivalent to the marginal utility that they anticipate from buying the good. Allocative efficiency is obtained when goods in society are produced at a level where price is equal to marginal cost. Cost–benefit analysis seeks to replicate the demand side of the market by using market observations (revealed preference studies) or laboratory experiments (typically contingent valuation studies or discrete choice experiments) to establish consumers' WTP for health care services.

Contingent valuation methods and discrete choice experiments typically involve asking people how much they are willing to forego (out of their private budget) in order to ensure access to a health care service. If a cost–benefit analysis demonstrates that WTP is higher than costs, this implies that allocative efficiency is improved if the health care service is introduced. For this conclusion to hold, additional resources must be taken from private funds. If it is instead a question of determining resource allocations within a predetermined health care budget constraint, it is necessary to evaluate all the specific programs that are competing for funds. Allocative efficiency (within the health care sector) is attained when the last dollar invested across all areas of health care services generate the same level of marginal utility.

One advantage of the cost–benefit approach is that it in principle can guide resource allocations across various sectors of society. Where CEA seeks to prioritize health care services within a given budget restriction, cost–benefit analysis could ideally indicate the size of the health care budget. The benefit measure used in cost–benefit analysis also has the advantage that it is broader and far less predetermined than the benefit measures in CEA. It rests on the notion that all preferences count, which necessarily opens up for a discussion of whether the goal of the health care sector is to serve needs or wants. The Achilles' heel of cost-benefit analysis in the context of health care is whether rational and robust preferences based on a full understanding of the merits of the health care services can be derived. More research into how best to ensure that respondents understand and adequately respond to the information that is provided to them is warranted. Also, measures of allocative efficiency, which rely on private interests only, may neglect to incorporate societal benefits that are not reflected in preferences of the consumers (externalities). To the extent that there are significant positive or negative externalities involved when providing a health care service (e.g., herd immunity), these should be valued and included in the cost–benefits analysis. The extension of allocative efficiency to encompass externalities is sometimes called social efficiency.

See also: Evaluating Efficiency of a Health Care System in the Developed World. Health and Its Value: Overview. Quality-Adjusted Life-Years. Theory of System Level Efficiency in Health Care. Willingness to Pay for Health

Further Reading

Boadway, R. and Bruce, N. (1984). *Welfare economics*. Oxford: Basil Blackwell.

Culyer, A. J. (1989). The normative economics of health care finance and provision. *Oxford Review of Economic Policy* **5**(1), 34–58.

Donaldson, C., Currie, G. and Mitton, C. (2002). Cost effectiveness analysis in health care: Contraindications. *British Medical Journal* **325**, 891–894.

Gerard, K. and Mooney, G. (1993). QALY league tables: Handle with care. *Health Economics* **2**(1), 59–64.

Hicks, J. (1939). The foundation of welfare economics. *Economic Journal* **49**, 696–712.

Hollingworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science* **6**(4), 203–218.

McKay, N. L. and Deily, M. E. (2008). Cost inefficiency and hospital health outcomes. *Health Economics* **17**(7), 833–848.

Olsen, J. A. and Smith, R. (2001). Theory versus practice: A review of 'willingness-to-pay' in health and health care. *Health Economics* **10**, 39–52.

Salkeld, G., Quine, S. and Cameron, I. (2004). What constitutes success in preventive health care? A case study in assessing the benefits of hip protectors. *Social Science & Medicine* **59**, 1593–1601.

Emerging Infections, the International Health Regulations, and Macro-Economy

DL Heymann, Centre on Global Health Security, Chatham House, UK, and Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, UK
K Reinhardt, Centre on Global Health Security, UK

The Economic Impact of Emerging Infections

By the end of 2009, the year in which Mexico first reported human infections with the H1N1 influenza A virus that then spread globally to cause a pandemic, 70 715 Mexicans had been reported with confirmed H1N1 infection of whom 1316 (~5%) had died. During this same period, though there were no official travel or trade bans from th,1e Mexican Government or international bodies such as the World Health Organization (WHO), Mexican tourism and trade in pork decreased both nationally and internationally. Temporal decreases in output from the pork industry contributed to a pork trade deficit of an estimated US$27 million, whereas an estimated loss of one million overseas visitors translated into an estimated economic loss of approximately US$2.8 billion.

The economic losses related to H1N1 outbreak in Mexico were clearly influenced by the unfounded perception by tourists and travel agencies that the risk of becoming infected with H1N1 was somehow greater in Mexico than elsewhere, even though the virus had spread throughout the world; and by a misunderstanding among pork trade partners that the pandemic was being amplified by infected pigs, despite the fact that it was human to human transmission, in which swine played no role, that was responsible for the global spread of the pandemic. In other countries, there were official recommendations, apparently based on this same misunderstanding that also caused negative economic impact. In Egypt, for example, slaughter of pigs was ordered by the Egyptian Government early in the pandemic, even though the H1N1 virus had already been demonstrated to be highly transmissible from human to human, and despite the recommendation of the World Health Organization for Animals that culling of pigs was not scientifically justifiable.

Countries around the world were affected as the H1N1 pandemic spread, and economies suffered. In Spain, for example, the direct economic impact of illness from H1N1 influenza on health services utilization, and indirect costs from work absenteeism, for example, has been estimated at €6236.00 per hospitalized patient. In Canada, it is estimated that the cost of the increased patient load to hospitals caused by H1N1 between April and December 2009 was Canadian$ 200 million.

The World Bank predicts that a pandemic caused by a different influenza virus, the highly infectious and virulent avian (H5N1) influenza virus, could cost the world economy as much as US$800 billion a year from direct patient costs, and indirect costs from lost lives, travel, and trade. The H5N1 virus is currently continuing to cause disease among poultry, but is only able to infect humans sporadically when they come in contact with infected chickens.

As influenza viruses are highly unstable, however, the H5N1 influenza virus could mutate or combine with other influenza viruses circulating in nature to a form that spreads easily from human to human, resulting in an influenza pandemic with much higher mortality than the H1N1 pandemic. To prevent such a scenario, attempts are being made to eliminate the H5N1 virus by culling entire flocks of infected poultry, mainly chickens. This precautionary measure, recommended by the World Health Organization (WHO) and the Food and Agriculture Organization, is causing lost revenue and poultry-replacement costs that have been estimated to be in billions of US dollars.

Emerging infections such as H1N1 and H5N1 influenza are the newly identified infectious diseases in humans caused by viruses that have breached the species barrier between an infected animal and a human. They are by definition new, and sometimes they are called novel infections. As they are new they are poorly understood, and their full potential to cause disease and death in humans is not known.

Unlike influenza, there are other emerging infections that cause human disease but are unable to spread from human to human. Economic cost associated with these infections is due to patient management and decreased work productivity while sick; and if there is death, from the lost years of work. An example is rabies: humans are infected by the bite of a rabies-infected animal, become sick and die, but do not spread the infection to other humans unless an organ is obtained from them postmortem, and grafted into another human. The direct cost of treating persons exposed to rabies has been estimated (conservatively) to be US$40 in Sub-Saharan Africa and US$49 in Asia, a cost that equals 5.8% and 3.9%, respectively, of the average annual per capita gross national income. Additional indirect costs attributed to persons with rabies occur because of death and permanent removal from the workforce. It is estimated that the economic impact from rabies each year in the United States is approximately $300 million, where an average of two human infections occur each year.

Bovine spongiform encephalopathy (BSE) is another example of an emerging infection that does not spread from human to human. BSE, or mad cow disease, was identified in the United Kingdom (UK) during the 1980s. To rid cattle populations of infection, precautionary culling of herds with infected cattle was required. When it was understood that humans could be infected with BSE from cattle and cattle products in 1996, culling activity increased, and the economic loss in the UK during the following year was estimated to be US$1.5 billion. Countries that had imported cattle from the UK also culled infected herds at a considerable economic loss.

Another emerging infection – monkeypox virus in the United States – was caused by human contact with infected

 doi:10.1016/B978-0-12-375678-7.00624-6

prairie dogs bought as pets. The prairie dogs had been infected with the monkeypox virus in pet shops by other animals imported from West Africa as exotic pets. The outbreak was stopped and there were no deaths. Though the overall direct costs for diagnosing and managing illness were not calculated, nor were the costs from the bans on pet sales by pet shops and their furnishers, they were significant to health insurance companies and to the trade in animals as pets.

Occasionally, an infection emerges at the human/animal interface, is able to spread easily from human to human, and then becomes endemic in human populations with long-term associated economic cost. HIV is one such emerging infection. Thought to have crossed the species barrier from nonhuman primates to humans sometime during the early twentieth century, it is spread from human to human mainly by intimate sexual contact. Owing to the long, symptom-free incubation period, HIV had already spread throughout the world's population by the time it was first identified in 1981. Since then the cumulative economic impact of AIDS on GDP has been estimated by various economists with a wide range of costs – one of these, the estimated direct costs in 2009 to achieve universal access to treatment and care alone was US$7 billion.

Severe Acute Respiratory Syndrome (SARS): A Case Study on the Economic Cost Associated with Emerging Infectious Disease Outbreaks

An outbreak of an emerging infection, Severe Acute Respiratory Syndrome virus, occurred in the Guangdong Province of China in late 2002. In China, SARS spread from infected persons to other family members and to health workers, and from them to others in the community, causing an outbreak associated with severe illness and death. In February 2003, when SARS was still unrecognized as a new and emerging infection in China, it crossed the border from the Guangdong Province to Hong Kong in a doctor, who had been treating patients with SARS. He himself had become sick, and during a one-night stay in a Hong Kong hotel spread SARS to other hotel guests. Before they had any major symptoms, infected hotel guests travelled by plane to other Asian countries, North America, and Europe where they became sick and spread infection to others.

SARS had never before been seen in humans. There were thus no vaccines, medicines, or predetermined measures that could be used for its control. As the virus continued to spread from human to human, there was concern that like HIV, it would become yet another endemic infection, sustaining itself indefinitely in humans. Precautionary measures to prevent international spread of the infection were immediately recommended by the WHO – it was first recommended that persons who were ill with similar symptoms and contact with geographic areas where outbreaks were occurring defer their travel until they were well.

These precautionary measures caused a decrease in international air travel from geographic areas where outbreaks were occurring. Concern and panic ensued, however, among populations from other geographic areas as well – clearly demonstrated in a decrease in passenger movements through international airports. The precautionary prevention measures recommending that persons who were ill with SARS-like symptoms postpone travel resulted in a decrease of passengers who were ill, but many well passengers perceived the risk of travel as being great. This resulted in a steady decrease in passenger movement, clearly shown in **Figure 1**, where

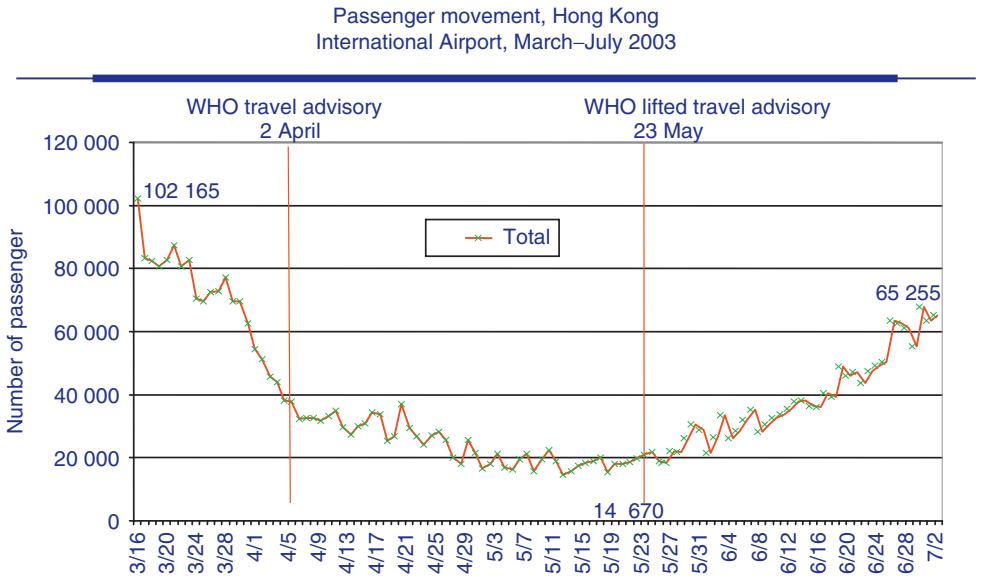


Figure 1 Passenger movement through the Hong Kong Airport from 16 March 2003, the day after the announcement of the SARS outbreak, to July 2003 when the outbreak was declared over. Passenger movement decreased immediately after the epidemic was announced on March 15, continued to decrease after a travel advisory to postpone travel was made by WHO, but increased again beginning 23 May when WHO lifted the travel advisory. Reproduced from Hong Kong International Airport and WHO.

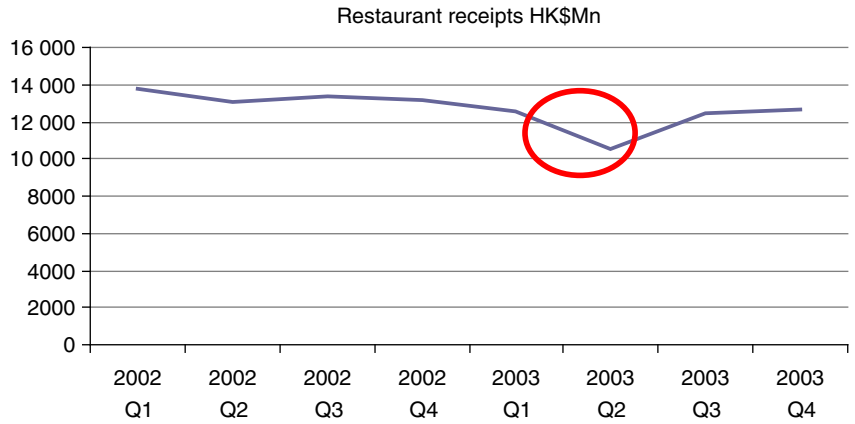


Figure 2 Restaurant receipts Hong Kong, 2002–03.

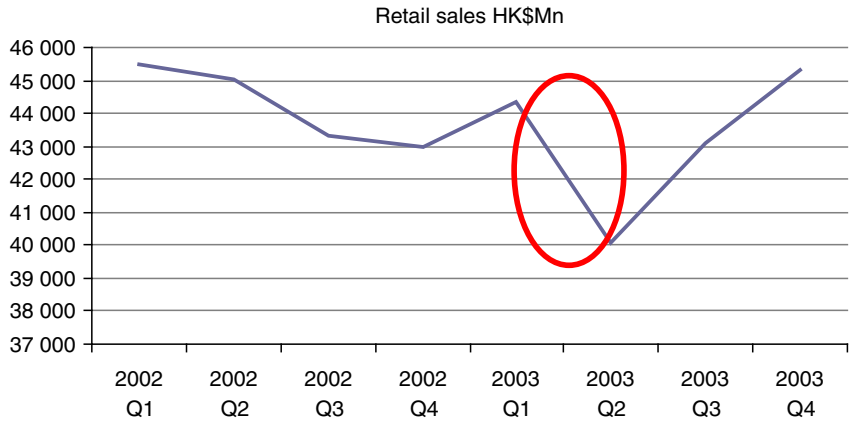


Figure 3 Retail Sales, Hong Kong, 2002–03.

passenger movements at the Hong Kong International Airport decreased soon after the outbreak was announced.

When SARS spread throughout a major housing complex in Hong Kong, among persons who had not been in contact with each other it was hypothesized that SARS might be spreading through an environmental factor such as an insect or water in addition to face to face contact. This led to stronger precautionary recommendations – to postpone or cancel travel to areas where outbreaks of SARS were occurring and a human contact could not be identified as a source of infection for each person with SARS. When WHO made this stronger precautionary recommendation on 2 April, a sustained decrease in passenger movement occurred in Hong Kong throughout the month of April and until 23 May, when the WHO removed the precautionary travel advisory.

Overall, Hong Kong International Airport had had an approximate decrease of 70% in passenger movements in April 2003 compared with April 2002, and aircraft movements decreased by an estimated 30%. In April 2003, the number of flights cancelled each day was approximately 164, representing more than 30% of all flights cancelled, and resulting in an estimated loss in landing fees of a minimum of $3.5 million per day.

During this same period, income from restaurants, hotels, and retail sales decreased because of panic and misperception of the risk among the Hong Kong population that resulted in decreased consumer activity. Figures 2–4 provide clear examples of the decreases in economic activity that occurred.

The SARS outbreak caused ended in July 2003, with 8096 reported cases from 37 countries of which 1706 (21%) were fatal. The Asian Development Bank estimated the economic impact of SARS at approximately US$18 billion in East Asia – around 0.6% of gross domestic product. However, fortunately recovery was rapid once international spread had been stopped.

The International Health Regulations and International Spread of Infectious Diseases

Attempts to limit the international spread of infectious diseases were first recorded in Venice in the fourteenth century when quarantine was used to keep ships and individuals at land border crossings in isolation for 40 days in an attempt to stop the spread of plague. Quarantine was widely used during the following centuries to attempt to limit the spread of plague and other diseases such as cholera, yellow fever, and smallpox; and during the nineteenth century a series of sanitary conferences within and between Europe and the Americas focused on these same four diseases demonstrated the concern. In the

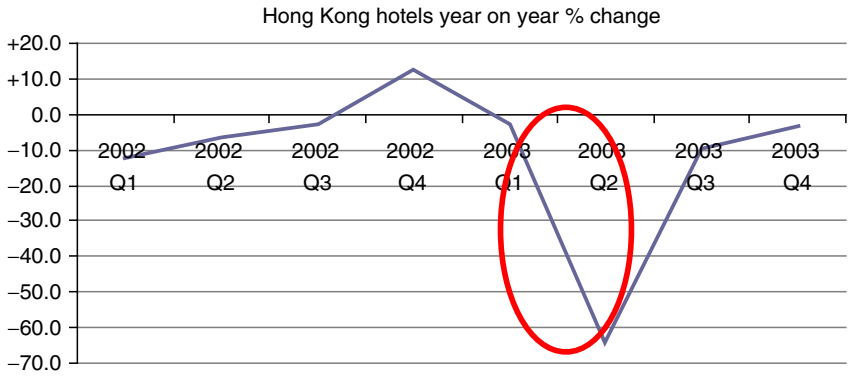


Figure 4 Year on year percentage of change in revenue, Hong King Hotels, 2002–03.

early twentieth century these sanitary conferences were broadened, under the League of Nations, to include all its member states. In 1969, the WHO Member States had agreed to a set of regulations aimed at ensuring the maximum security against the international spread of diseases with a minimum interference with world traffic.

The IHR 1969 were revised in 2005, incorporating many of the lessons learned during the SARS outbreak, and now ensure broader disease coverage, and in addition require countries to develop core capacities in public health laboratory and epidemiology in order to detect and respond to diseases where and when it occurs, and before it spreads internationally (Box 1).

Several disease threats have occurred since the revision of the IHR, including the H1N1 pandemic in 2009. The risk assessment when H1N1 first emerged was conducted by WHO and the IHR emergency committee. Though WHO recommendations based on this risk assessment clearly stated that travel and trade should continue as before, irrational trade and travel measures were imposed by several countries as described earlier in this article, and they resulted in the consequent economic losses described.

The outbreak of *Escherichia coli* (*E. coli*) that caused hemolytic-uremic syndrome in Germany occurred after the revision of the IHR as well. Though the outbreak resulted in an unexpected direct economic burden on the German health system, it also resulted in a severely negative economic impact on the European agricultural sector. Initial laboratory testing wrongly suggested that the outbreak was associated with consumption of salad greens and tomatoes imported from various countries neighboring Germany, and with consumption of cucumbers imported from Spain. Once this link was published in the mass media, the market for cucumbers fell and Spanish farmers began to experience losses of up to an estimated US$ 200 million per week. Polish, Dutch, and Italian farmers had similar losses, and German vegetable farmers had a drop in real income of 2.8%.

At the same time, Russia banned vegetable imports from the entire European Union (EU), an annual 600 million Euro market for EU farmers. As the outbreak investigation continued, however, it became clear that the outbreak was linked to ingestion of bean sprouts from an organic farm in Lower Saxony and the EU then compensated farmers in the European vegetable industry with 200 million Euros.

These two outbreaks suggest that the IHR 2005 are not completely effective in clear and effective risk communication, nor in preventing unnecessary negative economic impact. A recent outbreak of a novel coronavirus in the Middle East,

Box 1 The International Health Regulations

At the World Assembly in May 2003 based on lessons being learned from the ongoing SARS outbreak, a resolution was agreed by Member States of the WHO that helped to speed up the revision of the International Health Regulations (IHR).

The IHR first agreed by the WHO Member States in 1969, and had as a goal maximum prevention of the international spread of infectious diseases with minimal interruption of world traffic and trade. By setting out certain border requirements, and targeting four infectious diseases – cholera, plague, yellow fever, and smallpox (removed after eradication in 1980) – it was hoped that these four diseases could be stopped at international borders. However, countries often did not report these diseases when they occurred because of fear of irrational trade and travel measures and the severe negative economic impact that could occur.

In addition, as knowledge about emerging infections grew, it became clear that there were other infectious diseases of equal or greater potential for international spread than those that were covered by the IHR, and a revision of the IHR was begun in the late 1990s.

By the time of the SARS outbreak, a new way of detecting and responding to infectious disease outbreaks had been developed by WHO as a precursor to the revision of the IHR, and it was these ways of working that led to the coordinated global response to SARS. One of the major lessons learned was that strong national disease detection and response systems were of great importance in order that countries could detect and response to infectious disease outbreaks where and when they occur, thus preventing human suffering and death, and minimizing the risk of international spread.

This concept was incorporate into the revised IHR that now required all countries to develop a minimum core public laboratory and epidemiology capacity in order to detect and respond to outbreaks when and where they occur. The revised IHR also continue a requirement for reporting of disease outbreaks, and the requirement has been broadened to reporting all public health events of international concern (PHEIC) after risk assessment using a decision tree provided in the IHR.

however, gives cause for hope that the revised IHR do indeed offer a means of ensuring maximum security against the international spread of infectious diseases, while minimizing interference with travel and trade.

Reports of persons infected with this newly identified SARS-like virus were made to the WHO from countries treating patients with origins in the Kingdom of Saudi Arabia, Qatar, Jordan, and the United Arab Emirates. The initial reports originated at the time of religious pilgrimage for Hajj. An irrational response could have caused great confusion and a heavy economic and spiritual loss to pilgrims and to Saudi Arabia, which has increased its investment each year to provide for the health security of pilgrims.

An immediate and transparent risk assessment was made under the framework of the revised IHR, and the risk was then communicated widely. The Hajj was unaffected by the reports of the risk assessment, and surveillance of Hajj pilgrims for severe respiratory symptoms was conducted during the pilgrimage and after pilgrims had returned to their home country.

Only time will tell whether the new ways of working and communicating risk under the IHR will continue to help prevent unnecessary panic and confusion when an outbreak occurs and spreads internationally; and prevent the irrational reaction that increases their negative economic impact.

See also: HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Infectious Disease Externalities. Macroeconomic Effect of Infectious Disease Outbreaks. Water Supply and Sanitation

Further Reading

BBC. (2011a). *E. coli* cucumber scare: Spain angry at German claims. Available at: http://www.bbc.co.uk/news/world-europe-13605910 (accessed 03.10.11).

BBC. (2011b). *E. coli*: Russia bans import of EU vegetables. Available at: http://www.bbc.co.uk/news/mobile/world-europe-13625271 (accessed 03.10.11).

Bloom, E., de Wit, V. and Carangal-San Jose, M. J. (2003). ERD Policy Brief No. 42 – Potential economic impact of an avian influenza pandemic in Asia. Asian Development Bank. Available at: http://www.adb.org/Documents/EDRC/Policy_Briefs/PB042.pdf (accessed 12.10.11).

Heymann, D. L. and Rodier, G. (2004). SARS: A global response to an international threat. *Brown Journal of World Affairs, Winter/Spring* **X**(2), 185–197.

Smith, R. D., and Sommers, T. (2003). Assessing the economic impact of public health emergencies in international concern: The case of SARS. *Globalization, Trade and Heath Working Papers Series*. Geneva: World Health Organization

The World Bank. (2005). Avian flu: Economic losses could top US$800 billion. Available at: http://go.worldbank.org/E0YSLRS140 (accessed 12.10.11).

Empirical Market Models

L Siciliani, University of York, Heslington, York, UK

Introduction

This article reviews econometric techniques and studies aimed at characterizing the market structure in the health sector. It focuses on the following issues: (1) the effect of competition on hospital quality, efficiency, and prices (if they are not fixed by a regulator); (2) differences in behavior that arise from different types of ownership status (non-profit vs. for-profit); (3) the extent to which demand for healthcare responds to quality; (4) the effect of mergers on cost savings, prices, and quality; and (5) the use of report cards and their impact on quality and providers' incentive to select low-severity patients.

These research questions have potentially important policy implications. Governments can encourage or discourage competition, or regulate it. They can favor one ownership status over another by introducing favorable tax regimes or by making a certain ownership mandatory. They may forbid or allow mergers through antitrust authorities and legislation. They can make report cards mandatory and publicly available.

The article focuses on key theoretical predictions, econometric strategy, empirical specification, and possible biases which may arise in testing such predictions. It also summarizes the main empirical findings for each theme. Moreover, because space is limited, the focus is on the hospital sector. Therefore, issues related to insurance markets, the pharmaceutical industry, provider labor markets, and the market of nursing and care homes are not investigated.

Effect of Competition on Quality and Prices

The effect of competition on hospital behavior has been the subject of an extensive empirical literature. One key focus has been on testing the effect of competition on the quality of hospital care under two main institutional settings: (1) a fixed-price regime of the Diagnosis-Related Groups (DRG) type, where each hospital receives the same price to treat a patient with a given diagnosis (this is the case in Medicare in the USA and in many European countries); (2) a variable-price regime, where each hospital is free to set prices in a private competitive market (like in the USA) or prices are the result of a bargaining procedure between the purchaser of health services (a private or a public insurer) and the hospital.

Under the first regime, the standard prediction from economic theory is that higher competition should lead to higher quality. Because more competition makes the demand more responsive to a marginal increase in quality (and prices are fixed), hospitals have a stronger incentive to increase quality because it will attract a larger volume of patients and generate higher revenues. Under the second regime, the prediction is less clear-cut. More competition will also reduce price and the price-cost margin of each hospital, therefore, weakening the incentive to increase quality. This effect goes against the former one (in terms of higher demand responsiveness) so that competition may lead to an increase or a reduction in quality depending on the size of the two opposing effects. A similar ambiguous prediction is that if prices are easily contractible, whereas quality is not, more competition may lead to a large reduction in price at the expense of a large drop in quality.

The basic empirical strategy within a cross-sectional framework to test the above predictions is the following:

$$q_i = \alpha + \beta c_i + \gamma z_i + \varepsilon_i, \qquad i = 1, \ldots, N \qquad [1]$$

where q_i is the quality provided by hospital i, c_i is a measure of competition, and z_i is a vector of control variables which also affect quality (e.g., volume of patients treated to control for learning-by-doing, dummies for different types of hospital, etc.).

There are different ways to measure competition in the health sector, which involves two main steps. The first step involves the definition of the catchment area of each hospital i, which gives the geographical area covering the potential competitors of hospital i (the area over which the hospitals 'compete'). There are two main approaches to define catchment areas, which is based on: (1) a fixed radius, that draws a circle of 30 km (or an alternative distance of 20, 40 km) from the hospital; or a fixed travel time, that uses road maps to define a catchment area of 30 min travel time from the hospital (or alternative times: e.g., 20 or 40 min); (2) a variable radius technique, where the catchment area is based on the residence (as measured by their postcode) of the patients going to hospital i: the catchment area is defined, for example, on the residence of the 70% of patients living closest to the hospital (or an alternative proportion like 60% or 80% decided by the researcher). Note that not all patients are included (100%) because this would often imply that the catchment area of some hospitals includes the whole country, which is clearly unrealistic (there will often be at least one odd patient who traveled from far away or whose postcode is mistakenly recorded). Fixed radius models are simpler to compute but ignore the actual residence of the patients going to each hospital i. Variable radius models are more accurate because they address this problem but computationally more intensive. They also raise some endogeneity issues: hospitals with higher quality may have larger catchment areas. This is usually addressed by defining catchment areas on the basis of predicted rather than actual hospital choice. In practice, this involves estimating a multinomial logit model of a patient's choice as a function of distance and other key regressors. Predicted market shares are then computed for each hospital and used to compute a competition measure.

Once the catchment area has been defined, the second step involves measuring the degree of competition within this area. The simplest way to measure competition is to count the number of hospitals (N) within the catchment area. Equivalently, the degree of concentration can be measured by

$1/N$. However, this measure has the disadvantage that it implicitly assumes that all hospitals have the same size: the market structure of a duopoly where each hospital has 50% of the market can be quite different from one where one hospital has 90% of the market and the other only 10% of the market. In the latter case, the market is less competitive than in the former one because one provider has a dominant position.

A modified version of the simple competition measure is the number of hospitals in the catchment area divided by the population of the catchment area (P): the measure is therefore N/P. For a given number of providers, areas with larger population effectively imply a lower degree of competition.

A second measure which takes into account the different size of each hospital is the widely used Herfindahl Index (HI) define the market share of each hospital i as $s_i = \gamma_i/Y_i$, where γ_i is the number of patients treated by hospital i and Y_i is the total number of patients treated within the catchment area of hospital i. The HI is given by the sum of the square of each market share: $HI_i = \sum_{i=1}^{n} s_{ji}^2$. Note that if all hospitals are identical, then the HI_i is equal to $1/N$, and the two measures (HI and the reciprocal of the number of hospitals) coincide. However, if the market shares are different then the two measures will differ. Suppose there are only two providers ($N=2$) and that one hospital has 25% of the market whereas the other hospital has 75% of the market. Then, the HI is $1/4+9/16=0.81$, which is larger than 0.5 (the HI when each provider has 50% of the market). The idea is that an asymmetric market is a less competitive one. As mentioned above, one problem with the computation of the HI based on 'actual' market shares is that these can be endogenous if, for example, hospitals with higher quality have larger market shares. To address this problem, the HI is often computed on the basis of predicted market shares (based on multinomial choice models).

Quality of care is the other key variable in the empirical model described in eqn [1]. It can be measured in a variety of ways. The most common one in recent literature makes use of mortality rates for (emergency) patients with acute myocardial infarction, more commonly known as 'heart attack.' These are considered to be a marker of the quality in the hospital. Other measures include total hospital mortality rates (adjusted by casemix), mortality rates for patients with stroke, pneumonia, heart failure, and other specific conditions, readmission rates within a month of discharge, and infection rates. In general there have been mixed findings in the literature on the effect of competition on quality with prices either fixed or variable and endogenously determined (see Gaynor and Town (2011) for a detailed review).

A similar approach to the one described in eqn [1] can be used to estimate the effect of competition on prices charged by hospitals by replacing the dependent variable q_i with price p_i. The empirical evidence is mainly from the USA, for the market not covered by Medicare and Medicaid (where prices are fixed), and confirms the expected negative effect between competition and prices. There is limited evidence from Europe where prices are regulated (and therefore fixed) in several publicly funded systems: Competition on price occurs mainly in the private sector, which is often small and data on prices are difficult to collect.

Ownership

A long-standing question in the health economics literature is whether profit and non-profit hospitals differ in their behavior. Most of the literature has focused on differences in quality and efficiency (with more recent studies focusing on quality). A fewer number of studies has focused on differential incentives to upcode (also known as DRG creep) and to select more profitable patients. Regarding quality, on one hand non-profit hospitals may have an incentive to provide higher quality as they are under less pressure to increase profits; on the other hand, they are less responsive to demand variations. Standard economic theory also predicts non-profit hospitals to be less efficient because they cannot appropriate the financial surplus (or distribute it), they may have weaker incentive to keep costs down (or to be more efficient).

The typical basic regression for quality differences in a cross-sectional framework is the following one:

$$q_i = \alpha + \beta s_i + \gamma z_i + \varepsilon_i, \qquad i = 1, \ldots, N \qquad [2]$$

where q_i is the quality provided by hospital i, s_i is a dummy variable for hospital status and is equal to 1 if hospital is for profit, and z_i is a vector of control variables. Quality can be measured through mortality rates and adverse events such as surgical complications and medical errors. The empirical evidence from the USA is extensive but mixed. The recent review by Eggleston *et al.* (2008) find that whether for-profit hospitals provide lower or higher quality than non-profit ones depends on the specific context like the region, the data source, and the period of analysis. As an overall conclusion they suggest that as a whole quality seems to be lower among for-profit hospitals.

Some recent studies rely on a panel-data approach as opposed to a cross-sectional one, focusing on the effect of changes in ownership status over time (either from non-profit to for-profit or from for-profit to non-profit). This approach allows controlling for unobserved heterogeneity, i.e., the possibility that differences in quality between for-profit and non-profit hospitals simply reflect different location, catchment areas, casemix, or other unobservable variables. The econometric framework is therefore modified as follows:

$$q_{it} = \alpha + \beta n_{it} + \delta f_{it} + \gamma z_i + d_i + d_t + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T \qquad [3]$$

where n_{it} is a dummy variable equal to 1 from the time a hospital converts from for-profit to non-profit, f_{it} is a dummy variable equal to 1 from the time a hospital converts from non-profit to for-profit, d_i accounts for hospitals' fixed effects to control for unobserved heterogeneity, and d_t is a vector of year dummies to control for a time trend (e.g., health outcomes improve over time due to technology development).

Some literature finds that the change in status from non-profit to for-profit reduces quality as measured by higher mortality rates for patients with heart attacks (deaths at 1, 6, and 12 months). One potential problem with such approach is that the switch from non-profit to for-profit status may not be random. For example, it can be argued that hospitals with declining quality are more likely to change status. To address this issue, authors have interacted the dummy variables on

hospital conversion *(f, n)* with time dummies for the years preceding and following the conversion. This allows detecting whether the converting hospitals were already exhibiting a decline in quality before the conversion.

Others instead have addressed the issue by using a matching-estimator approach, for example, using propensity score matching to identify a control group which has a distribution of covariates that is in line with the distribution of the covariates in the treatment group. The estimation procedure first estimates the conditional probability of a hospital being for-profit for a given set of covariates (the propensity score) and then it matches each hospital (which switches from non-profit to for-profit) with a control hospital which has the closest propensity score. The covariates on which the hospitals are matched include hospital size, patient types, and financial state.

Most studies treat ownership as exogenous in eqn [2]. However, that may not necessarily be the case. For instance, patients may choose the type of hospital (for-profit vs. non-profit) based on how severe they are and this may generate endogeneity: the quality of care (the dependent variable) affects who goes to a profit versus a non-profit hospital. One strategy is to use an instrumental-variable approach with instruments that include the distance to the closest non-profit hospital, and the difference in the distance between the closest non-profit hospital and the closest hospital (regardless of being for-profit or non-profit). Distances will affect the choice between a for-profit and a non-profit hospital, but should not be correlated with patients' severity.

For-profit and non-profit hospitals may also differ in their incentive to upcode, i.e., to code patients in more remunerative fields. Payment systems of the DRG-type are complex and involve at least 500 different prices that depend on patient's diagnosis and treatment. There is evidence in the USA that for-profit hospitals tend to upcode more than non-profit ones. Moreover, private hospitals, regardless of the for-profit or non-profit status, may engage in cream-skimming of patients leaving the unprofitable ones to the public sector.

Regarding differences in efficiency, in his review of 317 published papers on frontier efficiency measurement, Hollingsworth (2008) concluded with some caution that public/non-profit hospitals tend to be more efficient than for-profit ones. The intuitive result that for-profit hospitals are more efficient than non-profit ones is therefore not confirmed in general. Efficiency is generally measured through parametric models, i.e., the estimation of stochastic frontiers, or nonparametric ones, i.e., data envelopment analysis. Some parametric studies focus on technical efficiency and derive efficiency scores by estimating the following production frontier model (within a cross-sectional framework):

$$y_i = \alpha + \beta x_i + \gamma z_i - e_i + \varepsilon_i\,, \qquad i = 1, \ldots, N \qquad [4]$$

where y_i is typically the number of patients treated by hospital i (weighted by DRG weight to control for different casemix of the hospital), x_i includes a range of inputs (number of beds, doctors, nurses), and z_i includes a range of control variables (ideally quality); e_i is hospital efficiency and ε_i is the error term. This model requires assumptions about the distribution of the efficiency term. The most common ones are the Half Normal, Truncated Normal, and Gamma. The efficiency scores derived following this methodology have been criticized for two main reasons: (1) they seldom control adequately for quality differences, so that efficiency scores may reflect higher quality; (2) they may be sensitive to outliers and the specific distributional assumptions of the efficiency term. The approach in eqn [4] has been extended to allow for multiple outputs (e.g., patients in different specialties, emergency vs. nonemergency patients, outpatients vs. inpatients). This can be pursued with a 'Shepard distance function' approach that ultimately involves using one output on the left-hand side (LHS) and the other 'normalized' outputs on the right-hand side (RHS) or a 'polar coordinates' approach using the Euclidian norm of the outputs on the LHS and polar coordinates angles on the RHS. These approaches can be criticized on the ground that output variables appear both on the LHS and the RHS of the regression model, possibly generating endogeneity.

Equation [4] focuses on technical efficiency. Other studies focus on allocative efficiency as well by estimating a cost frontier as opposed to a production frontier. In such a case the model is:

$$C_i = \alpha + \beta y_i + \gamma w_i + \sigma z_i + e_i + \varepsilon_i, \qquad i = 1, \ldots, N \qquad [5]$$

where C_i is total cost of hospital i, y_i is (a vector of) output, w_i is the (average) salary for different types of workers (doctors, nurses, administration), and z_i is a range of control variables (quality, whether the hospital has teaching functions, etc.). This approach has also the advantage of accommodating multiple outputs without any additional assumptions.

As mentioned above, stochastic frontier techniques have been criticized for imposing distributional assumptions on the efficiency term and to rely on these to disentangle efficiency from noise. These assumptions can be relaxed by using panel data and estimating models of the following type:

$$y_{it} = \alpha + \beta x_{it} + \gamma z_{it} - e_i + \varepsilon_{it}, \qquad i = 1, \ldots, N \qquad [6]$$

where e_i is a fixed effect at hospital level. The distributional assumptions are weaker. This approach still relies on having good control variables (e.g., on quality) so that e_i can be interpreted as efficiency as opposed to a control for unobserved heterogeneity (where efficiency is only one determinant). Once the efficiency scores are obtained, the second step simply involves regressing the efficiency scores on hospital's ownership type and other determinants.

Choice Models

At the heart of many health economic models is the assumption that demand of healthcare providers responds to quality. Providers with higher quality establish a good reputation and attract a larger number of patients. The estimation of the magnitude of the demand elasticity to quality has implications for policy design. For example, if hospital elasticity is high, policymakers will need to rely less on costly audits to ensure high standards of quality. Providers will have an incentive to provide high quality in order to attract patients and increase revenues. Similarly, one precondition for competition to encourage quality of care (already discussed above) is that demand responds to quality.

The assumption that providers' demand responds to quality has been tested empirically by modeling patients' choice of a hospital among a set of alternative ones. A common model is the conditional logit model which can be motivated within a random utility framework (McFadden, 1974). Suppose that the utility of patient j choosing hospital i is equal to $U_{ji} = \beta d_{ji} + \gamma q_j + \varepsilon_{ji}$, where d_{ji} is the distance between patient's j residential address and hospital i address, q_j is the quality of hospital j (e.g., mortality rates, readmission rates), and ε_{ji} is the unobserved component of utility. If ε_{ji} are independently and identically distributed, and follow type 1 extreme value distribution, then the probability of patient j choosing hospital i out of a total of N hospitals is given by:

$$p_{ij} = \frac{\exp(\beta d_{ij} + \gamma qi)}{\sum_{l=1}^{N} \exp(\beta d_{lj} + \gamma ql)}, \qquad i = 1, \ldots, N \qquad [7]$$

which is known as the conditional logit model.

The analysis is usually conducted for patients in need of a specific treatment (i.e., coronary bypass, percutaneous transluminal coronary angioplasty, kidney transplant, cataract surgery, hip replacement) or with a certain condition (i.e., acute myocardial infarction, pneumonia). A key regressor (or control variable) is the distance between the patient's residence (postcode) and the hospital, which in all models turns out to be the main predictor of patients' choice. The hospital choice is also affected by quality, as proxied by mortality rates, readmission rates, complication rates, and waiting times. Overall, this empirical literature finds that higher quality (as well as distance) increases the probability of choosing a provider, though the demand elasticities with respect to quality are small for most procedures and conditions.

To control for time-invariant unobserved heterogeneity, some studies estimate the conditional logit with panel data including hospital fixed-effects, therefore relying on variations in quality (mortality rates, readmission rates, waiting times) over time to identify the causal effect of quality on demand. One limitation of the conditional logit is that the relative probability of choosing any two hospitals is independent of any other alternative hospital (known as the independence of irrelevant alternatives). The logit models can also be extended to allow for latent classes (latent-class multinomial logit) and therefore allow the responsiveness of demand to quality to vary for different classes of patients (normally two), which are not observable to the researcher.

Mergers

A growing empirical literature has investigated the effect of mergers on efficiency (cost savings), prices, and quality. From a theoretical perspective, hospital mergers can lead to reductions in costs and an increase in efficiency through better management, exploitation of scale economies, and elimination of duplicate services. From an antitrust perspective, a merger also increases the market power of merging hospitals, which may allow them to increase prices at the expense of consumers. Therefore, one may expect price to reduce following a merger when the efficiency savings, which tend to reduce price, overcome the reduced competition effect, which tends to increase price. The lower degree of competition may also induce merged hospitals to skimp on quality because demand is less responsive to quality changes.

The basic econometric framework is the following:

$$y_{it} = \alpha + \beta m_{it} + \gamma x_{it} + d_i + d_t + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T \qquad [8]$$

where y_{it} is either cost, quality, or price, m_{it} is a variable equal to 1 from the time the hospital has merged (and 0 otherwise), x_{it} includes a range of controls. Note that for 'merging' hospitals, i refers to the sum of the costs of the two merged hospitals or the average price or quality in the merging hospitals.

One econometric problem with empirical studies evaluating the effects of mergers is that mergers may be endogenous: for example, a hospital merges because costs are high or quality is low (so that m_{it} depends on y_{it}). One way to account for such endogeneity is through the use of propensity score matching. This involves the estimation of a probit that models the probability of merging for each hospital i as a function of set of characteristics (the number of hospitals in the market, whether the hospital is for profit, non-profit, or a teaching hospital, etc.). Hospitals are then matched on the basis of predicted probabilities, i.e., the propensity to merge. Another potential econometric issue is that nonmerging hospitals may also react to mergers, for example, by also increasing prices or reducing quality, and may therefore not act as a good control group (Dafny, 2009). To address this issue, she uses as an instrument a variable which measures whether hospitals are colocated, the idea being that distance should be correlated with the probability of merging but not with the outcomes. 'Regression to the mean' may also be an issue if hospitals with high cost are followed by periods of low cost.

Most studies find that prices increase following a merger (Gaynor and Town, 2011). Dranove and Lindrooth (2003) find that in the USA mergers reduce hospital costs by approximately 14% during the 2–3 years following the merger. Previous studies have generally not found much evidence of cost savings. Ho and Hamilton (2000) find that mergers in California have no effect on the quality of care as measured by mortality rates for patients with heart attack and stroke, though readmission rates and early discharges for newborns increased in some cases. Gaynor Laudicella and Propper (2012) examine the impact of large number of mergers in England on a range of outcomes including financial performance, productivity, waiting times, and clinical quality. They find that mergers had no effect on quality.

Report Cards

Report cards are increasingly used in the healthcare sector to provide information on the quality of healthcare providers. They are intended mainly to help patients choosing the provider which matches better the needs of the patient, to improve choice and to encourage providers to increase quality in order to attract more patients. Typically, report cards provide mortality rates and readmission rates for specific conditions or procedures, coronary bypass being the most

common one. In the USA, the State of New York was among the first to introduce such cards and for this reason has been intensively investigated in the empirical literature. Report cards can be provided at hospital or at doctor/surgeon level.

Because report cards have been introduced in different states at different times (and never introduced in some states), their effect is often investigated within a natural experiment set up with some states in the USA acting as the control group. There is evidence that market shares may be influenced by report cards with providers with better reports having larger market shares. One potential adverse effect of report cards is that they may encourage providers to treat (select) patients with lower severity who are at a lower risk of mortality and readmission. Dranove *et al.* (2003) provide evidence for such selection behavior, observing that the introduction of report cards was followed by a reduction in the average severity of illness, as measured by hospital utilization before admission, with the severity of patients in teaching hospitals instead increasing.

Conclusion

This article has reviewed econometric techniques and studies aimed at characterizing the market structure in the hospital sector. A range of econometric techniques have been employed to investigate the effect of competition, differences in behavior by ownership status, demand responses to quality, mergers, and report cards. Several studies make use of natural experiments exploiting exogenous shocks (e.g., in the evaluation of competition or report cards). If control groups are not well defined, propensity score matching has been used to account for self-selection and create pseudo control groups (e.g., in the case of conversion of for-profit to non-profit hospitals and mergers). When natural experiments are not available, endogeneity caused by unobserved heterogeneity or reverse causality is an issue. In such cases, panel data and instrumental variables have been used (e.g., in the evaluation of for-profit vs. non-profit hospitals). Conditional logit models have been usefully employed to estimate the responsiveness of hospital demand to quality.

As a whole these studies suggest that the market structure matters in the health sector, though not always in the expected way, and that the results may differ depending on country, outcome measure, and econometric methodology employed.

See also: Competition on the Hospital Sector. Cost Function Estimates. Markets in Health Care. Models for Discrete/Ordered Outcomes and Choice Models. Quality Reporting and Demand

References

Dafny, L. (2009). Estimation and identification of merger effects: An application to hospital mergers. *Journal of Law and Economics* **52**(3), 523–550.

Dranove, D., Kessler, D., McLellan, M. and Satterthwaite, M. (2003). Is more information better? The effects of 'report cards' on health care providers. *Journal of Political Economy* **111**(3), 555–588.

Dranove, D. and Lindrooth, R. (2003). Hospital consolidation and costs: Another look at the evidence. *Journal of Health Economics* **22**, 983–997.

Eggleston, K., Shen, Y., Lau, J., Schmid, C. H. and Chan, J. (2008). Hospital ownership and quality of care: What explains the different results in the literature? *Health Economics* **17**(12), 1345–1362.

Gaynor, M. and Town, R. J. (2011). Competition in health care markets. In Pauly, M., McGuire, T. and Barros, P. P. (eds.) *Handbook of health economics*, chap. 9, pp 499–637. North-Holland: Elsevier.

Gaynor, M., Laudicella, M., Propper, C. (2012). Can governments do it better? Merger mania and hospital outcomes in the English NHS, The Centre for Market and Public Organisation, 12/281, Department of Economics, University of Bristol.

Ho, V. and Hamilton, B. H. (2000). Hospital mergers and acquisitions: Does market consolidation harm patients? *Journal of Health Economics* **19**, 767–791.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics* **17**, 1107–1128.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In Zarembka, P. (ed.) *Frontiers in econometrics*, no. 4, pp 105–142. New York: Academic Press.

Equality of Opportunity in Health

P Rosa Dias, University of Sussex, Brighton, UK

Background

Normative Context

In recent years, the concept of inequality of opportunity, rather than inequality of achieved states, has received growing attention in the economic literature. The simple advocacy of equal health, for example, fails to hold individuals accountable for their choices. This can be seen as significant limitation.

Equality of opportunity co-opts one of the sharpest ideas in the antiegalitarian arsenal: The notion of responsibility. By compensating for the impact of circumstances beyond individual control, yet holding individuals responsible for the consequences of their choices, equality of opportunity is an appealing compromise between strict equality of health and mere equity of formal rights. It has thus been increasingly advocated by policy makers, as is made clear in World Bank (2005) which focuses on the inequality issue.

This theoretical evolution reflects a number of recent developments in political philosophy, arguably prompted by the seminal work of John Rawls and Amartya Sen. Both Rawls' equality of social primary goods and Sen's proposed equality of capabilities move away from the social goal of equalizing subjective welfare. They propose that, once primary goods or capabilities are equally distributed, any residual inequality should be deemed a legitimate consequence of individual choice, hence of individual responsibility. Ronald Dworkin advanced this proposal by arguing that equality of welfare cannot be a valid equity criterion for it fails to make individuals accountable for their preferences, namely, those preferences they are happy to have. The problem thus becomes one of finding the distribution of resources that appropriately compensates individuals for their dissimilar endowments (physical resources, talents, and handicaps), while making them responsible for their preferences. This rationale leads Dworkin to propose the criterion of equality of resources, which attracted important criticisms, such as those raised by Richard Arneson and Gerald Cohen. Cohen shows that Dworkin's separation between preferences and resources can be intractable in practice: Should one be made responsible for childhood preferences that are chiefly instilled by one's social environment? This debate has prompted key progresses in social choice theory, which have rendered these new ideas operational within an analytical framework known as the equal-opportunity approach.

The Roemer Model of Equality of Opportunity

Equality of opportunity has been given different formal expressions in the social choice literature, such as in early proposals by Marc Fleurbaey and Walter Bossert. A related strand of research focuses on measuring opportunity sets, taking into account the intrinsic value of individual freedom in the ranking of social states. Despite the theoretical appeal of these contributions, they have proved too abstract to prompt related empirical work. Largely for this reason, the workhorse of the applied literature on inequality of opportunity in health has been the model proposed by Roemer (1998, 2002).

The Roemer model sorts all factors influencing individual attainment between a category of effort factors, for which individuals should be held responsible and a category of circumstance factors, which, being beyond individual control, are the source of illegitimate differences in outcomes. It should be noted that, in this framework, effort is not limited to human exertion and comprises all the determinants of health outcomes over which individuals have control. Also, the classification of the determinant of human achievement as either circumstances or effort is partly normative and partly informed by available empirical evidence. In the case of health, we may think of the outcome of interest as health as an adult (H) and define a health production function, $H(C,E(C))$ where C denotes individual circumstances and E denotes effort. Circumstances affect the health outcomes of individuals and social groups, directly and through their influence on effort factors.

The recent medical and economic evidence on the early determinants of health has emphasized the importance of a number of circumstantial factors. The fetal origin hypothesis stresses the role of parental socioeconomic characteristics as key determinants of *in utero* fetal growth which, in turn, condition long-term health. The life course models, which emphasize the impact of deprivation in childhood on adult health and longevity, and the pathways models, suggest that health in early life is important mainly because it will condition the socioeconomic position in early adulthood, which explains disease risk later in life. There is also evidence on determinants of health that, although affected by circumstances, are, at least partially, within individual control and therefore constitute effort factors in the context of the Roemer model. Lifestyles such as diet and physical exercise are good examples of such factors.

The Roemer model defines social types consisting of the individuals who share exposure to identical circumstances. Types can thus be defined using the set of observed individual circumstances in the data. In practice, it is up to the researcher to identify circumstances that lead to a meaningful partition of the population of interest. Factors such as parental socioeconomic background and region of birth have often been used by applied economists to partition the population, but other variables such as inborn cognitive ability and childhood health have also been used. It is assumed that the society has a finite number of T types and that, within each type, there is a continuum of individuals. A fundamental aspect in this setting is the fact that the distribution of effort within each type (F^t) is itself a characteristic of that type (t); because this is beyond individual control, it constitutes a circumstance.

 doi:10.1016/B978-0-12-375678-7.00210-8

In general, it is not possible to compare directly the levels of effort expended by individuals from different types because circumstances partly determine outcomes. For example, the number of times per week one does physical exercise is partly determined by individual choice (effort) and also influenced by parental background, social milieu, and peer pressure (circumstances). Thus, two individuals who exercise exactly the same number of times per week, may be interpreted as exerting very different levels of effort, depending on their circumstances. To make the degree of effort expended by individuals of different types comparable, Roemer proposes the definition of quantiles of the within-type effort distribution (e.g., the distribution of weekly frequency of physical exercise within each type): Two individuals from different types are deemed to have exerted the same degree of effort if they sit at the same quantile (π) of their type's distribution of effort. When effort is observed, this definition is directly applicable. However, if effort is unobservable, an additional assumption is required: By assuming that the average outcome, health in this case, is monotonically increasing in effort, i.e., that healthy lifestyles are a positive contribution to the health stock, effort becomes the residual determinant of health once types are fixed; therefore, those who sit at the πth quantile of the outcome distribution also sit, on average, at the πth quantile of the distribution of effort within this type.

How is the equality-of-opportunity policy characterized in this framework? Ideally, this policy should ensure identical health across types at identical levels of effort. Let us assume that, given our health production function, the highest health level attainable by type (t) given quantile level of effort (π) and policy (ϕ) is given by the indirect outcome function $v^t(\pi, \phi)$. In this setting, the equality-of-opportunity policy πth equalizes the highest attainable health level across types for identical values of π, i.e., $v^t(\pi, \phi^{\mathrm{E-opp}}) = v^{t'}(\pi, \phi^{\mathrm{E-opp}})$.

In addition, because the resources available for policy interventions are generally finite, one also needs to ensure that $\phi^{\mathrm{E-opp}}$is feasible. However, this poses a problem: As shown in Roemer (2002), it will not be possible, in general, to find an equality-of-opportunity policy that simultaneously satisfies the feasibility requirement. Thus, in practice, instead of literally equalizing v between types at each π, one maximizes the minimum value of v across types at each π.

But we are not finished yet. In general one is not interested in finding the equality of opportunity for a sole particular value of π: Healthcare policy does not usually apply only to those at say, the qth quantile of weekly frequency of physical exercise. The problem is that there are different optimal policies for different values of π, even if interest in the subset of efficient policies is restricted. So how is the equality-of-opportunity policy found? A number of compromise solutions have been suggested in the literature. The most widely used in practice (proposed by Roemer (2002)), consists of aggregating over all policies (each defined for a particular value of π) and giving each of them the same weight.

Ex Ante and *Ex Post* Inequality

So far, this account of inequality of opportunity has focused on inequalities between groups of individuals, called types, who share exposure to identical circumstances. This approach is the most prevalent in applied work and is known as the *ex ante* approach. The term *ex ante* refers to the fact that this approach can be used in cases where circumstances are known, but effort has not (yet) been exerted by the individuals.

There is, however, an alternative approach to the concept of equality of opportunity. Assume that effort is observed. The population of interest can thus be split into groups, known as tranches, which correspond to levels of exerted effort (e.g., number of times per week one does physical exercise). In this setting, inequality of opportunity corresponds to differences in outcomes within each tranche, i.e., amongst individual who have exerted the same level of effort. The source of unjust inequalities is still the variation across individual circumstances, but this line of research is known as the *ex post* approach, because it requires information on the level of effort already exerted by individuals.

An important question is the extent to which the *ex ante* and the *ex post* approaches are similar. Although they share points in common, they are fundamentally different. As mentioned earlier, equality of opportunity requires the elimination of differences in outcomes that are due to circumstances but not to effort. This is known as the principle of compensation. It should however be noticed that this principle of compensation leads to different compensatory policies according to whether one takes the *ex ante* or the *ex post* approach. In the *ex ante* case, compensation requires transferring resources from individuals in the most advantaged types to people in least advantaged ones. But in the *ex post* approach, the required transfers are within-tranche transfers, amongst individuals who exert the same level of effort. Thus, *ex ante* and *ex post* compensation are generally incompatible.

Another important aspect concerns the definition of fair rewards to effort. Individuals with the same circumstances are considered. The theory of equality of opportunity described so far is silent regarding the fair way of rewarding different levels of effort amongst such individuals. There is at present an intense debate on the way to combine the compensation principle with a suitable reward principle, but a definite solution has not yet been reached. Fleurbaey and Schokkaert (2012) describe a number of possible avenues for achieving this within the framework of equality of oportunity, although J. Roemer has recently argued that the definition of what constitutes fair rewards to effort should instead come from an ancillary theory, which limits the degree of inequality that is acceptable. In addition, Fleurbaey and Peragine have shown that the available options for combining the principles of compensation and reward depend vitally on whether the researcher adopts the *ex ante* or the *ex post* approach. Although, in general, the principles of compensation and reward are theoretically incompatible, this conflict can be avoided when one adopts the *ex ante* approach (but not the *ex post* one).

Partial Orderings and Inequality Measures

How can inequality of opportunity be identified in practice? A number of different approaches have been proposed, based on partial equality-of-opportunity orderings. The most widely

used in applied work defines equality of opportunity based on stochastic dominance conditions. The rationale is the following. Denoting by $F(.)$ the cumulative distribution function of health (CFD), a literal translation of the idea of equality of opportunity would correspond to the situation in which the distribution of health outcomes does not depend on social types, i.e., $F(.|t) = F(.|t')$. This condition is, however, unlikely to hold in any society and hence is too stringent to be applied to real data. Instead one could assume that the data be deemed consistent with the existence of inequality of opportunity when the social advantage provided by different circumstances can be unequivocally ranked by stochastic dominance criteria, i.e., $F(.|t) \succ_{SD} F(.|t')$. First-order stochastic dominance (FSD) holds for the whole class of increasing utility functions; thus if the distribution of health outcomes of type t FSD dominates that of type t', this means that all individuals with an utility function that is increasing in health (i.e., who prefer better health to worse health) would prefer the outcomes of type t to those of type t'. Although one may extend this partial ordering to second- and even third-order stochastic dominance criteria, most of the applied literature has been focused on first-order comparisons. These are better suited for the ordinal outcomes that are often used in health economics, such as self-assessed health. Moreover, in addition to their clear meaning in terms of welfare and preferences, these conditions have an important attractive feature: They are statistically testable in practice.

Partial orderings are useful but often inconclusive, hence complete orderings have been proposed to measure inequality of opportunity. In this literature, an analysis of inequality of opportunity in Brazil carried out by Bourguignon *et al.* (2007) prompted a number of methods collectively known as the parametric approach to the measurement of inequality of opporunity. The idea is intuitive. Earlier, the definition of the health production function of individual health outcomes, $H = f(C, E(C))$ was given. The same specification applies to the health outcomes of social groups. Thus, a parametric regression model can be used to estimate the counterfactual distribution of outcomes that would be brought about by assigning the same circumstances to all the individuals, i.e., $\tilde{H} = f(\tilde{C}, E(\tilde{C}))$. Inequality of opportunity can then be measured by an index $\Gamma = 1 - \frac{\tilde{H}}{H}$.

A different approach, known as nonparametric supposes that one replaces each individual outcome in H by one's type-specific mean (μ^t), obtaining the smoothed distribution of outcomes H^C. This eliminates, by construction, all within-type inequality, hence a relative inequality index $I(H^C)$ measures exclusively between-types disparities, which constitute inequality of opportunity. Alternatively, one may replace the outcome of each individual i outcome (h_i) by $\frac{\mu}{\mu^t} h_i$, where μ is the mean in the population of interest, obtaining the standardized distribution H^E. In this case, all the between-types inequality is eliminated, leaving solely within-type inequality. As a result, inequality of opportunity corresponds to the difference between the total inequality in health outcomes, $I(H)$, and the inequality measured by $I(H^E)$.

Two important practical issues arise in this context. The first concerns the choice of an appropriate inequality index I, given that, in general, the smoothing and standardizing approaches lead to different results. There is a class of measures (known as path-independent decomposable measures, for which these two approaches lead to the same result. Amongst this family of measures, the mean-log deviation has been very widely used in applied work. The second issue of interest is that of choosing between the parametric and the nonparametric approaches. Nonparametric methods are, in general, more robust in the sense that they do not depend on parametric assumptions. They are, however, more data-hungry: When the information on the circumstances set is rich the number of types increases, leading to data insufficiency. This is less of a problem for the parametric approach, which, in addition, allows the estimation of partial effects, namely, circumstance-specific inequality shares. Nonetheless, this comes at the expense of an increased reliance on structural assumptions.

Another index that has been increasingly used in applied work is the Gini-opportunity index proposed by Lefranc *et al.* (2009). This is a modified Gini coefficient that quantifies the inequality between the different types' opportunity sets. The area underneath a type's generalized Lorenz curve, and hence the value of its Sen evaluation function $A_j = \mu_j(1 - G_j)$ constitutes a cardinal measure of this type's opportunity set (G_j denotes the Gini coefficient and μ_j the average outcome within that social type). Thus, in the context of inequality of opportunity, one may rank types (not individuals) according to their respective values of the Sen evaluation. For any pair of types, denoted i and j, and starting from the one with the smallest value of the Sen evaluation function, the Gini-Opportunity index across types i to k is defined as: $G - \text{Opp} = \frac{1}{\mu} \sum_i^k \sum_{i<j} \left[p_i p_j (A_j - A_i) \right]$. This index gives the weighted average of the differences between the types' opportunity sets in which the weights (p) are the sample weights of the different types. The value of all these indices is highly sensitive to the number of types; this can be a problem because, as seen before, the number of types is, in practice, defined subjectively by the researcher.

It should finally be noted that a good measure of inequality of opportunity in health should be able to bring together multiple circumstances and, given that health is an inherently multidimensional concept, multiple dimensions of health outcomes. This also applies to the case of inequality of opportunity in healthcare, which incorporates a number of different dimensions, such as general practitioner visits and specialist visits. Rosa Dias and Yalonetzky (2013) have recently addressed this issue by drawing on the segregation literature and proposing new measures that are applicable when health (or healthcare) is proxied by a finite number of ordinal indicators.

Inequality of Opportunity in Health Economics: Theoretical Contributions and Empirical Evidence

Theoretical Contributions in Health Economics

It is possible to argue that inequality of opportunity is already the implicit equity concept in some earlier contributions in health economics, such as Alan Williams' fair innings argument and the Rawlsian approach to the measurement of health inequalities proposed in Bommier and Stecklov (2002). Yet, although the volume of applied research on

inequality of opportunity in health has grown rapidly over the past few years, the amount of theoretical work, has been comparatively smaller.

Fleurbaey and Schokkaert (2009) make an important contribution toward incorporating the analysis of inequality of opportunity in health in the broader framework of responsibility-sensitive egalitarianism. They propose analyzing inequality of opportunity in health within the framework of a complex structural model that encompasses simultaneously the demand for health, lifestyle and healthcare, labor supply, and income distribution. In this model, the health stock depends on a range of factors, encompassing the consumption of healthcare and other goods, job characteristics, socioeconomic background, genes, and unanticipated health shocks. Labor income is endogenous and depends on various factors including individual ability. The demand for healthcare also depends on multiple factors, including supply-side variables and individual demand for supplementary health insurance.

This model can be solved in two stages. First, individuals decide on their desired level of supplementary health insurance. Second, for that level of insurance coverage, they maximize utility subject to income constraints, time constraints, and to the supply of healthcare constraints. This allows for the joint determination of the demand for health care, consumption goods, and individual labor supply. Finally, armed with the optimal values for these, the optimal levels of health, income, and utility are endogenously determined by the model.

This complex structural model is the most encompassing framework proposed for the analysis of unfair health inequalities (including inequality of opportunity). However, the multiple and reciprocal causal relationships that it embodies poses serious operational challenges to the empirical identification of the model.

Another aspect that has received attention in the health economics literature relates to the fact that, in practice, it is often not possible to observe the full set of relevant circumstances influencing health outcomes. Fleurbaey has shown that this issue, known as the partial-circumstance problem, may bias the measurement of inequality of opportunity in health. At present, there has not yet been found a reliable way to derive theoretical bounds for this bias. Rosa Dias (2010) examines the practical relevance of this matter by proposing a simple behavioral model of inequality of opportunity in health that integrates Roemer's framework of inequality of opportunity with the Grossman model of health capital and demand for health. The model generates a recursive system of equations for the health stock and each of a series of effort factors such as the weekly consumption of calorific food, alcohol, and the weekly frequency of physical exercise. To take into account the role of unobserved heterogeneity, the system is then jointly estimated by full information maximum likelihood with freely correlated errors. The results suggest that, when unobserved heterogeneity in the set of circumstances is taken into account, the estimates of the recursive relationship between circumstances, effort, and health outcomes change considerably, thereby corroborating the empirical relevance of the partial-circumstance problem. García-Gomez *et al.* (2012) use an analogous estimation strategy to implement the framework of Fleurbaey and Schokkaert (2009), thereby modeling the channels through which circumstances affect health outcomes in adulthood. Armed with this behavioral model, García-Gomez *et al.* (2012) showed that distinguishing between these different channels is useful not only as a means of avoiding the partial-circumstances problem, but also in order to perform a sensitivity analysis of the results with respect to different normative positions regarding the factors that should be considered, i.e., circumstances and effort.

A different, although related, issue concerns the correct way to treat the partial correlations between circumstances and effort. Jusot *et al.* (2013) examine the practical relevance of this issue for the measurement of inequality of opportunity in health by applying a reduced-form approach to data from a large French survey. Interestingly, their results suggest that adopting fundamentally different normative approaches to this matter makes little difference, in practice, for the measurement of health inequalities.

Empirical Evidence

In recent years the number of applications of the inequality-of-opportunity framework to health has grown rapidly. Rosa Dias (2009) and Trannoy *et al.* (2010) examine the existence and magnitude of inequality of opportunity in health using, respectively, data from the UK and France. Employing the stochastic dominance testable conditions proposed by Lefranc *et al.* (2009), they find that, in both countries, there is clear inequality of opportunity in self-reported health between individuals of different parental background (defined according to the father or male head of household's occupation). Furthermore, these empirical applications show that shifting the focus from inequality in health to inequality of opportunity changes the results significantly: For example, in the case of the UK, Rosa Dias (2009) shows that an unusually rich set of circumstances that include parental background, childhood health, ability, and social development account for just approximately one-fourth of the total inequality in health.

These articles also show that inequality of opportunity in health is substantial in the countries studied: Trannoy *et al.* (2010) show that a hypothetical complete nullification of the influence of observed circumstances on health would, in the case of France, leads to a 57% points reduction in the self-reported Gini coefficient. Jusot *et al.* (2010) pursue this line of research further by using data from the 2004 Survey on Health Ageing and Retirement in Europe to compare the extent of inequality of opportunity in health across 10 European countries. Their results suggest that the magnitude of this type of inequality is markedly different between blocks of countries: Inequality of opportunity in self-assessed health is systematically higher in Southern Europe than in Northern European countries. In addition, this article makes clear that there are also differences regarding the most important circumstances in each of the countries.

Another important aspect concerns the evolution of inequality of opportunity in health over the lifecycle: Do circumstances affect health outcomes more heavily in the early years of life, young adulthood, or in old age? Rosa Dias (2009) provides some empirical evidence on this issue, using data from a UK cohort study; results from this study show that the influence of circumstances on self-reported health at 23, 33,

42, and 46 years of age is remarkably constant. This issue has been reexamined in greater depth by Bricard *et al.* (2012). This article proposes two alternative strategies for quantifying inequality of opportunity in health over the lifecycle. From an *ex ante* perspective, an aggregate measure of the lifetime health stock is estimated for each individual; inequality of opportunity in this aggregate health is then measured between individuals or groups. Alternatively, from an *ex post* perspective, health inequalities are measured across individuals at each stage of their lifecycle, before aggregating inequalities over the lifetime. Bricard *et al.* (2012) show that these two perspectives are grounded on different normative principles, and that they lead to different results when applied to real data.

Finally, an area that is, at present, receiving growing attention is the application of the inequality-of-opportunity framework to the normative evaluation of concrete policy interventions. Figheroa *et al.* (2012) propose a methodology to evaluate social projects from an equality-of-opportunity perspective by looking at their effect on the distribution of outcomes conditional on observable covariates. They apply this approach to the evaluation of the short-term effects of Mexico's well-known Oportunidades program on children's health outcomes. Jones *et al.* (2012) also proposes a normative framework, but designed for the evaluation of complementary policy interventions such as the health effects of educational interventions. This article grounds this proposal on Roemer's (2002) model of inequality of opportunity, and applies it to data from a large-scale UK educational reform. Although Figheroa *et al.* (2012) focus on the evaluation of short-run policy effects, Jones *et al.* (2012) center on their long-run impact on health and lifestyle.

Although considerable evidence on inequality of opportunity in health has been amassed, there are still important unanswered questions in this field. First, virtually all the available evidence relates to developed countries. It would be interesting to know more about the magnitude, causes, and the channels of influence of inequality of opportunity also in developing countries. Second, further research is needed on the impact of health policy on inequality of opportunity in health. Although over the past years much has been learnt about the size and evolution of this type of inequality, little is still known about the ways to tackle it effectively.

See also: Education and Health. Efficiency and Equity in Health: Philosophical Considerations. Fetal Origins of Lifetime Health. Impact of Income Inequality on Health. Intergenerational Effects on Health – *In Utero* and Early Life. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach. Measuring Vertical Inequity in the Delivery of Healthcare. Unfair Health Inequality. Welfarism and Extra-Welfarism

References

Bommier, A. and Stecklov, G. (2002). Defining health inequality: Why Rawls succeeds where social welfare theory fails. *Journal of Health Economics* **21**, 497–513.

Bourguignon, F., Ferreira, F. and Menéndez, M. (2007). Inequality of opportunity in Brazil. *Review of Income and Wealth* **53**(4), 585–618.

Bricard, D., Jusot, F., Tubeuf, S. and Trannoy, A. (2012). Inequality of opportunities in health over the life-cycle: An application to ordered response health variables. *Health Economics* **21**, 129–150.

Figheroa, J. L., Van de Gaer, D. and Vandenbossche, J. (2012). Children's health opportunities and project evaluation: Mexico's Oportunidades program. *CORE Discussion Papers 2012015*. Belgium: Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics* **28**(1), 73–90.

Fleurbaey, M. and Schokkaert, E. (2012). Equity in health and health care. In Pauly, M., McGuire, T. and Barros, P. P. (eds.) *Handbook of Health Economics*, vol. 2, pp 1004–1092. North-Holland: Elsevier.

García-Gomez, P., Schokkaert, E., Van Ourti, T. and Bago D'Uva, T. (2012). Inequality in the face of death. CORE *Working Paper 2012/24*. (in press).

Jones, A. M., Roemer, J. E. and Rosa Dias, P. (2012). Equalising opportunity in health through educational policy. *Health, Econometrics and Data Group (HEDG) Working Paper*. Chicago, USA: The University of Chicago Press.

Jusot, F., Tubeuf, S. and Trannoy, A. (2010). Inequality of opportunities in health in Europe: Why so much difference across countries? *Health, Econometrics and Data Group (HEDG) Working Paper 10/26*. (in press).

Jusot, F., Tubeuf, S. and Trannoy, A. (2013). Circumstances and effort: How important is their correlation for the measurement of inequality of opportunity in health? *Health Economics*. doi: 10.1002/hec.2896.

Lefranc, A., Pistolesi, N. and Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics* **93**(11–12), 1189–1207.

Roemer, J. E. (1998). *Equality of opportunity*. Cambridge, MA: Harvard University Press.

Roemer, J. E. (2002). Equality of opportunity: A progress report. *Social Choice and Welfare* **19**, 455–471.

Rosa Dias, P. (2009). Inequality of opportunity in health: Evidence from a UK cohort study. *Health Economics* **18**(9), 1057–1074.

Rosa Dias, P. (2010). Modelling opportunity in health under partial observability of circumstances. *Health Economics* **19**(3), 252–264.

Rosa Dias, P. and Yalonetzky, G. (2013). *Measuring Inequality of Opportunity in Health When the Health Variable is Discrete and Multidimensional*. Oxford: Oxford University Press.

Trannoy, A., Tubeuf, S., Jusot, F. and Devaux (2010). Inequality of opportunities in health in France: A first pass. *Health Economics* **19**, 921–938.

Van de Gaer, D., Vandenbossche, J. and Figueroa, J. L. (2012). Children's health opportunities and project evaluation: Mexico's Oportunidades program. *World Bank Economic Review* (forthcoming).

World Bank (2005). World Development Report. *Equity and Development*. Washington, DC: The World Bank.

Further Reading

Arneson, R. (1989). Equality and equal opportunity for welfare. *Philosophical Studies* **56**, 77–93.

Bossert, W. (1995). Redistribution mechanisms based on individual characteristics. *Mathematical Social Sciences* **29**, 1–17.

Checchi, D. and Peragine, V. (2010). Inequality of opportunity in Italy. *Journal of Economic Inequality* **8**, 429–450.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics* **99**, 906–944.

Dworkin, R. (1981). What is equality? Part 2: Equality of resources. *Philosophy & Public Affairs* **10**, 283–345.

Ferreira, F. and Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth* **57**(4), 622–657.

Fleurbaey, M. (2008). *Fairness, responsibility and welfare*. Oxford: Oxford University Press.

Fleurbaey, M. and Peragine, V. (2013). Ex-ante versus ex-post equality of opportunity. *Economica* **80**(317), 118–130.

Foster, J. and Shneyerov, A. (2000). Path independent inequality measures. *Journal of Economic Theory* **91**(2), 199–222.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Sen, A. (1980). Equality of what? In McMurrin, S. (ed.) *The tanner lectures on human values* 1. Salt Lake City: University of Utah Press.

Ethics and Social Value Judgments in Public Health

NY Ng, Yale School of Public Health, New Haven, CT, USA
JP Ruger, Yale Schools of Medicine, Public Health, and Law, New Haven, CT, USA

Glossary

Aggregation A process of adding up smaller parts to make a greater whole. In health policy the issue arises of how to weight the health experience of different individuals in arriving at a statement about the health of a population.

Autonomy The general ethical principle in medicine of respecting an individual's freedom from external interference and their right to self-determination.

Communitarianism The doctrine that individuals' welfare cannot be properly understood or measured without regard to their membership of a community and the roles they play in it.

Consequentialism The doctrine that the moral worth of an action, policy, etc. is to be judged in terms of its consequences.

Externality An externality is a consequence of an action by one individual or group for others. There may be external costs and external benefits. Some are pecuniary, affecting only the value of other resources (as when a new innovation makes a previously valuable resource obsolete); some are technological, physically affecting other people (communicable disease is a classic example of this type of negative externality); some are utility effects that impinge on the subjective values of others (as when, for example, one person feels distress at the sickness of another, or relief at their recovery).

Informed consent 'Consent' in general is usually legally grounded either on the principle that a physician has a duty of care or that a patient has a right to self-determination. In most countries the informed consent of patients to treatments is based on the idea of what information a reasonable person might expect to be told in a given situation. In the UK, however, informed consent is based upon what professionals regard as reasonable to provide and hence on what information in any given case a physician's peers would provide.

Utilitarianism The ethical doctrine, a variant of which underlies nearly all normative economics, which specifies utility (sometimes equated with 'happiness') as the principal moral good of society and the entity that humankind as a whole ought to maximize. The popular moral slogan for a society (of given population) to pursue under utilitarianism is 'the greatest happiness of the greatest number.'

Introduction

Public health, unlike medicine, is not about doctors treating individual patients. Public health is about population health. It is a collective social effort to promote health and prevent diseases – both communicable and noncommunicable – and disability that involves population surveillance, regulation of determinants of health (such as food safety and sanitation), and the provision of key health services with an emphasis on prevention. Because private actors lack sufficient incentive and ability to undertake population-wide measures, public health is a vital resource for which government is the crucial provider, enabled by its police powers and its ability to regulate, tax, and spend. The exercise of government powers for the health of its population raises ethical issues, such as public welfare, individual autonomy and freedom, privacy and confidentiality, just distribution of benefits and burdens, transparency, and public accountability. These ethical concerns sometimes conflict, pitting values against one another. How they should be balanced will vary on a case-by-case basis. This article discusses justifications for government action in public health, the tension between individual freedom and public health, issues of distributive justice in public health, and ethical guidelines for public health policymaking.

Justifications for Government Intervention

Given that the government is best placed to undertake the work of public health, what are justifications for public health policies?

Ethical Justifications

Public health has utilitarian and consequentialist aspects. In a utilitarian sense, its goal is to maximize public welfare through the protection and promotion of population health. From a consequentialist point of view, public health policies are justified and judged largely by their outcomes, achieved by means of acceptable procedures. Public health measures seek to minimize harm from communicable and noncommunicable diseases, from exposure to health-endangering substances and environments (e.g., cigarette smoke and poor sanitation), and from high-risk behaviors (e.g., substance abuse and unprotected sex). Welfare is promoted through policies aimed at encouraging and facilitating behavior conducive to health (e.g., hand washing, smoking cessation, education about the dangers of drugs, and unprotected sex), and establishing more healthful environments (e.g., smoke-free public spaces, mosquito extermination, and adequate nutrients).

In the course of protecting and promoting public health, government authorities have the responsibility to ensure that public health policies themselves do no harm, or at least that their harms are outweighed by their benefits. Public health policies are not entirely utilitarian, however, in that individuals are not considered expendable for the greater good. The rights of individuals are important considerations in the formulation and implementation of public health measures, as discussed later.

The protection of vulnerable groups is another ethical motive for public health action. Vaccination and nutrition supplements, for example, protect children from disease and malnutrition, and smoking bans in bars and restaurants safeguard the health of workers who may not otherwise have the leverage to demand a smoke-free environment. Publicly funded health services can in principle help address the health needs of those who cannot afford private medical care or insurance. Such measures also may contribute to reducing health inequalities, by bringing the health of vulnerable groups more in line with the general population. Reduction of inequalities can itself be considered an ethical justification, as people with equal status (e.g., citizenship) should not suffer from those types of health inequalities that are due to morally arbitrary reasons (e.g., birth into a poor family and other bad luck).

Economic and Other Justifications

Poor health has collateral effects. On an individual basis, illness, disability, and their associated expenses can lead to absenteeism and decreased productivity that diminish income, inability to pursue education, reductions in essential consumption such as food and shelter, bankruptcy, and poverty. High infant and child mortality may lead to the compensatory decision to have more children, which decreases resources available for investment in health and education for each child. High adult mortality leaves orphans with bleak prospects. On a societal level, employers and the health system also suffer economic losses from lower worker productivity and greater healthcare burdens. Poor population health can even be economically and politically destabilizing. A particularly grim example is the Human immunodeficiency virus (HIV)/Acquired immune deficiency syndrome (AIDS) crisis in Africa, which lowered life expectancy by decades in some countries, killing adult men and women in their prime productive years. This is economically devastating for individual families and can potentially have larger implications. If deaths cause an overall decrease in economic output, the tax base funding health, education, police, and the military would also shrink, thus diminishing the perceived legitimacy of government. Lower life expectancy discourages long-term investment in education; it also means fewer and less experienced civil servants, reducing government administrative capacity. Low income and low government capacity create incentive for crime, violence, and radicalism, which in turn may trigger more state repression. Foreign investment may be deterred by lack of productive workers and instability. Weak states are also more vulnerable to armed conflicts and terrorism, increasing regional and international security risks. Public health problems can stand as obstacles to economic, political, and human development. What can be achieved with a population debilitated and dying *en masse*?

Good population health, however, can be part of a virtuous cycle of development. Higher life expectancy provides higher returns to education and human capital investment; lower infant and child mortality helps lower fertility, which results in greater health and educational resources available per child. A healthier, more educated work force is more economically productive, and more capable to generate the tax revenue for crucial infrastructure and services that would further development and attract investments. The connection between public health and development is less pronounced in developed countries that have long attained a high standard of population health; in impoverished countries, however, public health is a key component of the fight against poverty. Generally speaking, the justification for government public health action is ample; it is the justifications for specific public health measures that tend to be more contentious.

Individual Freedom versus Public Health

Public health policies are population oriented. Because individual health – for example, whether one is vaccinated, infected, a smoker – affects the health of others, public health measures regulate individual behavior in order to achieve population health goals. Such policies apply broadly and are not tailored to specific individual circumstances. They typically mandate certain behaviors (e.g., vaccination) and prohibit others (e.g., congregating with others while infected with quarantinable diseases), and sometimes take individual choice largely out of the picture (e.g., water fluoridation). All raise questions about how individual autonomy and freedom should be balanced against public health interests.

Public health ethicists often invoke the 'harm principle,' which respects individuals' sovereignty over their bodies and actions as long as their actions do not harm others. Ethicists generally agree that the greater the intrusion on individual autonomy and freedom, the greater the public health benefit must be to justify the policy. The public health situation that most starkly pits individual freedom against population health is infectious disease control. The liberty of individuals and their right to associate with others are curbed by protocols to separate infected patients from the population to prevent exposing others (isolation), and to separate or restrict the activities of people who are not diagnosed as infected but who may have been exposed to infection or who may be ill without symptoms (quarantine).

Disease control in the age of globalization has global health implications. The conflict is no longer between individual freedom and domestic population health, but between individual freedom and global population health, as demonstrated by the rapid spread of HIV, Severe acute respiratory syndrome (SARS), and pandemic flu via air travel. The economic toll of outbreaks is also potentially significant; losses from the 2003 SARS outbreak have been estimated to run in the billions. Domestic efforts are an integral part of global outbreak prevention. Given the high health and economic stakes in disease containment, the isolation of infected

individuals to prevent spread of disease is fairly uncontroversial. Quarantine, which applies to those who are not evidently ill, is a more disputed practice, sparking debates on its necessity and effectiveness: Only a small number of quarantined individuals are likely to be actually sick, although rights and freedom are infringed for all individuals placed under quarantine. A 2006 study by Day *et al*. suggests that quarantine is likely to be more useful and justifiable when isolation is ineffective, or if disease can be transmitted asymptomatically, when the consequences of exposure to others are severe, fatal, and/or irreversible, or if there is an intermediate asymptomatic period that is not too short or too long.

Isolation and quarantine can be voluntarily observed or coercively imposed. To the extent feasible, public health measures should secure the voluntary compliance or participation of affected individuals, allowing individuals the autonomy of informed consent. The public health, legal, and ethical reasons for observing isolation or quarantine – and potential consequences for violating it – should be clearly communicated to affected individuals, such that they have the relevant information to assess individual and societal benefits, costs and risks, and to make the decision to comply. Should an individual refuse to comply, authorities should have a system in place to impose isolation or quarantine to protect public health. There may be circumstances in which the urgency and gravity of a public health crisis may make a complete informed consent procedure less practicable. For example, an outbreak in progress of a virulent, highly fatal disease like Ebola may require swifter separation of the infected and the exposed from the general population.

One person's infection has clear and direct negative health impact on others, but public health policies also concern activities like smoking, obesity, and the wearing of motorcycle helmets that are arguably 'lifestyle choices,' with more indirect (or minimal) negative externalities. Smoking is an individual activity that may cause lung cancer, emphysema, and other diseases for the smoker, but there is also substantial evidence for its harm to others through secondhand smoke. Illness from smoking and secondhand smoke can result in losses from lower economic productivity, and greater burdens on the health system. How should public health authorities weigh a smoker's right to smoke versus other people's right to a smoke-free environment? Do smokers really have full autonomous choice over smoking, given that nicotine is an addictive substance? Should smokers be refused tax-funded health services for smoking-related illness? To what degree should smoking be discouraged (e.g., through sin tax) or prohibited to protect especially vulnerable groups like restaurant workers, who are exposed to secondhand smoke, and the poor, among whom smoking is more common and difficult to stop?

Different people have different answers for those questions, reflected in the large variation in smoking regulations among the 50 US states and among countries worldwide. Such variation is also seen in laws governing the wearing of seat belts and vehicle helmets, the consequences of which are confined overwhelmingly to the individual making that choice. The fewer the negative public health externalities associated with particular behaviors, the more paternalistic the government regulation of these behaviors. Policies are paternalistic when they seek to protect or benefit individuals against their expressed preferences – for example, by legally requiring people to wear motorcycle helmets when they otherwise would not.

Paternalism comes in 'hard' and 'soft' versions. Hard paternalism interferes with choices of individuals who, according to Childress *et al*., are 'competent, adequately informed, and free of controlling influences' and is therefore hard to justify. Soft paternalism, however, deals with behaviors of individuals who are considered not competent, not adequately informed, or not free from external control to make that choice. For example, smokers may decide to smoke because they were insufficiently aware of the health consequences, and they may continue to smoke because they have become addicted to nicotine. Obesity may be exacerbated by food marketing and the pricing and availability of healthy versus unhealthy foods, among other factors. Such situations provide more valid grounds for government intervention, which may take the form of education, incentives (e.g., taxes or subsidies to influence price and therefore consumption), marketing restrictions, and even outright bans, if the benefits of strong regulation are deemed to outweigh the infringement of individual freedom. A 'libertarian' version of paternalism has been proposed by Thaler and Sunstein that would structure the choice environment such that people could more easily choose to act in their own best interest (e.g., placing healthy foods at eye level in the store), as a way to preserve greater individual freedom.

The privacy and confidentiality of individuals are also important factors to consider in public health policymaking. Certain conditions and diagnoses – such as HIV/AIDS or mental illness – may carry social stigma, or impede one's ability to gain employment or acquire health insurance if publicized. The right to privacy and confidentiality must be balanced against the need to collect and disseminate information to achieve valid public health goals, such as infectious disease contact tracing, providing patients with treatment, and screening to prevent transmission of diseases through blood or organ donation, or from mother to child.

Distributive Justice in Public Health

In the context of limited resources – which is always and everywhere – the question is how should resources be allocated? The distribution of benefits and burdens is another ethical consideration in public health policy. Resource allocation and policy application should be fair. Extermination of mosquitoes, for example, should not be implemented in some communities while excluding others; minority groups – such as homosexuals – should not be singled out for disease screening. Targeting programs and interventions could be justified if supported by empirical evidence, but the costs of targeting should be weighed against the benefits. Targeted intervention may be a more efficient way to reach particularly affected groups and may help reduce health inequalities, but it may also come with negative effects. Stigma may become attached to groups singled out for disease programs, and the health of the nontargeted groups and individuals may be compromised if they do not receive the relevant health

education and do not receive screening because they are not considered at sufficient risk. Where possible, a universal, voluntary screening policy should be implemented.

The use of sin taxes to discourage consumption of unhealthful products like cigarettes is another instance of a targeted public health policy. The sin tax affects smokers, and redistributes that revenue to the rest of the population. This unequal burden aims to discourage cigarette consumption, which benefits the health of smokers and those subject to their secondhand smoke. However, cigarette taxes may also disproportionately affect lower income and minority individuals, who are more likely to be smokers (at least in the US), which makes the tax regressive in practice. Just how regressive may depend on how the revenues would be spent (e.g., funding other tobacco control efforts? or folded into general revenues?). Again, public health authorities must balance the benefits against the costs.

The distribution of benefits and the allocation of scarce resources are important issues in designing publicly funded healthcare packages. What kind of services should state-funded healthcare packages include? How much emphasis should prevention receive relative to treatment? Should resources go toward improving average health, which can be done without special attention to people with special health needs, or should resources be devoted to reducing health inequalities, which implies greater resources to the least healthy to bring them closer to the general population? What should be done about people who have exorbitantly expensive health conditions with little prospect of big improvement?

The consequentialist orientation of public health and limits in resources make the balancing of costs and benefits a major concern in public health policymaking. Costs are weighed against benefits using methods such as cost-benefit, cost-effectiveness, and cost-utility analyses. Cost-benefit analysis translates all benefits into monetary units that account for direct (e.g., medical) and indirect (e.g., productivity) effects; cost-effectiveness analysis shows the cost of each unit of gain in health, as indicated by measures such as years of life gained or deaths averted. Cost-utility analysis presents costs associated with a subjective measurement unit that combines preferences for length of life with preferences for quality of life. These kinds of analyses are used in the hopes of maximizing health benefits while minimizing cost. The National Institute for Health and Clinical Excellence in the UK, for example, draws on cost-effectiveness analyses to help direct coverage of medicines and treatments under the National Health Service.

The use of such welfare economic assessments in public health policymaking is not without controversy. For instance, the US, despite extremely high healthcare costs, has so far rejected using such measures in health policy. Although welfare economic methods offer a way to maximize health value for money in an evidence-based fashion, they have other implications that can be politically and morally difficult to accept. These methods account only for aggregate welfare, without considering the distribution of benefits and burdens. They tolerate significant health inequalities. Inequalities may even be exacerbated for the disabled, old, and very sick, the health benefits for whom cost-utility analysis assigns less weight due to their reduced capacity to benefit from health resources. This goes against people's intuition, found in research, to prioritize resources for the sicker and the more disabled even though they are less able to benefit.

Aggregation problems can result when weighing a small benefit for many against a large – perhaps vital – benefit for a few, yielding counterintuitive assignments of priority to minor procedures such as tooth-capping ahead of a life-saving surgery for ectopic pregnancy, which Hadorn reported from the Oregon Medicaid experiment in which policymakers attempted to determine a Medicaid (state-funded healthcare for the poor) health package using cost-utility analysis. Welfare economic methods also treat all health conditions as directly comparable, but blindness and loss of limb, for instance, are arguably not comparable to cardiovascular disease or high blood pressure, which further suggests that those methods alone may not be sufficient to direct resource allocation. Efforts to include weights (e.g. age or distribution) and other modifications have not satisfactorily solved these problems.

Resource allocation issues go beyond healthcare. Because poverty and social class are strong predictors of health, some ethicists also argue that public health has a role in poverty reduction and improvement of social conditions – such as housing, education, sanitation, and female empowerment – in order to address the structural causes of ill health and to increase people's ability to protect health for themselves and others (e.g., more educated and empowered women are better able to secure nutrition for and prevent diseases in their children).

Public health-related distributive justice can take on a global dimension. Poor countries often have more acute resource allocation problems in that they have little resources to begin with, and what resources they have they must devote significant portions to servicing foreign debts. Because poor countries must often reduce social spending in health and sectors with impact on health in order to pay debts or to comply with loan conditions, wealthy creditor countries and international financial institutions such as the World Bank and the International Monetary Fund have been urged on moral grounds to forgive loans and reverse structural adjustment policies that hinder vital public spending, in addition to providing more assistance.

Conclusion

Broad questions of how resources should be allocated involve conceptions of what justice and equity entail, and what obligations a state has in ensuring the health of its populations – whether it should aim for a basic minimum standard or something higher, within the constraints set by resource availability and the needs of legitimate state duties besides health. On a global level, there are additional questions about the existence and extent of duties to redistribute resources between rich and poor countries. Different moral perspectives (e.g., humanitarianism, human rights, communitarianism, and realism) will have different answers for those questions.

For specific public health measures, conflicts in ethical concerns will vary on a case-by-case basis, but scholars have presented guidelines to help assess ethicality. One example of such guidelines is the 5 'justificatory conditions' formulated

by 10 ethicists in 2002. The satisfaction of these conditions would justify the pursuit of a given public health measure over competing ethical values. These five conditions are effectiveness, proportionality, necessity, least infringement, and public justification. The effectiveness condition requires the public health measure to have a good chance of protecting public health; proportionality demands that the probable health benefits exceed adverse effects. The necessity condition directs policymakers to show 'good faith belief' and plausible reasons for using their proposed approach over a less coercive alternative, that is, to show that a given degree of coercion is indeed necessary. Out of all effective, proportional, and necessary options, the option that least infringes other ethical values should be chosen. And policymakers should publicly offer justification for their public health measure as well as explanation and justification for infringement, in a transparent process that truthfully and fully discloses the risks, scientific uncertainty, and moral values to relevant parties and those who will be affected by the policy, whose input should also be solicited.

These five criteria are representative of basic elements of public health ethical guidelines, which also tend to advocate respect for individual privacy and confidentiality. A transparent, participatory public process to justify policy proposals and to deliberate the weighing of benefits, costs, and risks is appropriate for developing and evaluating both narrower public health interventions and more general public resource allocation. Allowing people to take part in the public health policymaking process can build and maintain trust in public health authorities; it also strengthens agency and autonomy, and gives fuller meaning to informed consent.

See also: Addiction. Advertising as a Determinant of Health in the USA. Alcohol. Cost–Value Analysis. Education and Health in Developing Economies. Fertility and Population in Developing Countries. HIV/AIDS, Macroeconomic Effect of. HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Illegal Drug Use, Health Effects of. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Infectious Disease Externalities. Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity. Macroeconomic Effect of Infectious Disease Outbreaks. Noncommunicable Disease: The Case of Mental Health, Macroeconomic Effect of. Nutrition, Economics of. Nutrition, Health, and Economic Performance. Priority Setting in Public Health. Public Health in Resource Poor Settings. Public Health: Overview. Quality-Adjusted Life-Years. Sex Work and Risky Sex in Developing Countries. Smoking, Economics of. Unfair Health Inequality. Water Supply and Sanitation. Welfarism and Extra-Welfarism

Further Reading

Anand, S., Peter, F. and Sen, A. (eds.) (2006). *Public health, ethics, and equity*. New York: Oxford University Press.

Bayer, R., Gostin, L. O., Jennings, B. and Steinbock, B. (eds.) (2007). *Public health ethics: Theory, policy, and practice*. New York: Oxford University Press.

Callahan, D. and Jennings, B. (2002). Ethics and public health: Forging a strong relationship. *American Journal of Public Health* **92**(2), 169–176.

Childress, J. F., Faden, R. R., Gaare, R. D., et al. (2002). Public health ethics: Mapping the terrain. *Journal of Law, Medicine and Ethics* **30**(2), 169–177.

Day, T., Park, A., Madras, N., Gumel, A. and Wu, J. (2006). When is quarantine a useful control strategy for emerging infectious disease? *American Journal of Epidemiology* **163**(5), 479–485.

Hadorn, D. C. (1991). Setting health care priorities in Oregon. *Journal of the American Medical Association* **265**, 2218–2225.

ten Have, M., de Beaufort, I. D., Mackenbach, J. P. and van der Heide, A. (2010). An overview of ethical frameworks in public health: Can they be supportive in the evaluation of programs to prevent overweight? *BMC Public Health* **10**, 638–648.

Kass, N. E. (2001). An ethics framework for public health. *American Journal of Public Health* **91**(11), 1776–1782.

Nord, E., Pinto, J., Richardson, J., Menzel, P. and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics* **8**(1), 25–39.

Ruger, J. P. (2009a). Global health justice. *Public Health Ethics* **2**(3), 261–275.

Ruger, J. P. (2009b). *Health and social justice*. Oxford: Clarendon Press.

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge*. New Haven, CT: Yale University Press.

World Health Organization (WHO) (2001). *Macroeconomics and health: Investing in health for economic development*. Geneva: WHO. Report of the Commission on Macroeconomics and Health.

Relevant Websites

http://www.academia.edu/177131/Public_Policies_Law_and_Bioethics_A_Framework_for_Producing_Public_Health_Policy_Across_the_European_Union
Academica.edu.

http://www.apha.org/NR/rdonlyres/1CED3CEA-287E-4185-9CBD-BD405FC60856/0/ethicsbrochure.pdf
American Public Health Association.

http://www.nuffieldbioethics.org/public-health
Nuffield Council on Bioethics.

Evaluating Efficiency of a Health Care System in the Developed World

B Hollingsworth, Lancaster University, Lancaster, UK

Introduction

The way economists look at the production of health care is to examine the relationship between the inputs into and the outputs from a production process as illustrated in Figure 1.

Figure 1 is a flow diagram showing how inputs such as medical staff and equipment produce health care, for example, the services offered by a hospital, and how use of these type of available health care inputs are converted into actual health itself, for example, curing a disease. Health itself, of course, is influenced by matters other than the health care system – such as housing conditions, education levels etc., which are often also accounted for in such models of how health is produced. Economists are interested in how one can make these production flows as resource efficient as possible because health care is very expensive, on average using up over 10% of developed countries GDP. To do this, the most efficient use of the inputs to these processes to produce the desired output is looked at – in most cases, to maximize health. In the top half of Figure 1, one can see how health care is produced given certain inputs, such as medical staff time. In the bottom half of Figure 1, health care becomes an input to a person's health, along with all the other things outside the health care system that contribute to health itself.

Mostly, research in this area has concentrated on the top half of Figure 1, as the inputs to, and the outputs from a health care organization can be measured, for example, a hospital. So, what sort of things would be inputs and outputs to the production of health care? It can be thought about in terms of a hospital, the most recognizable unit of health care production in a developed country, and the largest consumer of resources. Inputs include things like doctors and nurses, equipment and drugs, and capital, such as buildings and beds. Outputs are produced by the hospital – so, for example, numbers of patients treated – ideally adjusted in some way for the quality of care they produce – numbers of different operations undertaken, or diagnostic tests.

This article will describe how the relationship between inputs and outputs can be measured, and how information that improves the efficiency of how these services are delivered can be provided – the benefit being an improvement in the efficiency of production of service delivery and ultimately the production of patient health. It begins with a discussion of alternative techniques for measuring efficiency. Theoretical foundations are based on the pioneering work of Farrell (1957).

Two alternative approaches to measuring efficiency in the health care sector are described: data envelopment analysis (DEA), and stochastic frontier analysis (SFA). The article then describes how best to make use of techniques such as these in terms of a system of protocols and gives guidelines for how to provide the most appropriate information to those involved in policy making and service delivery.

Efficiency Measurement

In economic terms technically efficient combinations of inputs are those which use the least resources to produce a given level of output (for a given state of technology). Alternatively, technical efficiency (TE) may be defined in terms of maximizing output for a given level of inputs. By contrast, full

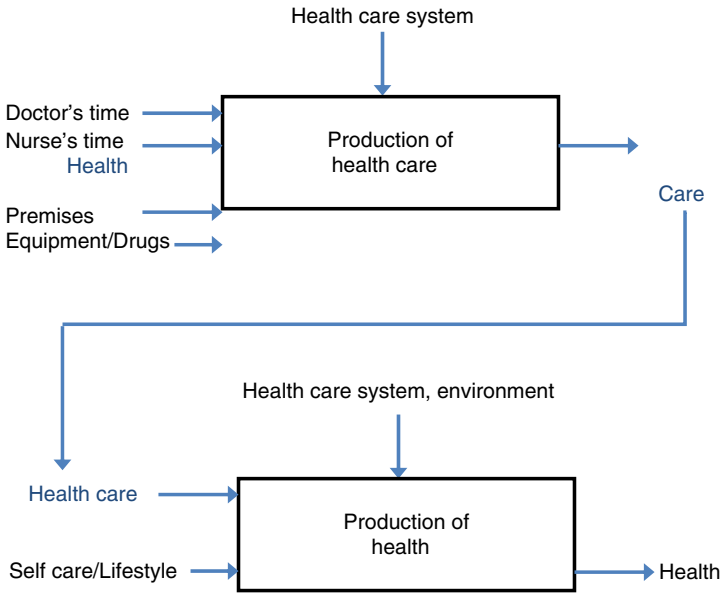


Figure 1 The production of health care and health.

 doi:10.1016/B978-0-12-375678-7.00204-2

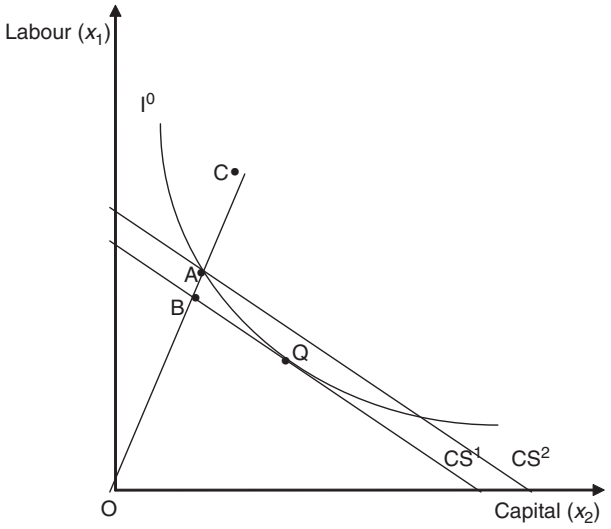


Figure 2 Radial efficiency measurement.

allocative efficiency is achieved by selecting combinations of inputs (e.g., mixes of labor and capital) which produce a given amount of output at minimum cost (given market prices for inputs), i.e., there can be no improvement in output by simply reallocating resources. The first measure looks at physical quantities, the second introduces a cost element.

Farrell's seminal work introduced two further concepts: radial measures of efficiency; and overall (economic) efficiency. These concepts are illustrated in Figure 2. The figure considers a simple example of producing a single aggregated output 'health care' from two inputs: medical labor (x_1) and capital (an example often used is beds)(x_2), The parallel lines (CS^1 and CS^2) represent isocost lines (which show relative input combinations that cost the same) and I°, an isoquant (simply a line drawn through a combination of input points used to produce the same level of output). Assuming a hospital chooses a desired health care output level y°, to be technically efficient they should choose a combination of inputs which lie on I°. Producing quantity y° using inputs at point C would be technically inefficient because the hospital could produce y° using both less labor and beds. Keeping the same mix of inputs, a hospital would be technically efficient if they are produced at point A, which lies on the isoquant. Farrell's measure of TE is based on the line OC, which passes through A and C. OC is often referred to as a radial measure of efficiency as it measures efficiency in terms of distance from the origin. TE at point C is given by the radial measure:

$$\mathrm{TE} = \mathrm{OA}/\mathrm{OC} \tag{1}$$

where TE must take a value greater than zero and less than or equal to one ($0 < \mathrm{TE} \leq 1$). If TE = 1, the hospital is technically efficient and is operating on the isoquant. If TE < 1 the hospital is technically inefficient.

If a hospital wishes to minimize costs, they will choose the combination of labor and beds at point Q where the isocost line CS^1 is tangential to I°, and where the combinations of inputs cost the least to produce the given level of output. If the hospital chooses an input mix (e.g., they may be legally obliged to have a certain number of doctors employed to offer certain services) along the line OC and is technically efficient they will produce at point A, which, lies on the isocost line CS^2. However, this implies they are not minimizing costs. The allocative inefficiency of choosing the input mix at point A (which is technically efficient) can be captured as the ratio of the costs of producing at A compared to the costs of producing at the allocative efficiency level, point Q, where the latter costs are given by the icocost line CS1 (the ray OA intersects this isocost line at B). This is the ratio:

$$\mathrm{AE} = \mathrm{OB}/\mathrm{OA} \tag{2}$$

where similarly AE must take a value greater than zero and less than or equal to one ($0 < \mathrm{AE} \leq 1$). When AE is less than 1 this implies that production is not allocatively efficient. AE can be interpreted as a measure of excess costs arising from using inputs in inappropriate proportions. If producing at 0Q the hospital would be technically and allocatively efficient, otherwise, if, for instance, a particular input mix is imposed on the hospital, it can achieve TE but not necessarily allocative efficiency.

Farrell's TE and AE terms can be combined to generate a measure of overall (economic) efficiency (OE) for production at point C:

$$\mathrm{OE} = \mathrm{TE} \times \mathrm{AE} = (\mathrm{OA}/\mathrm{OC}) \times (\mathrm{OB}/\mathrm{OA}) = \mathrm{OB}/\mathrm{OC} \tag{3}$$

where OE also lies in the range ($0 < \mathrm{OE} \leq 1$).

Empirical measurement of these concepts can now be considered.

Data Envelopment Analysis

DEA is by far the most common method for analyzing efficiency in health care. It has now been applied over 400 times in health care settings. DEA is a mathematical technique which makes use of linear programming methods. It is based on the idea of efficiency as the relationship between the outputs from an activity and the amount of inputs that the activity uses. In the simple case of a single output/single input firm a measure of TE can be defined as:

$$TE = \frac{y}{x} \tag{4}$$

where y = output and x = input.

The greater this ratio, the greater the quantity of output for a certain amount of input, as measured in natural (noncost) units. For a multiple-output/multiple-input firm, like a hospital which treats different types of cases using staff of different types, various equipment and so on, an overall measure of a hospital's TE is:

$$TE = \frac{\sum_r y_r}{\sum_i x_i} \tag{5}$$

where i is input, and r is output.

The problem with this is that inputs and outputs cannot be simply summed as they usually measure very different things, for example, numbers of doctors, and numbers of operating theaters). Rather, weights to each of the inputs and outputs are

given so that:

$$TE = 0 < \frac{\sum_{r=1}^{p} u_r \cdot y_r}{\sum_{i=1}^{m} v_i \cdot x_i} \leq 1 \qquad (6)$$

where: y_r = quantity of output r; u_r = weight attached to output r; x_i = quantity of input i; v_i = weight attached to input i; and p and m are the numbers of outputs and inputs. As is explained below, the weights are chosen so that $0 < TE \leq 1$. Thus, DEA is founded on an indicator of efficiency which can be calculated for each firm and, if u and v are fully flexible, is defined as the ratio of a weighted sum of the outputs relative to a weighted sum of its inputs.

The efficiency of any firm or unit, say a hospital (or nursing home, GP practice etc.), can be measured relative to other units within a peer group. Because the weights are unknown a priori, they must be calculated. Of all of the possible sets of weights which would satisfy all of the constraints, the linear program optimizes the ones that give the most favorable view of the unit. This is the highest efficiency score, the one that shows the hospital in the best possible light. This problem can be expressed as a fractional program. Such programs are difficult to solve, but can be reformulated into a straightforward linear program (LP) by constraining the numerator or denominator of the efficiency ratio to be equal to unity. This recognizes that in maximizing a ratio it is the relative values of the numerator and denominator that are important, not their absolute values. The problem then becomes to either maximize weighted output with weighted input equal to unity, or minimize weighted input with weighted output equal to unity. The output-maximizing LP is:

For h_0 in a sample of n hospitals,

$$\text{maximize} \quad h_0 = \sum_{r=1}^{p} u_r \cdot y_{rj0}$$

subject to:

$$\sum_{i=1}^{m} v_i \cdot x_{ij0} = 1 \qquad (7)$$

$$\sum_{r=1}^{p} u_r \cdot y_{rj} - \sum_{i=1}^{m} v_i \cdot x_{ij} \leq 0 \quad j = 1,\ldots,n$$

$$u_r \geq \varepsilon,\ r = 1,\ldots,p$$
$$v_i \geq \varepsilon,\ i = 1,\ldots,m$$

where: h_0 is the measure of relative TE of hospital 0, j is the reference set of $1 \ldots n$ hospitals, and ε is an infinitesimal.

In eqn [7], the denominator (weighted inputs) has been set equal to unity and the numerator (weighted outputs) is being maximized. One model must be solved for each hospital in the sample in turn, and can be solved using standard LP methods to give an efficiency score for each hospital.

The minimization rather than the maximization of this LP is simpler to solve and has a useful interpretation. If one now calls h_0 Z_0 to represent the opposite (or dual) measurement one is taking in a sample of n hospitals,

$$\text{minimize} \quad Z_0 - \varepsilon \sum_{r=1}^{p} S_r - \varepsilon \sum_{i=1}^{m} S_i$$

subject to

$$x_{ij0} Z_0 - s_i = \sum_{j=1}^{n} x_{ij} \cdot \lambda_j \quad i = 1,\ldots,m$$
$$\sum_{j=1}^{n} y_{rj} \cdot \lambda_j - s_r = y_{rj0} \quad r = 1,\ldots,p \qquad (8)$$

where: λ_j, s_r, $s_i \geq 0$ $\forall j$, i and r; λ_j are weights on units, sought to form a composite hospital to outperform j_o; s_i are the input slacks; and, s_r are the output slacks.

Essentially, the dual finds a set of weights for each hospital which minimizes an inefficiency measure subject to constraints. The hospital will be efficient if $s_i = s_r = 0$ and $Z_o = 1$, that is, a composite hospital cannot be constructed which outperforms it. This is the best that can be achieved in production terms using the combinations the hospital has available to it. If $Z_o < 1$ and/or $s_i > 0$, $s_r > 0$, the hospital will be inefficient. The composite hospital provides targets for the inefficient hospital and Z_o represents the maximum inputs a hospital should be using to attain at least its current output. The weighted combination of inputs over outputs for each hospital forms the production frontier. The hospitals which lie on this frontier, that is those which have a TE score of one using the weights of a reference unit, are called the 'peers' of the reference hospital.

DEA uses the assumption of either constant or variable returns to scale (CRS or VRS). The LP in eqn [7] or eqn [8] calculates the CRS production frontier. A VRS frontier is obtained by adding a further constraint to the dual of the LP:

$$\sum_{j=1}^{n} \lambda_j = 1 \qquad (9)$$

The extra constraint requires more units. Because the production function is not directly observable, DEA estimates a production frontier based on input and output data. The frontier maps the least resource use input combinations and is assumed to be convex to the origin. The DEA frontier is illustrated in Figure 3 and (like Figure 2) considers a simple, single output, two input example. The dots represent different producers and the quantities of inputs they use to produce the same given level of output. The DEA frontier ($1^{\circ}I^{0}$) consists of straight lines joining the points that represent the most efficient producers. Inefficient producers lie to the right of the frontier. The complete production frontier covering all levels of output can be inferred, and the analysis can be extended to cover both multiple inputs and outputs, and the assumption of CRS can be dropped.

Figure 4 illustrates DEA frontiers under CRS and under VRS. The frontier is drawn slightly differently to Figure 3 to introduce how the concepts VRS and CRS are important in DEA. The section AB of the VRS frontier exhibits increasing returns to scale (output increases proportionately more than inputs), BC exhibits CRS, and CD decreasing returns to scale (output changes proportionately less than the change in inputs). For a given hospital, G, the distance EF measures the

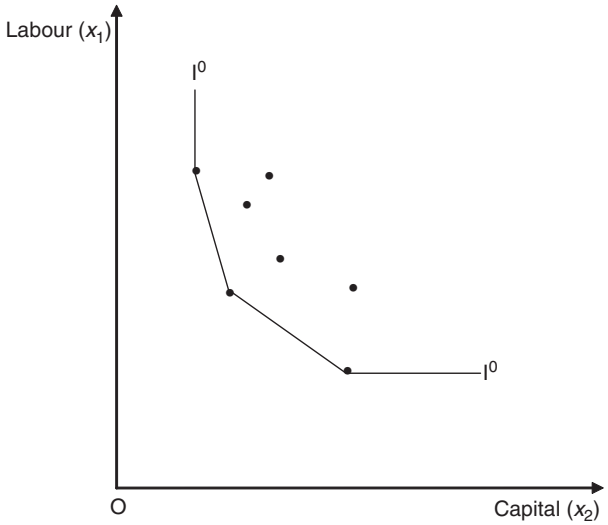


Figure 3 The DEA production frontier.

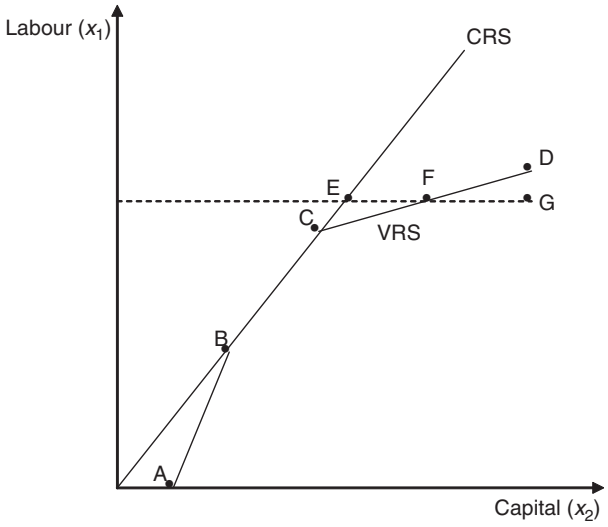


Figure 4 Constant and variable returns to scale under DEA.

effects of economies of scale in production, and FG measures 'pure' inefficiency. Clearly, more hospitals will be deemed to be efficient under variable returns to scale, as under an assumption of CRS any economies of scale are included in the measure of inefficiency.

DEA (in the formulation presented above) does not account for the influences of the distribution of medical case complexity (casemix) on producer efficiency in the production of health care. One approach to modeling the effects of casemix is to include the patient characteristics (for patients at different health care hospitals) as a type of input in the production frontier. However, this approach may be inconsistent with economic theory, as patients are not inputs which are transformed to make the final product (which in this case is a health care intervention). Instead, patients consume treatments to (hopefully) produce improvements in their health status.

The characteristics of patients and their illness will influence the production of health care in order to produce these health status improvements, hence patient illness differences (e.g., the intensity of a heart attack, or the stage of a cancer) may be better viewed as factors which shape the outputs rather than inputs in the production process. DEA models can incorporate this approach to patient illness characteristics (casemix factors) by modeling the effect of casemix on the overall production process by adjusting outputs by casemix group. Another method involves adding a second stage of analysis to the DEA approach. The first stage of the model involves running a DEA model based on physical inputs and treatment-based outputs to yield efficiency scores for units (say hospitals again), as shown above. The second stage then takes these efficiency scores and regresses them against hospital level casemix variables to assess the impact of the patients' socio-demographic and clinical characteristics on the production process and efficiency. This allows the inclusion of variables which do not fall neatly into the input–output analysis and potentially see if they have a significant impact on the efficiency scores obtained in the first stage, but there are many statistical issues with undertaking such second stage analysis (Fried *et al.*, 2008).

Some Limitations

Before proceeding, it is important to note that DEA has several major limitations which require some care on the part of those constructing models and others interpreting the results. There are major statistical issues to account for. The technique is deterministic and outlying observations can be important in determining the frontier (made up of the most efficient units). Closer investigation of these outliers is often warranted to ensure the sample is actually uniform in nature, i.e., one really is comparing like with like. Care must be taken in interpreting results as the DEA frontier may have been influenced by stochastic variation, measurement error, or unobserved heterogeneity in the data. DEA makes the strong and nontestable assumption of no measurement error or random variation in output. Small random variation for inefficient hospitals will affect the magnitude of the inefficiency estimate for that hospital. Larger random variation may move the frontier itself, thereby affecting efficiency estimates for a range of hospitals.

DEA is sensitive to the number of input and output variables used in the analysis. Overestimates of efficiency scores can occur if the number of units relative to the number of variables used is small. A general rule of thumb is that the number of units used should be at least three times the combined number of input and output variables.

DEA only provides a measure of relative efficiency in the sense that: a hospital which is deemed efficient using DEA is only efficient given the observed practices in the sample which is being analyzed. Therefore, it is possible that greater efficiency than that observed could be achieved in the sample.

The Malmquist Index

Efficiency can change over time, and DEA based Malmquist indices (named after a pioneering researcher in this area)

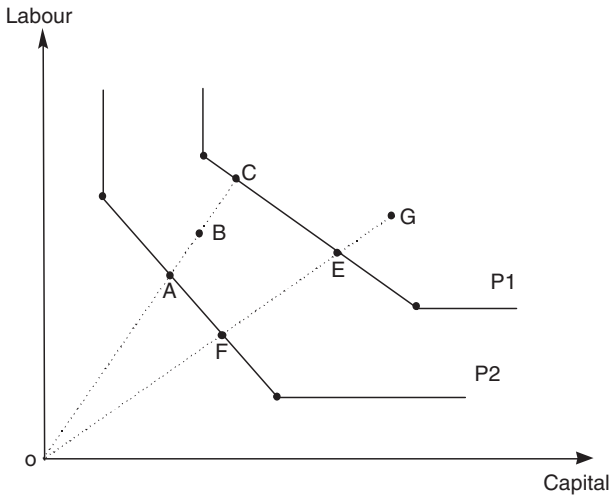


Figure 5 Malmquist index.

reused to measure this concept of productivity. The Malmquist productivity index (Fried *et al.*, 2008) is defined as (with reference to **Figure 5**, a two input, one output model, two time period, where G and B represent a hospital in two different time periods):

$$MPI = \left[\frac{OE/OG}{OC/OB} \times \frac{OF/OG}{OA/OB}\right]^{0.5} \tag{10}$$

The index is the geometric mean of two indices. In the first the production frontier of period 1 (P1) is taken as given and measures the distance of the two production points, G and B, from it. The second index is similar except the reference frontier is that of period 2 (P2). A score greater than unity indicates productivity progress as a hospital delivers a unit of output in period 2 using less inputs. In other words, the hospital in period 2 is more efficient relative to itself in period 1. Similarly, a score less than unity implies productivity regress and constant productivity is signaled by a unit score. The index can be decomposed:

$$MPI = \frac{OE/OG}{OA/OB}\left[\frac{OA}{OC} \times \frac{OF}{OE}\right]^{0.5} \tag{11}$$

The component outside the brackets is the ratio of TE in each period and measures efficiency change when moving from period 1 to period 2. It indicates whether the hospital gets closer to its production frontier, i.e., becomes more efficient (with a score greater than unity), or moves further away from the frontier, i.e., becomes less efficient (with a score of less than unity), or stays the same (with a unit score). The second component of the Malmquist index in eqn [10] captures technological change evaluated from both time periods, i.e. movements of the actual frontier itself – the technology with reference to which a sample operates. The frontier (i.e., technology) can progress (with a score greater than unity), regress (with a score of less than unity), or stay in the same position (with a unit score). Malmquist indices are increasingly used in health care.

Stochastic Frontier Analysis

SFA, see Coelli *et al.* (2005) has been used in a much smaller number of efficiency analyses in health care than DEA, but the number of papers is increasing. SFA on cross sectional data decomposes a regression error term into two parts. Given a model of the form:

$$y_i = \beta_i \cdot x_i + u_i + v_i \tag{12}$$

where y_i is the vector of outputs, x_i is the vector of inputs, β is the vector of parameters (of little interest in the context of these models) u_i is the one sided inefficiency term ($u_i \geq 0$ for all i), v_i is the two sided error term which is assumed to follow the usual classical linear regression model error term, and u_i and v_i have zero covariance. Note i,u,x,v all are now discussed with separate and new meanings to the equations in the DEA models above.

The first of the two error terms is a one-sided 'error' term that acts as a measure of inefficiency. By constraining this term to be one-sided, production units can only produce on or below the estimated production frontier. The second part is the 'pure error' term that captures random noise, and has a two sided distribution. The one sided constraint on the distribution of the inefficiency term allows a realized production frontier to be estimated, and each producer's efficiency to be measured relative to that frontier.

The use of SFA in the production of health care has received increasing attention over recent years. This is partly because of increased interest in efficiency measurement in general in health and health care, as discussed earlier, as discussed earlier but also because of advances in modeling techniques and increased computing capabilities.

To allow multiple outputs to be modeled (as outputs in health care are typically heterogeneous) researchers often estimate cost rather than production frontiers. Estimation of an SFA production frontier requires that all outputs can be meaningfully aggregated into a single measure. This assumption is questionable in the health context. However, costs can be easily aggregated into a single measure using monetary units such as dollars. The estimation of the cost frontier remains a valid method for examining productive efficiency as it is the dual of the production function. The cost frontier formulation of the model is:

$$c_i = \mathrm{f}(p_i, y_i, z_i) + u_i + v_i \tag{13}$$

where c_i is expenditure at hospital i, p_i is a vector of input prices, and z_i is a vector of producer characteristics which includes casemix variables. The inclusion of variables capturing casemix and producer characteristics in the model allows statistical testing of hypotheses concerning the relationship between these factors and producer efficiency.

The stochastic frontier model is estimated by maximum likelihood and requires that the researcher specifies an appropriate distribution for the inefficiency term. The most commonly adopted approach for cross-sectional data is to assume that u_i follows a half-normal distribution:

$$u_i^* = |u_i| \tag{14}$$

and

$$u_i^* \sim \mathrm{N}(0, \sigma_u^2)$$

Other distributions suggested for cross-sectional data include the exponential and gamma distributions. However, there are no strong a priori theoretical reasons for choosing any of the above distributions over each other. It has been argued that this has led to arbitrary and nontestable assumptions about the distribution of the inefficiency term, which are a potential source of model misspecification. Another approach adopted has been to use panel data which has the advantage that it requires no specific assumption about the distribution of u_i (Fried *et al.*, 2008).

Assumptions concerning the error term v_i in SFA may also be important. If the assumption of normality in the error term does not hold, and its distribution is skewed, inefficiency may be under or over estimated (Jacobs *et al.*, 2006). Because the error term v_i is assumed to show zero skewness, any skewness is attributed to the inefficiency term u_i. For instance, periodic capital repairs to a hospital may lead to a positive skew in total cost and hence in the error term. Under a stochastic cost frontier model this will result in inefficiency being detected, even if the hospitals studied are perfectly efficient. Conversely, a negative skew on the error term will bias the estimate of inefficiency downwards. Further, SFA may also reject the null hypothesis of no inefficiency too readily.

The SFA cost frontier is often estimated using a generalized functional form known as a 'translog' function, which allows the testing of a wide range of assumptions about the nature of the cost function, and does not impose restrictive a priori assumptions on its functional form. Translog multiproduct cost functions can also be used easily to test for the presence of economies of scale and scope. However, this approach requires a large number of degrees of freedom. In hospital studies, where sample sizes are often small, this may introduce measurement error and bias in inefficiency estimates through the inappropriate aggregation of inputs and outputs. An alternative approach is to impose a functional form which is less demanding on the data (e.g., Cobb-Douglas), but this may come at the price of introducing misspecification into the model.

Making Best Use of Efficiency Measures in Health Care

It has been postulated that efficiency measurement studies in health care are being produced at an increasing rate, but there is a limited amount of use of such studies in practical terms. Criteria have been suggested previously for assessing the use and usefulness of such studies, from the perspective of the supplier of such studies, and those who might make use of them (Hollingsworth, 2012).

Use and Usefulness Criteria for Suppliers and Demanders

Suppliers

1. Applied research needs to be placed in a policy context. One important element of any efficiency analyses is to get potential end users involved early on. This helps 'ownership' of the research from the users' perspective, and keeps the researcher on track. This may initially involve finding the right person, or group of people (having a number of people involved reduces risks, e.g., staff moving positions). Meetings to feedback results at various stages, and to different levels of users, for example, hospital managers, health department staff, will help make sure information is provided to those who want to use it. An advisory group to initially help set up model specification may be useful.
2. Hospital managers may have concerns about health authorities using efficiency measures as 'big sticks' and are generally interested in more detailed information on their specific unit, whereas health authority staff tend to be more interested in the overall picture and comparisons between hospitals. The researcher has to balance these views and providing all the information to everyone may help. One should also ask what information it would be useful to provide that the data/modeling is not providing right now, and try and accommodate this, or suggest means (e.g., extra data) which could help.
3. Hase the objective of giving end users the information been met? Surveying them, perhaps including a short report, may help refine the measures. Disseminate the results as widely as possible. Make sure users know the limitations of efficiency measures, and that they are a useful policy tool, not the useful policy tool. Results can be manipulated so full provision of information to all may be helpful.
4. Are the right questions being asked?
5. What is the underlying economic theory of production (or cost, does duality theory and the requirement for cost minimization as an objective really apply)?
6. Is the model specified correctly? Hasan extensive sensitivity analysis been undertaken? Ask the advisory group if there are any obvious omitted variables.
7. Are the data really good enough to answer the questions, particularly the output data?
8. Is there any data on quality of care? What will results using just quantity (throughput) data really show? Will any inefficiency be just made up of omitted quality data?
9. If quality data is available, how will it be weighed relative to quantity data, to avoid it being 'swamped' by relatively large numbers of throughput information? Unless carefully weighted, potentially vital information on quality may have little impact on results.
10. Is the sample inclusive enough,and is one comparing like with like? Exploratory analyses are useful. Just because all hospitals in the sample have the sample categorization, there may be a rogue specialist unit or teaching hospital that may confound the results. Frontier techniques are very susceptible to outliers. Sample size is also an issue.
11. If one is happy with the data and models, what techniques will be used, DEA, SFA or both? If there are multiple inputs/outputs, nonparametric techniques have an advantage (when comparing DEA and SFA) in terms of disaggregation (Coelli *et al.*, 2005). They allow one to feedback more detailed information on areas of inefficiency. Panel data techniques will also allow one to feedback more information, not only on what happens

between units, but also what happens over time. Looking at trends over time is more useful than a snap shot.
12. Is two stage analyses being undertaken, if so how are any statistical problems being accounted for?
13. Does one need to generate confidence intervals? Unless one is certain that the sample is all inclusive, then one might wish to account for sampling variation.

Demanders

Table 1 presents a checklist for assessing if an efficiency analysis should be judged as potentially useful. This (again) is a starting point, based on the Drummond *et al.* (2005) list for assessing economic evaluations. Suppliers of efficiency studies may also wish to take note of these points. The following two assessment questions asked by Drummond *et al.* (2005) are also pertinent here: Is the methodology appropriate and are the results valid; and if the answer to this is yes, then – do the results apply in this setting? As Drummond *et al.* (2005) acknowledge, it is unlikely every study can fulfill every criteria, but criteria are useful as screening devices to identify strengths and weaknesses of studies, and of course to identify the value added by comprehensive extra analysis of this nature.

Table 1 A checklist for assessing efficiency measurement studies

1. Is the question well defined, and answerable?
   - Are the inputs and outputs clear?
   - Is there a particular viewpoint stated (whose objectives are accounted for – managers, Government policy makers, patients?), is any decision making context established?
2. Is a comprehensive description of the sample given?
   - Can you tell if any relevant comparator units are excluded?
   - Is the sample strictly comparable, are there potential outliers?
3. Are the quality and quantity output data clear and comprehensive?
   - Where do the data come from, who collected them, and why?
   - Are quantity data case mix adjusted?
   - Are quality data useful, for example, can individual patients be followed through the system?
4. Are all the relevant inputs and outputs included?
   - Is the range wide enough to answer the research question?
   - Do they cover all relevant viewpoints (e.g., hospital mortality may be of interest to patients, scale of operation to policy makers, and range of services to managers).
   - Are there measures of physical quantities of inputs as well as costs (although in a number of contexts costs alone may be appropriate)?
5. Are inputs and outputs measured accurately in appropriate units?
   - Are all resources used relevant to the analysis accounted for?
   - Are any data omitted? If so what is the justification?
   - Are there any special circumstances, which make measurement difficult, for example, joint use of staff? Were these circumstances handled appropriately?
6. Were inputs and outputs (or objectives) valued (or weighted) correctly?
   - Were the sources of all values clearly identified? for example, market prices for inputs, case mix weights?
   - Was the value of outputs appropriate? Were the right weights placed upon the relationship between quantities (and qualities) of outputs?
7. Were analyses over time undertaken?
   - Were values (and outputs) adjusted to present value?
   - How are the specific techniques justified, for example, are random or fixed effects models used, how is scale accounted for, how is efficiency decomposed?
8. Do techniques add incremental value?
   - For example, is data envelopment analysis used? Or stochastic frontier analysis? Which cross sectional or panel data (over time) techniques are used?
   - Are the techniques used justified clearly, for example, what incremental value do they add beyond how efficiency is currently measured?
9. Was allowance made for uncertainty?
   - Were appropriate statistical analyses undertaken?
   - Were sensitivity analyses performed, which dimensions are tested?
   - Were the results sensitive to the statistical/sensitivity analysis?
10. Did the presentation and discussion of study results include all issues of concern to users?
    - Were the conclusions based on an overall measure, or individual comparisons of efficiency?
    - Were the results compared with others who have investigated the same question?
    - Did the study discuss the generalizability of the results to other settings?
    - Did the study allude to other important factors in the decision or choice under consideration, for example, ethical issues, or access issues, or equity?
    - Did the study discuss issues of implementation, such as the feasibility of adopting efficiency changes, given existing operational constraints, and whether freed resources could be redeployed to other more efficient programmes?

Source: This checklist relies heavily on Box 3.1 in Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. and Stoddart, G. (2005). *Methods for the economic evaluation of health care programmes*. Oxford, UK: Oxford University Press.

Summary

The number of studies which seek to measure health service efficiency and productivity continues to increase quite dramatically. Research in this area should be reviewed carefully and the results of studies interpreted and used cautiously, as it is still an area under development. Estimated results can be sensitive to changes in the basic assumptions and specifications of the models used, and the characteristics of the environment in which the units operate. Thus, as concluded previously, the results may only be valid for the units under investigation raising generalizability issues.

A number of criteria are suggested for judging whether research published in this area is potentially useful in a policy context. It should be noted that, as with the original economic evaluation criteria on which they are modeled, these criteria should be used as a means to interpret results, not a checklist for dismissing the usefulness of individual studies on a generic basis. What is of no use to one user may be very useful to another, working from a different viewpoint in a different health system.

In terms of 'best practice' for undertaking efficiency studies, it may be that the use of multiple techniques might help indicate trends in inefficiency. If the multiple techniques (parametric and nonparametric, including techniques which can account for multiple objectives) point to the same inefficient organizations, and the organizations cannot sensibly explain them away (i.e., omitted variables and policy shocks), then perhaps some form of inefficiency is being picked up. Of course it may be that in certain circumstances one method is obviously more useful: for example, when there are multiple outputs, SFA may not be appropriate because of problems with having to aggregate variables. Justification of the method used is sometimes difficult at present as there are few criteria for which is 'best,' although in practice different measurement methods often show similar results. Another danger at present is relying on exact numbers: small differences in inefficiency may not truly reflect inefficiency, and should be viewed with caution. Trends over time may be more reliable.

As economists the basics of what is meant by efficiency should be kept in mind. However, not only must one decide how efficiency and productivity is measured (efficiency changes over time in the context here), but also why, and how important it is relative to other societal objectives in terms of the delivery of health care. These are all questions left to be answered in a research context.

See also: Efficiency in Health Care, Concepts of. Theory of System Level Efficiency in Health Care

References

Coelli, T., Rao, D. S. P., O'Donnell, C. J. and Battesse, G. (2005). *An introduction to efficiency and productivity measurement*. New York: Springer.

Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. and Stoddart, G. (2005). *Methods for the economic evaluation of health care programmes*. Oxford, UK: Oxford University Press.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society: Series A* **120**(3), 253–281.

Fried, H., Lovell, C. and Schmidt, S. (2008). *The measurement of productive efficiency and productivity growth*. New York: Oxford University Press.

Hollingsworth, B. (2012). Revolution, evolution, or status quo? Guidelines for efficiency measurement in health care. *Journal of Productivity Analysis* **37**(1), 1–5.

Jacobs, R., Smith, P. C. and Street, A. (2006). *Measuring efficiency in health care*. Cambridge, UK: Cambridge University Press.

Further Reading

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science* **6**(4), 203–218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics* **17**(10), 1107–1128.

Hollingsworth, B., Dawson, P. and Maniadakis, N. (1999). Efficiency measurement of health care: A review of non-parametric methods and applications. *Health Care Management Science* **2**(3), 161–172.

Hollingsworth, B. and Peacock, S. (2008). *Efficiency measurement in health and health care*. UK: Routledge.

Fertility and Population in Developing Countries

A Ebenstein, Hebrew University of Jerusalem, Jerusalem, Israel

Glossary

Demographic dividend Term describing the benefit to a country of having a large working population following a fertility slowdown.

Demographic transition Theoretical model used to explain population changes over time from a context characterized by high fertility and mortality rates to low fertility and mortality rates.

Dependency ratio An age–population ratio of those typically not in the labor force and those typically in the labor force. It is calculated by dividing the number of people younger than 15 years and older than 65 years by the number of people aged 15–64 years.

Hypergamy Marriage into an equal or higher caste or social group.

Missing women Pattern of high sex ratios in census data indicating sex discrimination toward females.

Patrilocality Custom in many societies with son preference that adult children live with the husband's parents.

Replacement rate The number of children each woman needs to have to maintain current population levels.

Sex ratio Ratios of males to females in the population.

Introduction

In the mid-twentieth century, many developing countries experienced a 'demographic transition': a transition from a society in which women had many births and many infant deaths, to a society with lower fertility and lower infant mortality. This pattern was particularly pronounced in China and India, which enjoyed rapid improvements in public health and steep declines in death rates among infants and children. In the early 1960s, following sharp declines in infant mortality which had exceeded 100 per 1000, the total fertility rate (TFR) – the number of children a woman would have in her lifetime at prevailing age-specific rates – of both countries exceeded six births per woman, resulting in massive young cohorts. Government policies and changing social norms led to rapid fertility decline in the 1970s in China and in the 1990s in India, leaving both countries with massive cohorts born during their respective baby booms, and much smaller cohorts before and after. This peculiar age structure is associated with a set of advantages and challenges that will be discussed later in this article.

A similar story has begun to play out in sub-Saharan Africa, where recent declines in mortality have led to a rapid increase in population growth. Much of Africa's population is extremely young, posing a challenge in the short run but possibly aiding economic growth in the long run. Africa's age structure is also affected by the human immunodeficiency virus (HIV)/acquired immune deficiency syndrome (AIDS) epidemic, which generally affects young adults, leaving children and the elderly behind to fend for themselves. This has resulted in a very young age distribution in Africa, similar to the situation in China and India in the 1970s and 1980s. The lesson of China's and India's present may be useful for Africa's future.

The rapid fertility decline in China and India was also accompanied by an alarming pattern: the 'missing girls' phenomenon. The combination of traditional son preference, the need to reduce fertility, and the diffusion of ultrasound technology led to a sharp increase in the sex ratio at birth in both countries. Scholars estimate that more than 100 million girls are missing worldwide, 80 million of which are due to sex discrimination in China and India alone. Both countries are at the cusp of an explosion in the sex ratio of the adult population, which may have important implications for society in general, and health in particular. Recent increases in China's syphilis rate have alarmed policymakers, and the dynamics of both countries' populations could generate a challenging scenario for public health officials.

In this article, the author examines the causes and consequences of these population patterns, focusing on health as an outcome. The author begins in Section A Modern History of Fertility in Developing Countries with a general overview of fertility trends that gave rise to rapid demographic transition. The experiences of China, India and Africa are examined, as each are at a different stage of the demographic transition. In Section Demographic Transition and the Implications for Economic Growth and Public Health, issues related to where each country finds itself in the context of its demographic transition are examined. For China, the most pressing concern is to provide old age support for its rapidly aging population. For India, the challenges the country faces in providing medical care to its large young population are described. How the African experience with HIV/AIDS will shape its country's future, in light of the disease's pronounced effect on the age distribution is examined. In Section Missing Women and Implications for Public Health, the focus is on the impact of China's and India's skewed sex ratios on health in a variety of contexts, including its impact on sexually transmitted infections (STI), care for infants, and other pathways, such as the emergence of a large unmarried elderly population. In Section Conclusion, the author concludes with a brief discussion of policy recommendations for public health planning in the developing world as it relates to the demographic patterns observed.

A Modern History of Fertility in Developing Countries

The demographic transition involves four stages. In the first stage, society is characterized by high birth and death rates that keep the population in balance. All human populations

 doi:10.1016/B978-0-12-375678-7.00110-3

are believed to have had this balance until the late eighteenth century, when this stage ended in Western Europe. Developing countries found themselves in this predicament of high birth and death rates until the twentieth century.

In the second stage of the demographic transition, the death rate drops due to improvements in food supply, sanitation, and access to medical care, leading to lower infant mortality rates and longer life spans. The size of the population grows rapidly during this phase, and the decline in death rates among infants and children result in a very young population. In the third phase, birth rates fall due to several factors. These include increased access to contraception, reduced need for farm labor, and increased participation of women in the workforce. A key factor in lowering the fertility rate is a growing recognition among parents that births will likely survive to adulthood, reducing the need for very high fertility to compensate for high child death rates. This gives way to the fourth phase, where countries experience low birth and low death rates, and balance reemerges, slowing population growth.

The Demographic Transition in China, India, and Africa

The current phase of each region analyzed is shown in Figure 1, showing China near the conclusion of its transition, India in the transition process, and Africa which is yet to experience transitional fertility decline.

China

In China, the demographic transition narrative fits the country's population history tightly, and the country has now entered the last stage. In China, throughout the 1960s the TFR exceeded six births per mother. This rapid population growth alarmed Chinese officials, and the Communist Party subsequently enacted a series of fertility control policies, including new restrictions on women having more than two children during the 1970s. These early policies were immensely successful and from 1970 to 1980, the TFR fell from 5.8 to 2.3 births per woman. Family planning officials were instructed to enforce an even stricter policy starting in 1979, when China instituted its one-child policy. Under this policy, China's TFR declined to 1.5, below replacement and among the lowest rates in the world.

In the short run, the benefits to China's fertility program are indisputable. At present, the fraction of China's population that is in their working years (ages 15–64) is 73.5%. This has contributed to the country's stellar growth record, which, in turn, has been an important factor in the improvement in health outcomes. Recent estimates from nationally representative surveys put life expectancy at birth at 74.8 years for females and 72.8 for males, levels that approach those of the world's more developed countries. However, a crisis is looming. The size of the country's population aged 60 and above will increase dramatically in the coming years, growing from 200 million in 2015 to more than 300 million by 2030. The challenges stemming from this rapid population aging is discussed in the next section.

India

The Indian population narrative is similar to China's, but occurred roughly two decades later. Between 1951 and 1976, India's crude death rate dropped by more than half, from 28.6 to 13.8 – and the crude birth rate only fell by a quarter, from 45.9 to 34.4. This period featured rapid population growth, and India's improvements in infant health continued during the 1980s and 1990s.

The population explosion has left India with a very young population, and on the cusp of becoming the world's most populous nation – possibly by 2020. At present, more than half of India's population is under 25 and more than 65% is below the age of 35. In recent years, Indian fertility has slowed, partly due to government mandates and partly through the normal mechanisms highlighted in the demographic transition framework, such as increasing female education, which has led to wider take up of contraception. Birth cohorts in recent years are smaller than in the previous decade, as reflected in Figure 1. Still, India's explosive population growth for several decades has left the country with an extremely young population.

As a result of this currently favorable age distribution, India is currently enjoying its demographic dividend, with economic growth exceeding 7% every year since 1997. The country continues to enjoy a low dependency ratio, with 65.2% of the population in their working years. However, the country still lags behind developed countries in life expectancy. Life expectancy at birth for men is 66.1 years and for women 68.3 years, reflecting challenges in providing adequate health care to its massive population. The country has also struggled with providing sufficient primary and secondary education. Further investments in health and human capital can position the country to continue cashing in its demographic dividend. However, although India is still decades away from facing an aging population, the country will almost certainly face challenges similar to those that China will face, albeit in a delayed fashion.

Sub-Saharan Africa

During the 1980s, the population of sub-Saharan Africa grew at a rate of 3.1% per year, the highest of any developing region. The population growth occurred due to rapid mortality decline and only moderate fertility decline. In 1970 Africa's TFR was 6.7. By 1990 it had declined 12% to 5.9 with an additional decrease of 24% to 4.5 by 2010. However, childhood mortality rates declined more rapidly, with the under-five mortality rate declining from 180.6 to 125.3, a 31% decrease, between 1980 and 2010. The combined impact of rapid declines in mortality and more modest declines in fertility have left sub-Saharan Africa with a very young population, with 44% of the population under the age of 15. If the Indian and Chinese precedent is followed, it is reasonable to expect that fertility will begin to level off in Africa, though when this will occur is unclear, and less effective government fertility regulations imply that intervention will need to come from voluntary family planning participation. Should Africa succeed in encouraging faster fertility decline, the region may enjoy its demographic dividend earlier. In any scenario, however, the population should continue to grow at robust rates for many years, leaving the continent with a very young

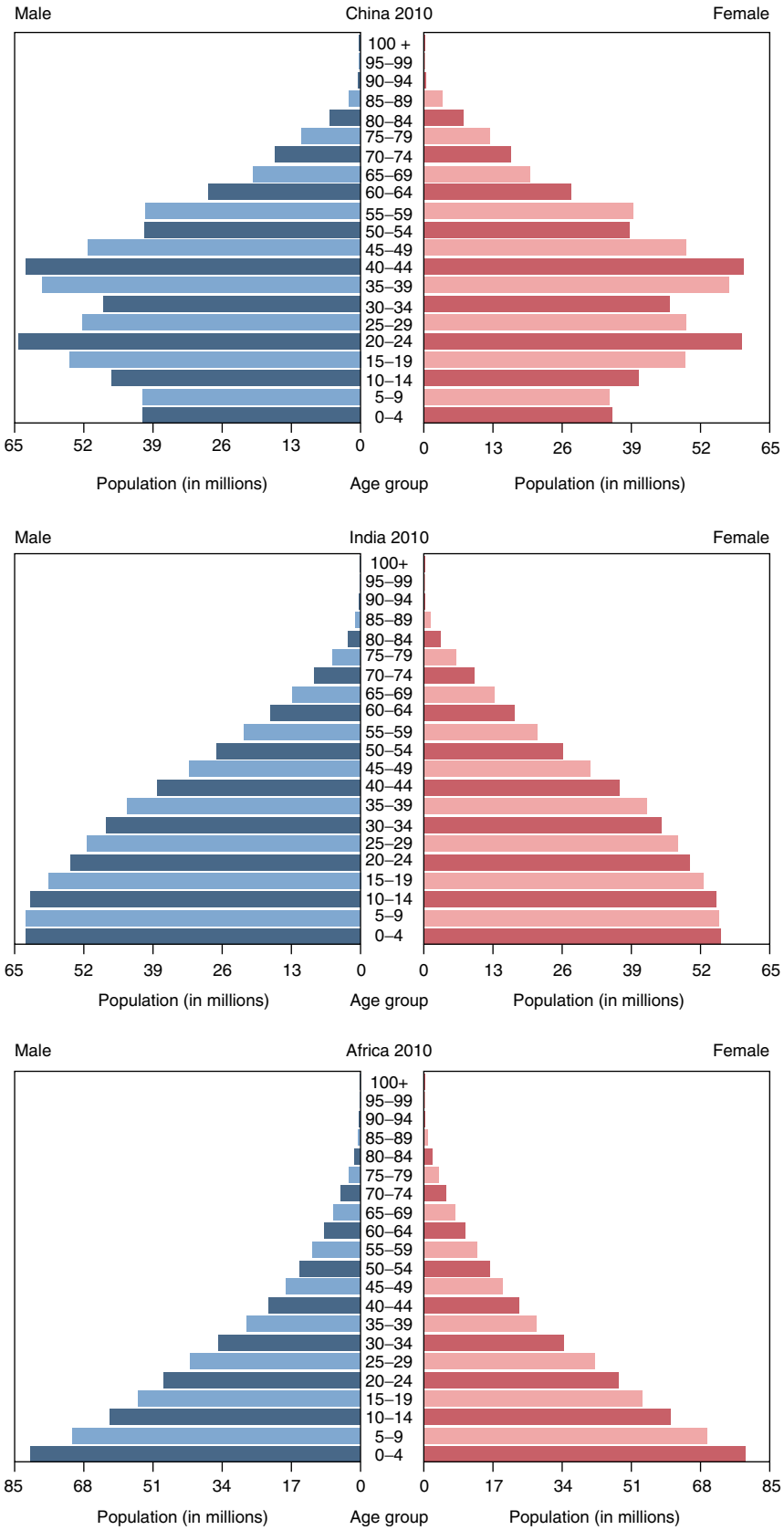


Figure 1 Age pyramid in China, India, and Africa – 2010. US Department of Commerce (http://www.census.gov/population/international/data/idb/informationGateway.php).

population in the coming decades. This could prove to be a boon to economic growth, as the eventual fertility decline and subsequent population aging will leave Africa with a huge working population. Some policymakers, however, fear that poor management of African economies may leave them unable to capitalize on the favorable age structure.

However, as shown in Figure 1, the massive young cohorts in Africa may pose a challenge in the near-term, as the region grapples with a high dependency ratio. Note that this is in part related to the consequences of the HIV/AIDS epidemic, which has resulted in millions of deaths to people who are in their prime working years, as the disease peaks in prevalence among individuals between ages 20 and 49. There is little reliable national-level data describing the distribution of deaths by cause for sub-Saharan Africa, and the World Health Organization's mortality database lists HIV-related causes for only one sub-Saharan nation (South Africa). An examination of cause-specific death data available for two countries, Tanzania and South Africa, revealed an increase in the probability of dying between ages 15 and 50 from HIV-related causes of up to 127% for males and 153% for females. Recent evidence indicates, though, that deaths from HIV have begun to plateau, which is an encouraging sign that the epidemic will not continue to worsen. However, for several high-prevalence countries such as Botswana and Zimbabwe, HIV has shortened life expectancies by several decades. A lack of further progress containing HIV could prevent the region from enjoying the benefit of its favorable age distribution, should the population of workers continue to suffer from high mortality.

The Missing Girls of China and India

As China and India experienced rapid fertility decline, many parents were unwilling to complete fertility without having a son. The value of sons is in part religious, as both Confucianism and Hinduism designate the son as having the responsibility to perform certain rites. However, a primary explanation for son preference is the custom of patrilocality, practiced in both countries. Patrilocality refers to the firmly-entrenched cultural norm for elderly parents to coreside with their adult son, and for a woman to 'marry in' and assist him in this function. Patrilocality is the custom in almost every country with missing women. In a world without social security and with limited ability among individuals to generate financial wealth, this is the primary method of guaranteeing support in one's old age. In this context, it is perhaps unsurprising that parents have resorted to sex selection in a period of fertility decline, when parents will have to rely on fewer children to care for them in their old age.

When Amartya Sen first coined the term 'missing girls' in a 1990 New York Review of Books article, it was unclear exactly how these women went missing. Although some presumed that daughters suffered higher mortality rates throughout childhood, later scholarship documented that infanticide and sex-selective abortion were the primary explanations, with the latter becoming increasingly prominent after ultrasound's diffusion in China in the late 1980s and early 1990s in India.

Historically, Chinese and Indian parents discriminated against girls on birth and throughout childhood to ensure the survival of a son to adulthood. However, this practice was muted during the baby boom of the 1960s, which allowed the vast majority of parents to have an adult son without engaging in sex selection. However, in both China and India, increasingly strict enforcement of fertility limits put parents in a more difficult position. Strict enforcement of China's one-child policy throughout the 1980s forced parents to curb fertility. In India, overzealous promotion of family planning occurred through activities such as sterilization camps, and the country later adopted a two-child limit for public officials. In both countries, the need to have a son at an early parity became paramount. Following the introduction of ultrasound technology, parents were able to identify the sex of the fetus after 4 months of pregnancy, a technology that significantly lowered the time and psychic cost of engaging in discrimination against girls. A steep rise in the sex ratio at birth was observed in both countries in the 1990s, and has remained disturbingly high. As shown in Figure 2, this increase was concentrated among births following daughters, when parents would have felt compelled to have a son but be in violation of the one-child policy.

The most recent census data for both countries indicates that the sex ratios are at the highest levels ever recorded for each country. The naturally occurring sex ratio at birth is 106 (106 boys for 100 girls). In China, the 2005 Chinese Population Survey and the 2010 census reported that the sex ratio at birth was 118 and 119 males to females respectively, suggesting that the distorted sex ratios will continue to be a problem well into the twenty-first century. In India, the problem is somewhat less severe, though still shocking in magnitude. India's 2011 sex ratio among ages 0–6 was 109 as a ratio of males to females, representing deterioration from the 2001 sex ratio of 108. In Northern Indian states with strong son preference such as Haryana and Punjab, the ratios are similar to those in China, with reported sex ratios of 120 and 118, respectively. This long running problem has left both countries with extremely distorted sex ratios among the young. In China, there were nearly 25 million more boys than girls under 20 in the 2010 census.

Demographic Transition and the Implications for Economic Growth and Public Health

As the large cohorts born during the second phase the demographic transition enter their prime working years, a window of opportunity is provided for rapid economic growth, as slowing fertility yields a large mass of workers. However, as these cohorts enter old age, they place pressure on the system; the large mass of elderly, with smaller population cohorts before and after them, represents a challenge.

In this section, the author briefly describes a set of unique challenges facing China, India, and sub-Saharan Africa, related to the demographic transition in each context. In China, how the country will deal with a large elderly population without extensive pension programs is examined. In India, the challenges with providing health care to its large, poor, and rural population is discussed. In sub-Saharan Africa, the focus is on the most pressing concerns in the area of public health, which

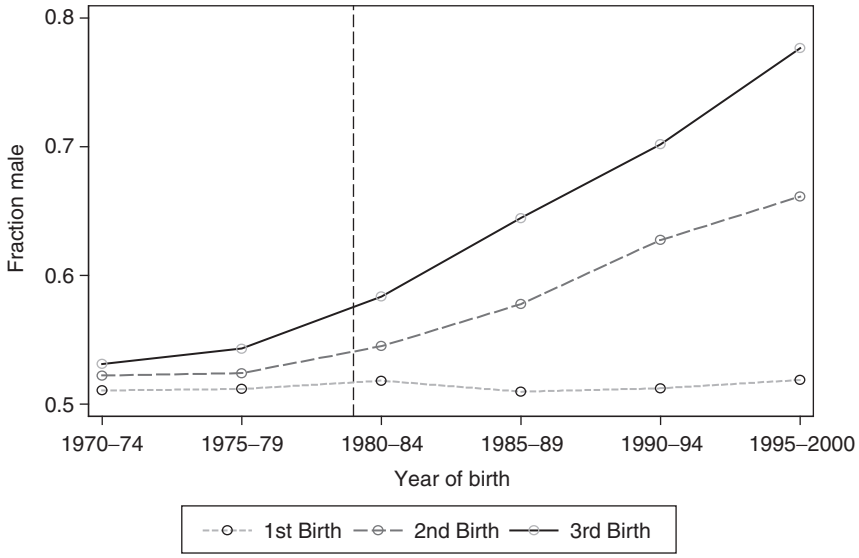


Figure 2 Sex ratios at birth following daughters, China 1980–2000. China census 1982–2000. Sample restricted to mothers ages 21–40. Vertical line indicates year of introduction of China's one child policy (1979). Reproduced from Ebenstein, A. (2010). The 'missing girls' of China and the unintended consequences of the one child policy. *Journal of Human Resources* **45**(1), 87–115. © 2010 by the Board of Regents of the University of Wisconsin System. Courtesy of the University of Wisconsin Press.

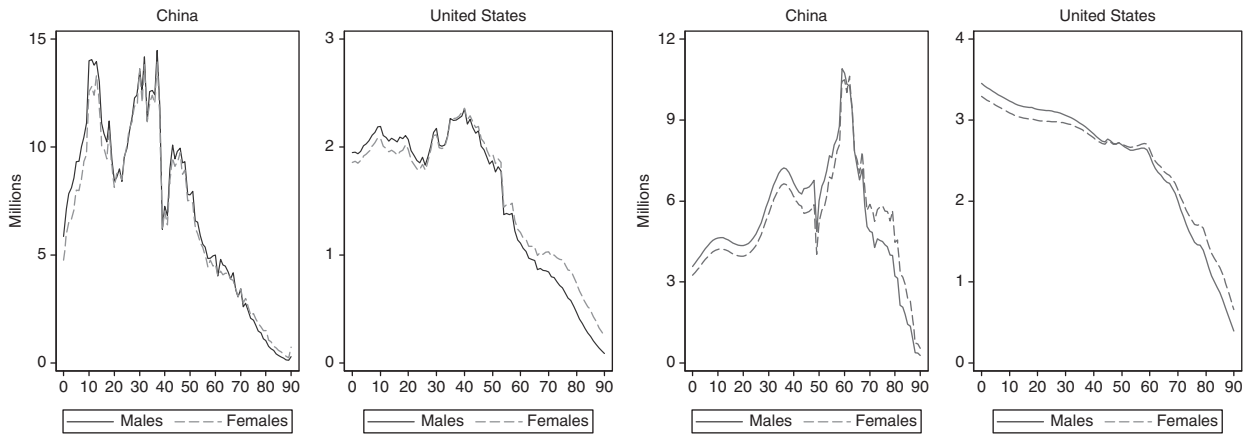


Figure 3 Age distribution in China and the US in 2000 and 2050. Results for China based on 2000 census and simulations. Results for the US taken from 2000 census and projections by the social security administration (2007). Mortality rates for China are based on Banister and Hill (2004). Reproduced from Ebenstein, A. and Sharygin, E. (2009). The consequences of the 'missing girls' of China. *World Bank Economic Review* **23**(3), 399–425, with permission from Oxford University Press.

are to lower infant and maternal mortality, and provide wider access to contraception.

China

In China, the chief implication of the age distribution is that the country has to rapidly prepare for a heavy burden on each worker to support multiple retirees. For example, the one-child policy has resulted in a 4-2-1 problem, where four grandparents turn to two adult children for support, who only have one child of their own, leaving a great burden on each young person to provide old age support. The need for pension programs in China is acute, but programs are limited. The rural pension programs attracted reasonable participation rates, especially among individuals without sons, but complications in implementing the programs prevented their expansion. The massive expansion in the elderly population forecasted has already led many to call for a relaxation of the one-child policy. However, government officials have ignored these proposals and called for an extension to the policy in its most recent five-year plan.

China's age distribution is highly skewed, relative to the US (Figure 3). China experienced two baby booms: the first in the 1960s, and the second in the late 1980s, when the earlier boom cohorts had children. However, in the wake of government-mandated fertility control, each successive cohort in China is now smaller than the last. The magnitude of

China's baby boom cohort dwarfs that of the US's that occurred following World War II. Although the US is anticipated to converge to a normal population distribution with a modest fraction of elderly in the population, China is predicted to have a massive population of retirees. This will place pressure on the system to provide for these retirees later in the twenty-first century. Forward-thinking policy would dictate that the government access funds from the current generation of workers to provide for the future generation of retirees, as it seems unlikely that the next generation of workers will be able to support the large population of retirees.

India

In India, a critical challenge is how to provide proper care to the massive young, poor, and primarily rural population. India's young population, if provided proper access to education and health care, should allow the country to be highly productive for several decades. New initiatives have been launched in India, such as the National Rural Health Mission, which will serve to increase access to medical professionals in India's rural areas. Challenges have also plagued the expansion of rural health insurance. While 70% of India's population lives in villages, less than 2% is insured. Issues of cost sharing and access to services have made insurance either not financially viable or unattractive.

In many rural areas, there is an insufficient supply of properly trained physicians. In areas with skilled physicians, absenteeism is a challenging issue. It has been estimated that absenteeism can be as high as 40% among primary health providers and among teachers. They found absenteeism rates were related to the quality of infrastructure, and doctors were often working more hours at private facilities instead of publicly accessible facilities. This highlights the challenge of making medical services affordable and available.

Sub-Saharan Africa

Sub-Saharan Africa faces a set of unique challenges in the context of its demographic transition. The two primary issues are the need to (1) lower infant and maternal mortality and (2) expand access to contraception. Maternal mortality in sub-Saharan Africa with 500 deaths per 100 000 births is twice as high as in the next highest region, South Asia with 220 deaths per 100 000 births. More than half of all maternal deaths worldwide occur in sub-Saharan Africa. Likewise, under-five mortality exceeds 100 deaths per 1000 births, higher than in any region in the world. Although both of these rates have declined from even higher levels, they both represent challenges to development. High childhood mortality rates prevent the proper allocation of parental resources to children who will survive, and high maternal mortality rates leave many children without proper parental support. Both represent challenges necessary for sub-Saharan Africa to overcome in order to exit the poverty trap.

Sub-Saharan Africa's high fertility rate also poses a challenge for policymakers. For the region to enjoy a demographic dividend, fertility must be slowed. Fertility rates are highly negatively correlated with female educational attainment. This occurs through several channels affecting both desired family size and access to contraception to achieve the desired family size. Higher female education is associated with later marriage, greater autonomy of women in the household and over their fertility choices, and perhaps most importantly, higher opportunity costs of childbearing due to foregone wages. More educated women also have greater knowledge of an access to contraceptives, which is also partly responsible for lower fertility among more educated women. As such, increasing female education may be an effective policy tool for lowering Africa's fertility rate. In light of recent evidence that fertility declines in Africa are stalling, policy makers may wish to consider more proactive strategies for lowering fertility.

Missing Women and Implications for Public Health

In this section, the author examines how China's and India's 'missing girls' will affect public health in the coming years. The focus is on a set of health issues that have been examined by scholars that are related to the high sex ratios in Asian countries.

Unmarried Men in China

China is on the cusp of a dramatic deterioration in men's marital prospects. As shown in Figure 4, the sex imbalance between potential spouses is forecast to be at its worst by 2025, when the cohorts with the highest sex ratios (those born under the one-child policy) reach adulthood. China's one-child policy in combination with legislation regulating minimum age at marriage generates a problematic scenario. As birth cohorts age, they find that each successive generation is smaller than their own, giving rise to a 'kite-shaped' age distribution common in many Asian countries. It has been estimated that the fraction of men aged 25 and older who fail to marry will exceed 5% by 2020 and 20% by 2030. In the most optimistic scenario simulated, where the sex ratio returns to normal immediately, the share of men who fail to marry in 2060 will stabilize just below 10%. In light of historical patterns of hypergamy in China, it will likely be the men of lowest status who fail to marry, and the poorest regions of the country will have the highest rates of bachelorhood. This will generate a challenging situation for providing old age support at the local level as the population of 'bare branches,' or men who fail to marry and represent bare branches on the family tree, increases.

Trends in Sex Work and Sexually Transmitted Infections

Prostitution in China is widespread and has increased dramatically in recent years. Following Deng Xiaoping's campaign for economic reform in 1978, the market for sex work increased dramatically, as migration of both men and women to urban areas provided both increased demand and supply. Current estimates indicate that between 3 and 10 million women participate in this market, a steep increase from the hundred thousand estimated as recently as 1989. Informal prostitution rackets are common throughout China,

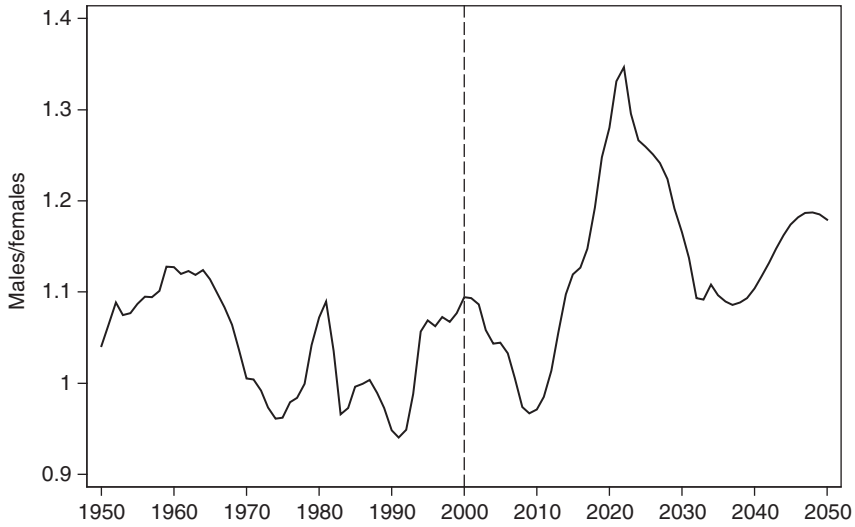


Figure 4 Sex ratio of the marriage market in China, 1950–2050. The marriage market is defined as men aged 22–32 and women aged 20–30. The sex ratio for each year is calculated using data from the 2000 census, modeling population changes with age-sex-year specific mortality rates. The population is simulated forward from 2000 using fertility assumptions described in Ebenstein and Sharygin (2009) and a sex ratio at birth of 1.09 from 2005 and beyond. The vertical dotted line indicates the year 2000. Reproduced from Ebenstein, A. and Sharygin, E. (2009). The consequences of the 'missing girls' of China. *World Bank Economic Review* **23**(3), 399–425, with permission from Oxford University Press.

sometimes involving high-school girls. However, government response is generally limited in China. Authorities attempt crackdowns through controversial 'shame parades' where Chinese prostitutes are forced to endure the embarrassment of being marched down a public street. In spite of these efforts, most scholars believe that the government is unwilling or unable to seriously tackle the problem.

In a parallel and alarming trend, China has experienced a steep increase in the syphilis infection rate, with maternal transmission rates to newborns increasing by a factor of five between 2003 and 2008 in Shanghai. Although sex work may often have ambiguous welfare consequences, in the Chinese context, the concern is clear. Chinese men visit prostitutes frequently and they are reluctant to wear condoms, which are in combination a cause for concern. The low condom use rates, lack of institutional will to reduce prostitution, and the rising sex ratio will likely create challenges as men fail to marry. In light of evidence that many women participate in prostitution while being married in general and in China in particular, this is a serious concern for the future, as concurrent sexual relationships may speed the diffusion of HIV and other STIs.

Patterns in Breastfeeding

The differential fertility behavior after the birth of sons and daughters also manifests itself in subtle ways in India. In a recent paper, it was shown that boys are breastfed for longer than girls. The mechanism is not explicit gender discrimination among living children, but driven rather by the fact that sons are often the last child. Because breastfeeding makes women less fertile, mothers looking to have another child, as is often the case after a female birth, will discontinue breastfeeding their daughters sooner than after sons. As such, boys are treated to longer durations of breastfeeding, which is documented to have important health implication in India, where drinking water is often unsafe relative to breast milk. The difference in duration for boys and girls is shown in Figure 5, and it is estimated that this explains 14% of the excess child mortality for girls relative to boys. Although historically parents exhibited explicit bias in allocation of resources to boys over girls, now developing countries are faced with more subtle but no less problematic forms of discrimination.

Sex Ratios and Social Unrest

An additional concern in China is that the high sex ratios will lead to social unrest. There are several reasons for concern over having millions of surplus males, including the possibility for China to seek out an armed conflict, as occurred in the nineteenth century following a prior episode of elevated population sex ratios. One 2007 study focuses more narrowly on the incidence of crime rates and exploits timing of the implementation of the one-child policy by province, which generates variation in sex ratios regionally. Modest effects of the adult sex ratio on violent crime and property crime were found, with the rise in sex ratios responsible for roughly one seventh of the overall rise in Chinese crime rates during the period 1988–2004. The possibility that unmarried men will generate social unrest is very plausible, and has been advanced in popular media such as newspapers. Unfortunately, the literature is scarce as the hypothesis will not be fully testable using the Chinese experience until the cohorts with extremely skewed sex ratios reach adulthood, which will occur in the next decade. This is, however, an important issue that will need to be monitored.

The Gender Gap and Female Suicide

Chinese suicide rates exhibit several unique and alarming patterns. Suicide rates in China are twice the international

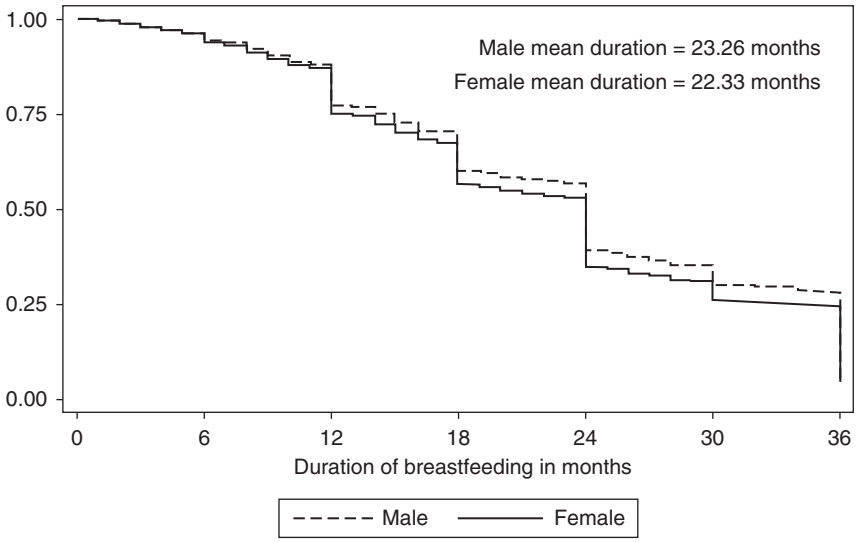


Figure 5 Breastfeeding duration by gender in India. The figure plots the proportion of children, by gender, who are still being breastfed at the duration (age) given on the horizontal axis. Reproduced from Jayachandran, S. and Kuziemko, I. (2011). Why do mothers breastfeed girls less than boys? Evidence and implications for child health in India. *Quarterly Journal of Economics* **126**(3), 1485–1538, with permission from Oxford University Press.

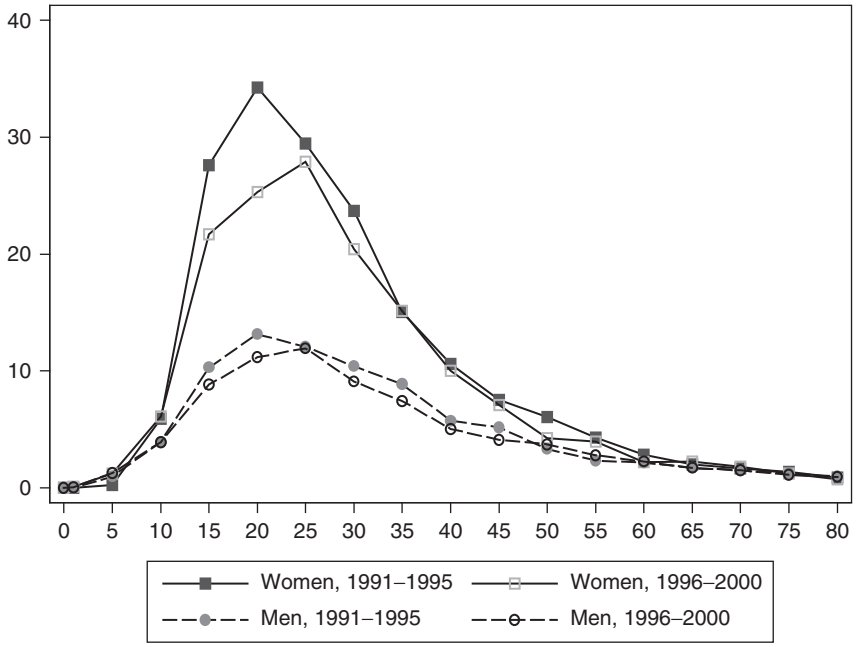


Figure 6 Percentage of rural deaths by age due to suicide, China 1991–2000. Author's Calculations from Chinese Disease Surveillance Points (1991–2000). Reproduced from Jayachandran, S. and Kuziemko, I. (2011). Why do mothers breastfeed girls less than boys? Evidence and implications for child health in India. *Quarterly Journal of Economics* **126**(3), 1485–1538, with permission from OUP.

average, and are nearly six times higher in rural China than urban China. China is also the only country where suicide rates are higher for women than men, with suicide accounting for nearly a third of deaths to young women in rural areas. In recent years, female suicide rates have declined sharply, with no parallel decrease for men, as shown in Figure 6. What explains these striking patterns in Chinese suicide? And what role has the rapid economic and social change in China played in the decline in suicide rates among women? India has also had challenges dealing with suicide among farmers, often after poor harvests, and high rates have been observed among the young. Among men, 40% of suicides were among people aged 15–29 but for women, it was nearly 60%. These patterns indicate that women continue to have difficult lives in these countries with traditional son preference. The high suicide rates in China and India among young women speak of a welfare gap by gender that has led to a serious public health concern, and is an area for future research.

Conclusion

The developing world is characterized by extreme population patterns. The rapid demographic transition of China and India has left both countries primed to capitalize on their favorable age distribution in the short run, but with challenges in the long run. Africa is now at the cusp of its own fertility decline, provided proper family planning is implemented it could likely begin to enjoy its own demographic dividend. The role that fertility change has played in determining economic outcomes in these countries is important, and will continue to be so as they each deal with the unique challenges associated with population aging, providing access to health care, and lowering mortality rates.

The high sex ratios in Asia also represent a complicated policy issue, as they relate to a set of health challenges in a wide range of contexts including crime, old age support, and prevention of STI. The impact of missing women on the future health status of these populations is not yet clear, as the cohorts born following the introduction of ultrasound technology have not yet reached sexual maturity. However, it is certain that this will be an important and challenging issue in the coming decades, and in the near future in China.

The policy lessons of the history of China and India are important for countries earlier in their demographic transition, such as those in sub-Saharan Africa. Sharp changes in fertility can generate rapid economic growth, and pull a country from a poverty trap. However, a highly skewed age distribution also generates a new set of challenges. For policymakers, it is critical to capitalize on the opportunity presented by having a large working population. This requires investment in education and health, to ensure these cohorts are productive. Eventually, these cohorts will age and represent a large responsibility, as will occur in China's near future. As such, it is critical to prepare for population aging during the period of demographic dividend. These lessons will be important as India and sub-Saharan Africa enter the next stage of their respective demographic transitions.

Acknowledgments

The author thanks Susan Schwartz and Elisheva Mochkin for excellent research assistance.

See also: Abortion. Fetal Origins of Lifetime Health. HIV/AIDS, Macroeconomic Effect of. HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Sex Work and Risky Sex in Developing Countries. Social Health Insurance – Theory and Evidence. What Is the Impact of Health on Economic Growth – and of Growth on Health?

References

Banister, J. and Hill, K. (2004). Mortality in China 1964–2000. *Population Studies* **58**(1), 55–75.

Ebenstein, A. and Sharygin, E. (2009). The consequences of the 'missing girls' of China. *World Bank Economic Review* **23**(3), 399–425.

Further Reading

Ainsworth, M. (1996). Fertility in sub-Saharan Africa. *World Bank Economic Review* **10**(1), 81–84.

Bloom, D., Canning, D. and Sevilla, J. (2003). *The demographic dividend: A new perspective on the economic consequences of population change*. Santa Monica, CA: RAND Population Matters.

Bongaarts, J. (2010). The causes of educational differences in fertility in Sub-Saharan Africa. *Vienna Yearbook of Population Research* **8**, 31–50.

Bongaarts, J., Buettner, T., Heilig, G. and Pelletier, F. (2008). Has the HIV epidemic peaked? *Population and Development Review* **34**(2), 199–224.

Caldwell, B. K. and Caldwell, J. C. (2006). *Demographic transition theory*. New York: Springer.

Coale, A. and Banister, J. (1994). Five decades of missing females in China. *Demography* **31**(3), 459–479.

Ebenstein, A. (2010). The 'missing girls' of China and the unintended consequences of the one child policy. *Journal of Human Resources* **45**(1), 87–115.

Ebenstein, A., Dasgupta, M. and Sharygin, E. J. (2012). The socio-economic implications of son preference and fertility decline in China. *Population Studies* (in press), doi:10.1257/aer.103.5.1862.

Edlund, L., Li, H., Yi, J. and Zhang, J. (2007). Sex ratios and crime: Evidence from China's one-child policy. *Review of Economics and Statistics* (in press).

Hudson, V. and den Boer, A. (2004). *Bare branches: The security implications of Asia's surplus male population*. Cambridge, MA: MIT Press.

Jayachandran, S. and Kuziemko, I. (2011). Why do mothers breastfeed girls less than boys? Evidence and implications for child health in India. *Quarterly Journal of Economics* **126**(3), 1485–1538.

Klasen, S. and Wink, C. (2003). 'Missing women': Revisiting the debate. *Feminist Economics* **9**(2–3), 263–299.

Liu, M. and Finckenauer, J. O. (2010). The resurgence of prostitution in China: Explanations and implications. *Journal of Contemporary Criminal Justice* **26**(1), 89–102.

Parish, W. and Pan, S. (2006). Sexual partners in China: Risk patterns for infection by HIV and possible interventions. In Kaufman, J., Kleinman, A. and Saich, A. (eds.) *AIDS and social policy*, pp 190–213. Cambridge, MA: Harvard University Asia Center.

Patel, V., Ramasundarahettige, C., Vijayakumar, L., et al. (2012). Suicide mortality in India: A nationally representative survey. *Lancet* **379**, 2343–2351.

Rele, J. R. (1987). Fertility levels and trends in India, 1951–1981. *Population and Development Review* **13**(3), 513–530.

Sen, A. (2010). More than 100 million women are missing. *New York Review of Books* **37**(20), 61–66.

Tucker, J. D., Sheng-Chen, X. and Peeling, R. W. (2010). Syphilis and social upheaval in China. *New England Journal of Medicine* **362**(18), 1658–1661.

Zeng, Yi, Ping, Tu, Baochang, Gu, et al. (1993). Causes and implications of the recent increase in the reported sex ratio at birth in China. *Population and Development Review* **19**(2), 283–302.

Relevant Websites

www.popcouncil.org
Population Council.

www.prb.org
Population Reference Bureau.

http://www.un.org/en/development/desa/population/index.shtml
United Nations Population Division.

Fetal Origins of Lifetime Health

D Almond, Columbia University and NBER, New York, NY, USA
JM Currie, Princeton University, Princeton, NJ, USA
K Meckel, Columbia University, New York, NY, USA

Glossary

Complement A good or service whose demand rises or falls as the price of another good falls or rises is said to be a complement.

Elasticity of substitution A measure of the degree to which one input in a production process can be replaced by another without reducing the output rate. Technically, it is the proportionate change in the ratio of the amounts of two inputs to their marginal productivities.

Flow A variable having an interval of time dimension: so much per period. Compared to a stock, which is the value taken by a variable at a particular date.

Health production function A function showing the maximum impact a variety of variables can have on a person's or people's health.

Human capital The stock of human skills embodied in an individual or group. In terms of value, it is usually measured as the present value of the flow of marketed skills (for e.g., the present value of expected earnings over a period of time). It is determined by basic ability, educational attainment and health status, among other things.

Inputs The variables that generate outputs in a production function. It includes capital, labor and the quality of such variables (e.g., health status).

Stock The value taken by a variable like health, or the services of a piece of machinery at a particular date, compared with a flow, which is a variable having an interval of time dimension: so much per period.

Introduction

Recent work in economics suggests that adverse health shocks experienced *in utero* can have long-lasting effects. Studies have linked fetal health to a variety of outcomes in adulthood, such as schooling, labor market activity, and mortality. These studies have also identified a broad array of 'nurture shocks,' including ambient pollution levels, infectious disease, and mild nutritional deficits, that can generate long-lasting consequences.

The fact that maternal health has such important consequences for the child stands in stark contrast to conventional medical wisdom of the early twentieth century, which held that the womb effectively protects the fetus. For example, during the 1950s and 1960s, expectant mothers were routinely told it was fine to drink and smoke. Policymakers felt there was little cause to aim health policy at pregnant women.

Recent findings by economists on the fetal origins of adult outcomes should help change policymakers' focus. Environmental regulation that decreases the exposure of pregnant women to pollutants, for example, may have important ramifications on the educational attainment of their children. However, understanding the exact mechanisms that tie fetal health to later-life outcomes remains a developing area of research.

Early Evidence

The 'thalidomide episode' in the late 1950s and early 1960s was a watershed event in establishing the importance of the *in utero* period. Thalidomide was licensed in 1957 and widely prescribed to pregnant women for morning sickness until 1961, when it was identified as the cause of an epidemic of severe birth defects such as missing arms and legs. This episode revealed that the fetus was more vulnerable than previously thought, and led researchers to wonder: Could shocks to maternal health have other long-term health effects?

Several aspects of this historical episode facilitate analysis of the causal effects of fetal malnutrition. First, the famine was unexpected, so the Dutch were unable to stock up on food or leave the country in anticipation. Second, it was sudden, meaning that researchers can clearly identify which children were *in utero* during the famine versus those that were unaffected. The fact that food supply was adequate beforehand means that children born shortly before the famine serve a good control, or comparison, group. Finally, famines tend not to occur in countries with good vital statistics data systems in place, the Netherlands being an important exception.

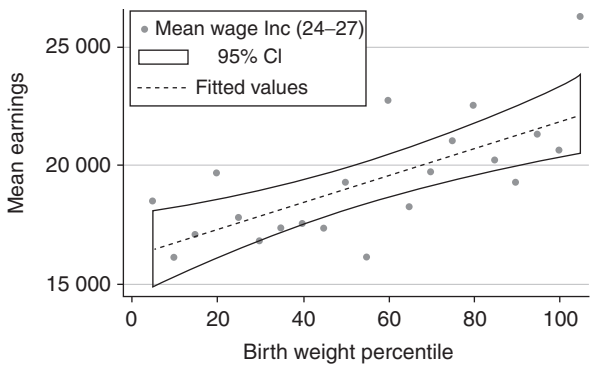


Figure 1 Mean wage earnings (age 24–27) vs. birth weight in the US, using data from the US National Longitudinal Survey of Youth.

Epidemiologists found widespread effects of the 'Hunger Winter' on maternal and fetal health. These studies show that the famine affected fertility, weight gain during pregnancy, maternal blood pressure, and infant birth weight. Results on the long-term effects on children *in utero* during the famine were initially somewhat mixed, in part because birth weight did not always seem to mediate the long-term damage (as many expected). As the affected birth cohorts aged, a more consistent pattern of adult health damage has emerged, including chronic health conditions like coronary heart disease, glucose intolerance, hypertension, and obesity.

Motivated by this evidence (and perhaps the initial controversy surrounding it) economists wondered whether adverse conditions *in utero* might: (1) affect outcomes traditionally studied in economics, such as schooling, employment, wages, and retirement, and (2) extend to a broader range of *in utero* environmental influences.

In Figure 1, wage earnings are plotted against birth weight using data from the US National Longitudinal Survey of Youth. This survey began with young people between the ages of 14 and 21 in 1978. Children born to women in this cohort have now been followed into young adulthood. As the figure shows, there is a positive correlation between birth weight and mean earnings. Descriptive findings like this encouraged economists to believe that there might be a causal relationship between fetal health and human capital.

The finding that test scores were lower in low-birth weight children was surprising as epidemiologists had posited fetal 'brain sparing' mechanisms, whereby adverse *in utero* conditions were parried through a placental triage that prioritized neural development over the development of other parts of the body.

Economists have subsequently explored the idea that fetal insults manifest later in life with numerous studies. In these studies, economists such as Janet Currie, Douglas Almond, and Michael Greenstone have looked at both the effects of large natural experiments, like the 'Hunger Winter,' as well as smaller, every-day shocks, such as pollution. Some studies compare across siblings – where one is affected by the shock but the other is not – whereas others compare across affected and unaffected cohorts. Before these studies are reviewed, a simple framework to help organize concepts will be discussed.

Conceptual Framework

One reason economists have become interested in the fetal origins hypothesis is that it holds important implications for the modeling of human capital development. In the classic health production framework, developed in 1972 by economist Michael Grossman of City University of New York, health behaves like a physical stock that serves as both an investment good and a consumption good. In this classic framework, the impact of shocks to the stock of health fades away over time. This model is applicable to many scenarios – if a child suffers a broken bone, it can heal as time passes. More formally, the formula for the health stock at time t in Grossman's model is often written as:

$$H_t = (1-\delta)H_{t-1} + I_t$$

where I_t represents investments in health capital and δ represents the depreciation rate. So, if health capital depreciates and is responsive to new health investments, then the effects of shocks to health capital tend to also depreciate over time, so that events further in the past will have less-important effects than more recent events.

Figure 2 shows how persistent a 25% negative shock to the birth endowment would be given alternative annual depreciation rates δ. Even under the lowest annual depreciation rate of 5%, half of the endowment shock is gone by the mid-teen years. For the higher depreciation rates of 10% and 15%, one

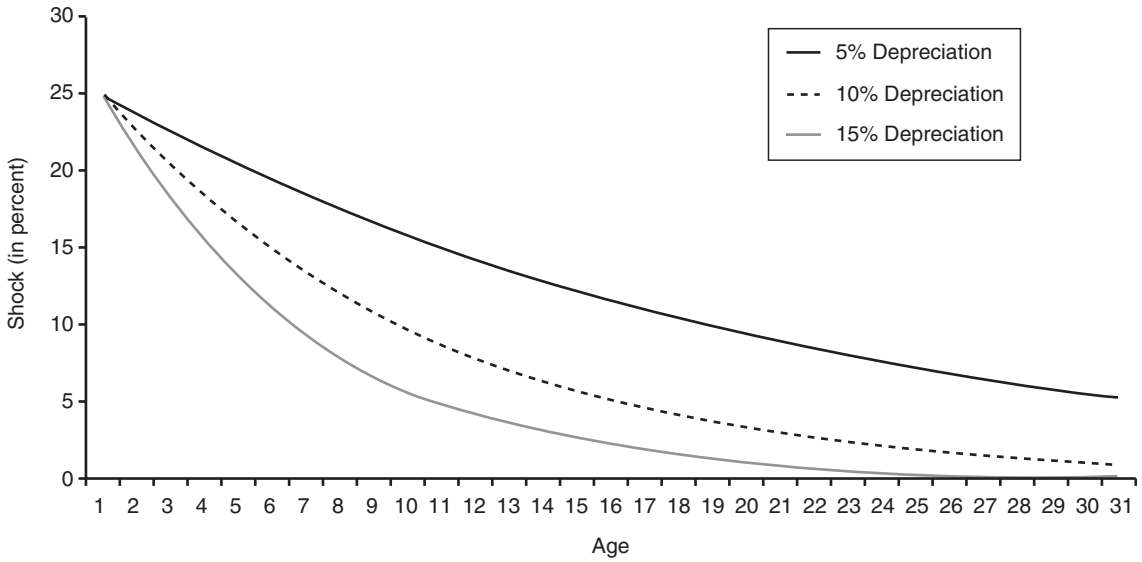


Figure 2 Shock persistence by age in the Grossman framework. Reproduced with permission from Almond, D. and Currie, J. (2011b). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives* **25**(3), 153–172.

would be hard-pressed to detect any lingering effects of the shock after age 30.

More formally, in the simplest two-input constant elasticity of substitution model capital and labor inputs are replaced with investments *in utero* and those occurring during the rest of childhood, writing:

$$H_{\text{adult}} = A[\gamma I_{\text{prenatal}}{}^{\Phi} + (1-\gamma) I_{\text{postnatal}}{}^{\Phi}]^{1/\Phi}$$

By allowing for varying complementarities between investments in different periods, the model is able to generate a number of rich theoretical predictions. If fetal and childhood health are complements, for example, this underscores the persistent importance of a 'good start,' as opposed to the 'fade out' implication of the Grossman model. This would occur, for example, if healthier newborns benefit more from breastfeeding or other nutrition. An extreme version of this technology includes perfect complementarity, whereby investments made *in utero* restrict the maximum level of lifetime capacity. Further, by allowing different dimensions of capacity to affect the productivity of investment, cross-capacity complementarities can shape investment decisions. For example, one might expect good childhood health to facilitate the development of cognition.

Empirical Evidence

Empirical evidence shows that investments in early childhood explain much of the variation in adult health. An intuition is that if early investments are especially effective and have had a longer time to feed through the dynamic system, their effect might be especially persistent. That said, it may be useful to distinguish conceptually between an early-life health shock and responsive investments: actions made in response to health shocks. What is observed in adulthood combines the effect of the shock and the responsive investments, should they exist. For example, there may be individual or institutional responses to health shocks, such as government aid following an earthquake. Importantly, families may provide investments that either remediate or reinforce shocks experienced *in utero*. Hence, when examining longer-term outcomes, it is important to keep in mind that these can represent both biological and social factors.

The 1918 Influenza

An influential study in the field of fetal origins research is Douglas Almond's paper on the 1918 Influenza Pandemic. Almond, a Professor of Economics at Columbia University, linked *in utero* exposure to the Influenza Pandemic to deteriorations in human capital accumulation and labor market activity decades later. Like the Dutch 'Hunger Winter' associated with the Nazi occupation of the Netherlands the Influenza Pandemic was sudden, short, unexpected, and widespread, providing an appealing research design.

Almond used data from the US Census, which record quarter of birth in some decades, to identify which infants were exposed to the flu. Although the Census does not tell us which mothers were infected, the flu was widespread enough that roughly one-third of infants born in early 1919 had mothers who contracted influenza while pregnant. As a control group, those born in early 1918 had essentially zero prenatal exposure to the 1918 pandemic. Figure 3 shows the high school graduation rates by birth year as recorded in the 1970 Census.

Further, Almond also used variation across US states in the severity of the pandemic to construct a second, difference-in-differences estimate of the pandemic's effect. Both econometric approaches yield large estimates of long-term effects. Despite the brevity of the health shock, children of infected mothers were approximately 20% more likely to be disabled and experienced wage decreases of 5% or more, as well as reduced educational attainment. These results have now been

Figure 3 High school graduation rates by birth year as recorded in the 1970 US Census. Reproduced with permission from Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of *in utero* influenza exposure in the post-1940 US population. *Journal of Political Economy* **114**(4), 672–712.

replicated using data from other countries including Great Britain, Brazil, and Taiwan.

Identification

The fact that the fetal origins hypothesis applies to a well-defined developmental period means that it lends itself well to testing. In particular, the hypothesis predicts that later-life health outcomes should be worse only for those cohorts whose pregnancies overlapped with the shock. This means that economists can compare outcomes among these affected cohorts against two other cohorts: the cohort that was about to be conceived when the shock occurred (and is therefore too young to be affected) and the cohort that was already born at the time of the shock (and is therefore too old to be affected prenatally).

Still, seeking to quantify *in utero* effects through such comparisons gives rise to several problems. First, most birth cohorts are neither exposed to an identifiable shock *in utero*, nor were born just before or just after such a shock (and thus cannot serve as good controls). Rather than looking at all the data on births, the researcher is immediately pushed to looking at particular episodes in which an identifiable shock occurred and then attempting to draw defensibly generalizable conclusions from these episodes.

Second, the ideal shock would be shorter than the length of gestation, so as to differentiate between fetal and early-childhood exposure and perhaps stages of gestation. Many important prenatal factors, however, may last longer than pregnancy or may indeed shift permanently (e.g., the beginning of the US Food Stamp Program during the 1960s). The effect of fetal exposure may still be identified but constitutes the additional effect on top of any early-childhood effects. In general, it can be more challenging to isolate the effect of 'early-childhood' exposure because it is both less well defined and longer than the prenatal period.

Finally, one needs to be able to link data on adult outcomes to data on the affected cohorts. Economists have been creative in linking large-sample cross-sectional datasets back to ecological conditions around the time of birth. Most often, they have used information on when and where a respondent was born to link that person back to *in utero* health conditions. This has enabled economists to consider historical events featuring relatively well-defined start and/or end points. But many prominent datasets, such as the Current Population Survey, do not include information on where someone was born or the precise date of birth. As a result, many interesting and policy-relevant experiments linked to a certain time and place may never be analyzed.

In the next section, the empirical evidence in the context of these issues will be discussed.

Evidence from Sudden Shocks

A number of studies use sudden shocks like the 1918 Influenza to study the fetal origins hypothesis. These types of episodes often provide clean identification through sharp timing and, if far enough in the past, allow the researcher to examine outcomes over the full lifecourse, including mortality. A drawback is that predictions associated with large-scale or historical events may be difficult to generalize.

Large-scale shocks that have been studied in association with fetal origins include: a prenatal iodine supplementation program rolled out across Tanzania in the 1980s (by Field and colleagues), radioactive fallout from Chernobyl (by Almond and colleagues), and ambient temperature and rainfall shocks during pregnancy (by Maccini and Yang and colleagues). Outcomes examined include many different measures of health and human capital.

Identification in these studies is often based heavily on birth timing vis-à-vis the shock. Where possible, robustness is assessed by comparing effects within a certain time period across locations that experienced differing severities of the shock. Thus, the researcher is able to control for seasonal events that might coincide with the timing of the shock. Further, some datasets include a sibling link, allowing the researcher to control for fixed characteristics of families, including selective uptake of the treatment, though it is of course possible for parents to treat some siblings differently than others.

The studies referenced above produced a number of interesting findings. For example, the study on iodine supplementation found large and robust educational impacts – on average approximately half a year of schooling, with larger improvements for girls. Health measures, in contrast, appeared to be unaffected by this intervention. Subsequent work by Adhvaryu and Nyshadham has considered whether postnatal investments made by parents seem to respond to the iodine supplementation program, finding that parents reinforce iodine-related cognitive increases. Similarly, Chernobyl radiation in Sweden seems to have had its largest impact on human capital formation, not on health per se, suggesting the possibility of parental response to health endowment at birth.

Longer Natural Experiments

Many potential pathogens are more persistent than the shocks considered in Section Evidence from Sudden Shocks. Recent research has sought to maintain identification while considering slower-moving experiments, for example, to ambient pollution levels. Empirical evidence shows that these insults often have large effects on fetal health. Such findings are of particular interest because these exposures are often more common and generalizeable than with sudden shocks. A case in point is to consider the impact of slower-moving climate change as opposed to weather shocks, where adaptations and responses may differ.

As before, studies have also considered longer-term changes in the infectious disease burden. Infections can affect fetal health by diverting maternal energy toward fighting infection, by restricting food intake, or through negative consequences arising from the body's own inflammatory response. These studies have exploited variation in infectious disease in the US across seasons and states, including policy-related improvements in malaria in South US. Results show that reductions in infectious disease *in utero* lead to improvements in mortality and schooling later in life. For example, estimates show that early-life malaria can account for a quarter of the difference in long-term educational

attainment between cohorts born in malaria-afflicted states and non-afflicted areas in the early twentieth-century US.

There is evidence that some milder health shocks such as relatively low-level exposures to every-day contaminants as automobile exhaust and cigarette smoke also have negative effects on fetal health (see studies by Janet Currie, Michael Greenstone, Kenneth Chay, and others). Yet there has been little research to date linking fetal exposures to future outcomes. An exception is a study by Saunders that links the US recession of the early 1980s to reduced pollution and, through increased fetal health, improvements in high school test scores years later. Pollution levels experienced by these cohorts were high when compared to today but low when compared to many developing countries, such as China.

Studies found that being *in utero* during the annual Ramadan fast is associated with a broad spectrum of damage later in life, both to health and human capital. Daytime fasts that fall during early pregnancy have been found to have particularly large effects, despite being relatively mild when compared to famine events previously analyzed. This effect may arise because some pregnant women may fast without knowing they are pregnant.

Finally, a number of recent papers consider the effects of aggregate economic conditions around the time of birth on fetal health. Here, health in adulthood tends to be the focus (rather than human capital), and findings are less consistent than in the studies of nutrition and infection described above. One problem may be that the shocks are more diffuse in terms of timing so comparisons are less sharp. (A notable exception considers the effect of income shocks from crop blight across France.) A second issue is that the mechanism is less clear as economic downturns may affect fetal health through multiple pathways including effects on nutrition, smoking, and stress. Research by Van Den Berg and colleagues found that those born during economic downturns in the Netherlands had shorter lives, whereas a study by Cutler and colleagues on cohorts born during the Dust Bowl era in the US did not find any long-term effects.

Further Issues: Measurements of Fetal Health

All of the previously discussed studies show this maternal health shocks can be transmitted to the fetus. The most commonly used measure of fetal health is birth weight, but it may not be a particularly comprehensive or sensitive measure. In studies of the Dutch famine, for example, cohorts who were exposed to famine during the first half of pregnancy were found to have relatively normal birth weight but later showed evidence of health effects such as incipient heart disease.

Birth weight is, however, the most widely available measure of fetal health and there has been no convergence on an alternative, superior measure. An ideal metric would be sensitive for (even latent) fetal insults at all stages of pregnancy, be easy to measure, and be available for all mothers (or at least a large sample of mothers) in a cohort at the time of birth. Finding this measure of health at birth would obviate the need for data on later-life outcomes, enabling the researcher to examine current shocks rather than having to focus on those far in the past.

The lack of an ideal measure of fetal health has not, however, prevented economists from addressing the fetal origins hypothesis. This may be because economists are accustomed to considering many variables to be latent – like the potential wages of non-workers. On a practical level, economists' focus on identification strategies enables them in many circumstances to sidestep the question of finding a better measure of fetal health.

Further Issues: Bias from Selective Prenatal Mortality

A final issue to beconsidered is that of fetal mortality. Depending on the severity of a given shock, it may be that some fetuses die in response. Given that this type of selective mortality is unobserved in most birth data, researchers may underestimate fetal health shocks if the fetuses with higher baseline health are the ones that survive (but are 'scarred'). This becomes a serious problem if the negative scarring effects are sufficiently strong among the survivors to overwhelm the positive effects of selection.

Although this issue has been acknowledged outside of economics, economists have contributed by devising ways to model unobserved fetal mortality somewhat more formally. Such an exercise can be used to help quantify the selective effect due to mortality, and thereby isolate the 'scarring' effect of prenatal health conditions.

Conclusion

This article has summarized the current state of economics research on the fetal origins hypothesis. This hypothesis states that many important adult health and labor market outcomes may originate with fetal health conditions. Leveraging large-scale datasets and the sharp predictions associated with *in utero* exposure, economists have confirmed the link between fetal health and later-life outcomes. These results may hold true not only for large shocks but also for relatively mild and common shocks, such as reductions from already relatively low levels of air pollution and seasonal infections. Understanding the exact propagation mechanisms and how best to design remedial policies remain important research areas.

Acknowledgement

Almond was supported by NSF CAREER award #0847329.

See also: Education and Health. Intergenerational Effects on Health – *In Utero* and Early Life. Macroeconomy and Health. Nutrition, Economics of. Nutrition, Health, and Economic Performance. Smoking, Economics of

Further Reading

Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 US population. *Journal of Political Economy* **114**(4), 672–712.

Almond, D. and Currie, J. (2011a), Human capital development before age five. In Ashenfelter, O. and Card, D. (eds.) *Handbook of labor economics*, ch. 15, vol. 4b, pp 1315–1486. North Holland: Elsevier.

Almond, D. and Currie, J. (2011b). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives* **25**(3), 153–172.

Barker, D. J. (1990). The fetal and infant origins of adult disease. *BMJ* **301**(6761), 1111.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics* **122**(1), 409–439.

Chay, K. Y. and Michael, G. (2003). The impact of air pollution on infant mortality: Evidence from the geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* **118**(3), 1121–1167.

Currie, J. and Rosemary, H. (1999). Is the impact of shocks cushioned by socioeconomic status? The case of low birth weight. *American Economic Review* **89**(2), 245–250.

Currie, J., Stabile, M., Manivong, P. and Roos, L. L. (2010). Child health and young adult outcomes. *Journal of Human Resources* **45**(3), 517–548.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy* **80**(2), 223–255.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *PNAS: Proceedings of the National Academy of Sciences* **104**(33), 13250–13255.

Kermack, W. O., McKendrick, A. G. and McKinlay, P. L. (1934). Death-rates in Great Britain and Sweden: Some general regularities and their significance. *Lancet* **31**, 698–703.

Relevant Websites

http://users.nber.org/∼almond/
Douglas Almond's web page at the National Bureau of Economic Research.

http://www.jenni.uchicago.edu/
James Heckman's web page at Chicago.

http://www.princeton.edu/∼jcurrie/
Janet Currie's web page at Princeton.

http://www.thebarkertheory.org/
The Barker Theory.

Global Health Initiatives and Financing for Health

N Spicer, London School of Hygiene and Tropical Medicine, London, UK
A Harmer, University of Edinburgh, Edinburgh, UK

Glossary

Aid architecture It refers to actors, institutions, systems and approaches at the international and country level that are concerned with the transfer of financial, technical, and human resources from donors to recipient countries.

Alignment A term used when donors design their development priorities and programs to be consistent with those of a recipient country, for example by using country procedures and institutions rather than those that are externally introduced.

Bilateral donor It refers to an agency that manages the transfer of aid from one country to another.

Country ownership It refers to recipient country leadership in development priorities and programs. An absence of country ownership suggests limited capacity in the government of the recipient country or overly prescriptive donor programs.

Fungibility A term used to describe the substitutability of one entity for another. For example, (1) money is fungible, in that a ten dollar bill is equivalent to ten one dollar bills, (2) In aid policy, the phenomenon of external funding intended for one purpose but ultimately used by a recipient government for another.

General budget support The money given directly to a recipient country government, generally to the ministry of finance or equivalent that is channelled into the general public spending budget.

Global health initiative (GHIs) The international initiatives for raising and disbursing additional financing for infectious diseases such as human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) tuberculosis and malaria, and for immunization and strengthening the health systems of low- and middle income countries. GHIs share a common set of functions: to finance, resource, coordinate, and/or implement disease control globally.

Harmonization An attempt in making uniform or mutually consistent the rules and other arrangements of different jurisdictions or international organizations and initiatives. It may refer to financial, organizational or procedural arrangements, including aid programmes and global health initiatives.

Health systems strengthening The procedures for supporting a country's health care system through means such as leadership training, principles of good governance, quality assurance in service delivery, affordable financing arrangements, investment in evaluative skills and in other human and clinical resources for health.

Multilateral agencies It refers to agencies representing multiple countries working together on a given issue which includes the United Nations, the World Health Organization, the World Bank, and the World Trade Organization.

Vertical programs The programs that tackle one or few diseases or health issues; an approach often contrasted with horizontal programs which tackle multiple diseases or health issues – usually at the primary healthcare level.

Introduction

Recent years have seen important shifts in global development assistance for health (DAH). Global health initiatives (GHIs) – consisting of bilateral donor and multilateral programs, and global public–private partnerships – have mobilized significant new financing for health programs, and equate to a considerable proportion of overall overseas development aid (ODA) for health in many low- and middle-income countries (LMICs). This has enabled a dramatic scaling up of health interventions, especially for HIV/AIDS. GHIs emerged from shifts in thinking about DAH in the 1990s/early 2000s, which was hitherto characterized by donor-prescribed projects and programs financed principally by bilateral donors and the World Bank. The shift in policy focus from international to global health, an increasing number of global financial actors, and the pressing need to meet persistent and newly emerging global health threats meant that a new response was required to coordinate global efforts to raise more money for health. Although GHIs share a common set of functions: to finance, resource, coordinate and/or implement disease control globally, the term global health initiative encompasses a range of financing and implementing entities (bilateral and multilateral actors, and global public–private partnerships) with diverse governance arrangements and programmatic foci.

In this article the focus is on four of the largest GHIs: the Global Fund to Fight AIDS, Tuberculosis and Malaria (Global Fund), the Global Alliance for Vaccines and Immunizations (GAVI Alliance), the President's Emergency Plan for AIDS Relief (PEPFAR), and the World Bank's Multi-country AIDS Program (MAP, which ceased financing in 2008). **Table 1** summarizes the main features of these GHIs and other key initiatives and partnerships. The discussion is restricted to these four GHIs primarily because evidence is now beginning to emerge from empirical studies of their effects on country health systems – particularly the Global Fund; also because there is fairly limited data beyond these four large initiatives. GAVI, launched in 1999, was the first of the GHIs to disburse substantial funds at the global level, shortly followed by the MAP. High expectations surrounded the launch of the Global

Table 1 Examples of major global health initiatives

Global health initiative	*Institutional type*	*Date established*	*Total financing*	*Disease/health issue focus*
GAVI Alliance	Public–private partnership	1999	$4.5 B by 2009	Immunizations, prioritizing pneumococcal and rotavirus
Global Fund to Fight AIDS, Tuberculosis and Malaria	Public–private partnership	2002	$18.1 B by 2010	HIV/AIDS, tuberculosis and malaria
Multi-country AIDS Program (World Bank)	Multilateral program	2000	$3.1 B (total World Bank financing for HIV/AIDS programs 1989–2009)	HIV/AIDS
President's Emergency Plan for AIDS Relief (PEPFAR)	Bilateral program	2003	$19 B 2004–08	HIV/AIDS
Stop Tuberculosis Partnership	Public–private partnership	2001	Secretariat has received $396 M (2001–09) in cash contributions	Tuberculosis
Roll Back Malaria	Public–private partnership	1998	Not available	Malaria
Global Alliance for Improved Nutrition (GAIN)	Public–private partnership	2002	Total donations (2003–10) $133 M	Malnutrition
International AIDS Vaccine Initiative (IAVI)	Public–private partnership	1996	Revenue for the period 2006–09 $354 M	Vaccines to prevent HIV infection and AIDS

Fund in 2002: the initiative aimed to raise consciousness about important health issues, attract new partners, leverage substantial new funds, benefit from economies of scale in drug procurement, and promote coordination through pooling funds. There was, however, some reversal of the multilateral models of GAVI, MAP, and the Global Fund when PEPFAR was launched by the Bush Administration in 2003, a move criticized for operating in parallel to other actors and initiatives, and for adopting a prescriptive approach to determining the content of HIV/AIDS programs.

Reflecting the experimental nature of these new financing mechanisms and their sheer size, decision makers are inevitably curious about what impacts – both positive and negative – they have on recipient countries. There is an emerging literature on the effects of global initiatives and partnerships – most of which focuses on the largest HIV/AIDS initiatives – the Global Fund, PEPFAR and World Bank MAP, although there are also several studies on the GAVI Alliance. In this paper current knowledge on GHIs is reviewed, focusing on issues of healthcare financing. The achievements are reflected on, and also on the real and potential challenges that these initiatives create or reveal.

To What Extent Have Global Health Initiatives Increased Health Financing?

At the beginning of the 1990s, DAH was $5.7 billion. By the end of the decade, it had risen to just under $10 billion. A decade into the new century and DAH is pushing $25 billion annually, an increase of 124% in ten years. A 2010 report published by the Institute for Health Metrics and Evaluation discerns shifts in the balance of financial contributions to global health from traditional multilateral funders to GHIs. However, since 2007–08 when growth in DAH reached a peak of 17.5%, the rate of funding has been slowing down. In 2008–09 it dropped to just 6%. The proportion of bilateral funding has increased from 30% in 2001 to 45% in 2010, boosted by PEPFAR. So too has the proportion of funding from the Global Fund: from just 1% in its inaugural year to 11% by 2010. During the same period, UN agencies' contributions have shrunk sharply from 24% in 2001 to 14% in 2010. The World Bank's contribution has seen a dramatic reduction from 17% of total DAH in 2001 to just 5% in 2010.

For disease-specific health interventions, the Global Fund has punched well above its weight, and funding for HIV/AIDS, tuberculosis and malaria has increased dramatically. In 2009, this GHI disbursed just over $1.35 billion for these diseases. Financing HIV/AIDS, tuberculosis and malaria inevitably benefits maternal, neonatal, and child health (MNCH), and in this respect the Global Fund and GAVI have also contributed sizeable sums. In sub-Saharan Africa, for example, HIV/AIDS, tuberculosis and malaria are responsible for 52% of deaths among women of childbearing age and malaria alone accounts for 16–18% of child deaths. As funding for HIV/AIDS and other infectious diseases increased, funding for health systems and populations has experienced a corresponding decline. Between 1992 and 2003, funding for HIV/AIDS increased from 8% to a third of all commitments; during the same period, aid for population health experienced precisely opposite fortunes, decreasing from 32% to 8% of donor aid.

Does this shift mean that financing for specific diseases is displacing – or 'crowding out' – much-needed funding for other health priorities such as health system strengthening or non-communicable diseases; or conversely, has increased financing for specific diseases had a knock-on effect and increased funding for other health priorities? In terms of displacement of funds, there are multiple trends that indicate possible HIV/AIDS displacement effects, such as an increasing share of donor health and population funds. But there are also indications that HIV/AIDS funding is raising other health funding levels, particularly for control of other infectious diseases, though not for non-communicable diseases. At the

same time that funding from global health financing partnerships is increasing, a widening mismatch between ODA and health need is becoming apparent, with high visibility global health problems and measurability of outputs being major drivers of funding.

Neither is it clear whether additional funding for health has been used in the manner intended by funders – namely on the health sector. The term used to describe the phenomenon of funding intended for a specific purpose but ultimately used for another is fungibility, and this is typically used when governments receiving donor funding reduce their own spending on the same health issue and therefore aid substitutes rather than increases local funding. Evidence whether financing from global initiatives and partnerships results in governments reallocating funds to other health areas, or indeed to non-health programs is inconclusive: in some cases domestic finances stay the same or decreases, in other domestic financing increases. For example, in Ghana there is no evidence that Global Fund support had led to deductions in government or other donor financing, whereas in Tanzania receipts of external financing for HIV/AIDS and tuberculosis had led the government to reallocate resources away from the health sector.

GHIs and Innovative Financing

To achieve the Millennium Development Goals, developing countries will have to spend approximately $60 per capita by 2015, or 100% more than they are currently spending. It is unrealistic for many countries to achieve this increase. In 2001 members of the Organization of African Unity (OAU) met at Abuja, Nigeria. The resulting 'Abuja Declaration' committed all members of the OAU to ensure that at least 15% of the domestically financed government expenditure went to health. Even if low-income countries were able to meet their Abuja commitments and divert 15% of government budget to health very few of them would generate enough funds to meet the $34 per capita threshold that the Commission on Macroeconomics and Health in 2001 deemed sufficient to meet basic health needs. Admittedly, this $34 has now appreciated to approximately $50, and some countries would not achieve that target even if 100% of the government budget was diverted to health. DAH from multiple donors, including GHIs, will go some way towards filling this gap, but in addition, GHIs – particularly GAVI and the Global Fund- have championed innovative mechanisms for raising even more funds. Tasked with the challenge of identifying a range of innovative ways to raise money for health systems, a Taskforce for Innovative International Financing (TIIF) was established up in 2008 through the auspices of the International Health Partnership. It identified a tax on airline tickets, a currency transaction levy, and levies on other products and services such as mobile phone use, amongst other innovative ideas (Table 2). If brought to fruition, these mechanisms could increase ODA by $10 B. Through these innovations GHIs are proving to be essential vectors for new ways of raising much-needed money.

These issues are discussed in the 2010 World Health Report which notes, if donors honored their international pledges, external funding would double and there would be no need for innovation (http://www.who.int/whr/2010/en/index.html).

Is Financing from Global Health Initiatives Predictable?

In September 2008, development agencies met in Accra for the Third High Level Forum on Aid Effectiveness. Here, there were promises to increase the predictability of aid to enable developing countries to effectively plan and manage their short- and medium-term development programs. Unpredictable aid makes it difficult for countries receiving financial assistance to budget and implement their development agenda efficiently. Indeed, lack of predictability can shave off substantial value of aid and is believed to be one of the biggest constraints on its effectiveness. There is a fundamental missmatch between medium to long-term development strategies of recipient country governments (which would often include employing more doctors and nurses), and many donors, including GHIs, relatively short-term funding commitments. Typically, donors only commit aid 12 months in advance, and levels of aid can vary greatly from year to year. This weak alignment runs counter to funders' stated commitment to country ownership, undermining governments' authority to manage their health development programs.

For full details of Accra for the Third High Level Forum see: http://www.oecd.org/dac/aideffectiveness/thirdhighlevelforumonaideffectiveness2–4september2008.htm

A further problem is that unpredictable aid can increase fiscal and monetary instability, which in turn can lead to inflation and macroeconomic disruption. Ensuring macroeconomic stability is the raison d'être of the International Monetary Fund (IMF), an international financial institution that lends money to ailing economies. IMF loans typically come with a set of economic conditions – such as raising interest rates – derived from a set of economic principles sometimes referred to as the 'Washington consensus.' One controversial principle is the insistence on low – single-digit – inflation. The twin goals of raising interest rates and disinflation come with a high 'sacrifice ratio' (the amount of GDP growth a government 'sacrifices' to achieve the prescribed low level of inflation). As a country's economy cools, negative consequences for health become apparent from resulting cuts in health spending and wage ceilings for health workers. Early experiences from countries in sub-Saharan Africa revealed tensions between IMF loan conditions and GHI funding for health. In Uganda, disbursement of a large tranche of Global Fund money ($201 M) was delayed in 2002 because of concerns by the Ugandan finance Minister – on the advice of the IMF – that receiving such large amounts of 'additional' funds would increase the value of the Ugandan currency and render its economy less competitive. In Kenya, the heath workforce was reduced by over 30% during the 1990s in response to IMF loan conditions, and was only able to use Global Fund and other global health initiative financing to hire new nurses after intense pressure from international nongovernmental organizations and strong leadership from the Kenyan Ministry of Health.

Do GHIs Commit Aid More Predictably Than Bilateral Donors?

It is suggested that GHI funding commitments are generally more predictable than bilateral commitments. Indeed, GAVI's

Table 2 Innovative financing mechanisms championed through GHIs

Financing innovation	*GHI*	*Established*	*Funding source*	*Amount raised*	*Type of innovation*	*Health issue*	*Aim*
UNITAID	Global Fund, Clinton Foundation	2006	Tax on airline tickets	$1 B	Drug purchase facility; market intervention	HIV/AIDS, malaria	Decrease the price of medicines for priority diseases; increase the supply of drugs and diagnostics
International Finance Facility for Immunization (IFFIm)	GAVI	2006	Bonds issued in capital markets	$3 B	Front-loading cash from long-term donor commitment	Vaccine development; immunization service utilization	To rapidly accelerate the availability and predictability of funds for immunization
Debt2Health	Global Fund	2007	Debt relief	Germany has cancelled €40 M debt with Pakistan and ¢50 M with Indonesia	First ever trilateral debt relief arrangement involving a multilateral organization	AIDS, tuberculosis and malaria	Using debt swaps to free up domestic resources for Global Fund approved programs
AMC	GAVI	2005	Front-loaded financing from donors	Cost to GAVI – according to Light (2011) initial claim of $180 M yr^{-1}, likely to cost $576 M yr^{-1}	Advanced market commitment from donors to purchase pre-agreed quantity of vaccine. But some argue it is a 'large volume surplus contract' not a true AMC (Light 2011)	Pneumonia	To stimulate the development and manufacture of vaccines for developing countries

third strategic commitment was to improve the predictability and sustainability of financing for national immunization programs. According to the OECD, the Global Fund had a predictability ratio of 82% (where 100% meant that a donor disbursed the same amount as it initially planned). However, disbursement is tied to a country's performance, and so this can have a negative effect on predictability of financing. In an effort to address problems associated with short-term funding cycles, the International Finance Facility for Immunization (IFFIm) was launched by the GAVI Alliance in 2006. The IFFIm is an innovative mechanism through which national donors raise money up-front by issuing bonds which are paid back over 23 years. So far IFFIm has raised more than US$3 billion for the GAVI Alliance's immunization programs. A total anticipated IFFIm disbursement of US$4 billion is expected to protect more than 500 million children through immunization (Table 2).

Some aid modalities are more suited to predictable funding than others. General budget support – aid channeled directly into the budget of a recipient country – is arguably more effective than other modalities as it avoids project-based inefficiencies and is easier to align with country priorities. It does, of course, run the risk of mismanagement of funds in countries with weak economic governance. Budget support is not without its own problems – some of which go against other measures of aid effectiveness such as country ownership including, it can be argued, that budget support allows donors direct access to country decision making. Although there are positive examples of PEPFAR disbursements in sub-Saharan Africa, including Mozambique, Uganda, and Zambia, others argue that PEPFAR has been less predictable than other donors. Although GAVI, the Global Fund and the World Bank have been able to secure multi-year replenishments, long-term pledges, and innovative financing arrangements to accumulate funds, other GHIs, such as PEPFAR, are constrained by legal restraints on their primary funders. Although the Global Fund has contributed to more predictable financing through its shift to general budget rather than program support, the premium the partnership places on performance has had an adverse effect on predictability. Indeed the Global Fund's requirements for frequent reporting were a major burden on recipients that caused delays in disbursement and resulted in the perception that its money is unpredictable. Indeed the Fund's temporary suspension of grants in Uganda had a negative effect on perceptions of predictability by recipients which led sub-recipients to favor PEPFAR funding that was seen as more quickly disbursed and predictable.

Detailed country case studies of PEPFAR and Global Fund financing flows can be found on the Center for Global Development's website (http://www.cgdev.org/).

Do GHIs Disburse Aid on Time or Are Delays or Interruptions Commonly Experienced?

Difficulties have been reported drawing down Global Fund and MAP finances because of problems created by certain countries' low absorptive capacity, and also because of performance-based funding conditions. In contrast, PEPFAR has disbursed finances more quickly since these finances are not routed through government implementers and do not rely on government systems. There are mixed experiences from different countries on timeliness of GHI funding. In Kenya, PEPFAR disbursements were reported as timely, whereas the Global Fund grant application process was lengthy and complex. In Haiti and the Central African Republic delays between Global Fund grant approval and disbursement were experienced. In the Central African Republic this stemmed from human resource constraints delaying the reporting required to trigger disbursements. The Global Fund delayed disbursements in Laos because of the country's weak financial monitoring and evaluation systems, and interruptions in Global Fund disbursements to nongovernmental sub-recipients in Kyrgyzstan were reported as a key reason for intermittent HIV/AIDS service delivery.

Are Global Health Initiatives Financing Sustainable Health Programs?

GHIs have aimed to provide short- to medium-term finances with the intention of stimulating increases in longer term financing for health programs from country governments or other domestic sources. However, in countries with high levels of external financing from GHI vertical programs serious concerns have been raised about increasing aid dependency, while few or no strategies are in place for longer term financing.

Country evidence is thin on whether they are stepping up domestic financing in parallel with GHI financing leading to sustainable programs. In some countries such as Ethiopia, Mozambique, Uganda and Zambia GHI financing is linked to reductions in domestic financing for focal diseases programs, and in Haiti – which received substantial PEPFAR and Global Fund financing – it is expected that when these grants finish focal health programs will need to end. The problem is not confined to low-income countries. A study in the middle-income country of Georgia showed that scale-up of HIV/AIDS, tuberculosis and malaria programs financed by relatively modest Global Fund grants led to government diverting financial resources to non-focal disease healthcare priorities. At the same time rising recurrent cost requirements in focal service areas aggravated the potential for longer term funding shortages with the government unlikely to be in a position to replace GHI financing.

The Global HIV/AIDS Initiatives Network website provides extensive resources on the country effects of GHIs including county case studies and a searchable database of research-based evidence (http://www.ghinet.org/).

GHIs are increasingly seeking to make investments in longer term health systems strengthening (HSS) interventions, thereby creating a more tangible legacy of their programs. For example, PEPFAR invested US$ 640 million to systems strengthening work including health worker training in 2007. Global Fund financing has supported a range of HSS strategies including those relating to strengthening human resources for health and has expanded support of HSS in Global Fund applications. However, the imperative of the initiative to rapidly disburse finances and demonstrate their impacts is reflected in the tendency for programs to place most attention on in-service

training, task shifting and expanding the numbers of lower cadre workers, and in some countries on the recruitment of nongovernmental workers on short-term contracts, rather than training and recruiting new highly skilled health workers. In those countries nongovernmental organizations acting as implementers of HIV/AIDS Global Fund financed HIV/AIDS programs were reported as heavily dependent on Global Fund financial support, which jeopardized their long-term existence.

Before 2010, it seemed as though the Global Fund was in a strong position to continue to fund countries' health needs. However, in 2010 cracks began to appear in the strength of donor support for the Fund. Donors committed far less to the Fund for the period 2011–13 than was expected or hoped for. However, towards the end of 2011 in the following year, the Global Fund announced that it had insufficient funds to finance any new projects until 2014. This was a catastrophe for the Fund. In mid-May 2012 the Global Fund was able to release $1.6 billion to spend on new projects – still far less than was anticipated. The future of the Fund is now uncertain, although under the new leadership of Mark Dybul confidence may be returning. Despite its dramatic reversal of fortune, it is nevertheless true that before 2010, the Fund had generated massive scale-up of new funds. These have had undeniably positive effects on health.

Is Financing from Global Health Initiatives Harmonized and Aligned?

The proliferation of global health actors, including new GHIs, has heightened concerns about the lack of harmonization of health programs, and poor alignment between GHI programs and country priorities, systems and procedures. This concern is central to the aid effectiveness agenda that recognizes that while substantial new resources are being mobilized for focal health issues and disease areas, this aid may not be used as effectively as it might. Indeed, GHI funding may have some damaging effects on recipient countries with fragile health systems. The principles articulated in the Paris Declaration on Aid Effectiveness and the Accra Agenda for Action have sought to galvanize global commitments to improve the ways aid was disbursed by ensuring that aid is better harmonized and aligned, more predictable, based on country ownership and demonstrate greater mutual accountability; an agenda embodied in the health sector with the launch of the International Health Partnership in 2007. This raises the question – have GHIs stepped up to the expectations of Paris and Accra?

GHIs have embraced a disease-specific 'vertical' financing approach to target particular health issues, in part because this enables donors to demonstrate a link between their financial inputs and impacts. In this context many commentators agree that the expectation that GHIs would simplify aid architecture has not been achieved. Country experiences reveal misalignment between predominantly vertical GHI programs and country priorities and/or country disease burdens. Duplication and lack of coordination have inevitably stemmed from the introduction of parallel initiatives and donor programs, although experiences vary between initiative and recipient country and have improved over time. For example, Global Fund, PEPFAR and World Bank MAP programs each adopted different procedures for procurement and disbursement of drug supplies, and the Global Fund's requirement for a country coordinating mechanism (CCM) differed from the requirement of the World Bank and was perceived as a Global Fund rather than country-owner structure.

It is widely accepted that high transaction costs of the Global Fund and PEPFAR, and indeed other donors, impose different reporting procedures that place substantial demands on fragile country health systems, including institutional capacities and staff. PEPFAR's imposition of rigid budget allocations to prescribed interventions had undermined the initiative's commitment to country ownership, lack of transparency and lack of willingness to coordinate with other donors. Global Fund programs were reported as not engaging with pre-existing country coordination structures such as SWAps, and this reinforced vertical tendencies against government priorities to integrate health interventions at the primary healthcare level, as experienced in Georgia. In other countries in Africa, Global Fund, GAVI and PEPFAR financing is believed to be driven by global agendas that gave recipients limited flexibility to allocate finances according to their own priorities.

Nevertheless there are improvements in some countries: the Global Fund has fared well in terms of use of country procurement systems, improving alignment between Global Fund programs and national priorities and having greater country ownership than other donors, although less well in terms of alignment with national M&E systems and country cycles. Country studies reveal improved alignment between Global Fund programs and health reforms in Benin and Ethiopia, and engagement in SWAps in Mozambique and Malawi. In Rwanda, Global Fund financing allowed greater country ownership of focal disease programs than other external financing; the CCM had enabled country actors to make resource allocation decisions that were in line with country priorities. There is also some evidence that PEPFAR's programs were becoming better aligned with national plans over time.

To What Extent is GHI Financing Transparent?

Although the Five-Year Evaluation of the Global Fund suggests that Performance-Based Funding (PBF) has contributed to a culture of accountability, it also accepts that the approach has 'evolved into a complex and burdensome system,' and there remain weak monitoring, evaluation and information systems limiting the PBF approach. Similarly in the Central African Republic and Rwanda the introduction of PBF by initiatives including the Global Fund and GAVI served to improve performance, transparency and management thereby fostering accountability and reducing waste. In Haiti, PEPFAR and Global Fund financing made grantees more efficient, accountable and strengthened administrative and managerial capacity, as had Global Fund financing in Ukraine and Kyrgyzstan, although in Kenya and Kyrgyzstan performance-based monitoring had delayed grant disbursement. PEPFAR funding practices were reported as lacking transparency in some countries such as Rwanda.

The Global Fund launched a major review of its progress in 2006, known as the Global Fund Five-Year Evaluation. This multi-country assessment of the health impacts of the Global

Fund, including health systems effects can be found at: http://www.theglobalfund.org/en/terg/evaluations/5year/

Conclusion

Global health initiatives have raised and disbursed substantial new financing for major diseases and health issues. Although there are clear benefits of this increased financing in terms of significant programmatic scale-up, GHIs have revealed and in some cases aggravated weaknesses within fragile health systems. Particular concerns remain about the longer term legacies of these initiatives on the countries they aim to benefit. There are multiple problems: first, the global financial crisis puts at risk donors' commitments to make longer term financial pledges to GHI programs, threatening to undo the important gains so far; second, the ability of initiatives to strengthen country health systems in the longer term has been limited by their vertical, disease-specific nature; and third, recipient governments' strategies to scale up domestic financing to supplement or replace external health support have been restricted by international loan conditions that indirectly restrict domestic spending on health, thereby jeopardizing the sustainability of focal disease programs beyond the life of current GHI financing.

Generating evidence on the effects of GHIs is not without methodological problems: GHIs and other donors have financed complex, multi-level country programs making it difficult to attribute the effects of a single initiative or program and findings are often context-specific and quickly out of date in the context of evolving, multiple financing streams. Considerable evidence is derived from mixed quantitative–qualitative studies and the synthesis of cross-country qualitative evidence, approaches that are not as universally accepted as traditional quantitative study designs. GHIs have introduced new models of financing major health programs, yet the contrast between different models – global public–private partnerships in the form of the Global Fund and GAVI Alliance, the multilateral World Bank MAP and the bilateral PEPFAR initiative – reflects what is very little global consensus about which financing models are best. Nevertheless all four initiatives have demonstrated their willingness to learn from and respond to emerging evidence, and a number of promising 'course corrections' over their relatively short lives have been apparent.

The global health arena is a dynamic one and GHIs have become pivotal actors. There are discussions on establishing a joint GAVI Alliance, World Bank and Global Fund Funding Platform for HSS, and there have been calls to amalgamate major GHIs programs to form a Global Health Fund to coordinate global funding for broader health programs. Evidence will be needed to capture and assess the impacts of these and other changes, as GHIs evolve and effects of the global financial crisis become fully apparent.

See also: Development Assistance in Health, Economics of. HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. International Movement of Capital in Health Services

Further Reading

Biesma, R., Brugha, R., Harmer, A., et al. (2009). The effects of global health initiatives on country health systems: A review of the evidence from HIV/AIDS control. *Health Policy and Planning* **2009**, 1–14.

Brugha, R. (2008). Global health initiatives and public health policy. In Heggenhougen, K. and Quah, S. (eds.) *International encyclopedia of public health*, vol. 3, pp. 72–81. San Diego: Academic Press.

Dodd, R. and Lane, C. (2010). Improving the long-term sustainability of health aid: Are global health partnerships leading the way? *Health Policy and Planning* **25**, 363–371.

Edstom, J. and MacGregor, H. (2010). The pipers call the tunes in global aid for AIDS: The global financial architecture for HIV funding as seen by local stakeholders in Kenya, Malawi and Malawi. *Global Health Governance* **IV**, 1.

Farag, M., Nandakumar, A., Wallack, S., Gaumer, G. and Hodgkin, D. (2006). Does funding from donors displace government spending for health in developing countries? *Health Affairs* **28**(4), 1045–1055.

IHME (2010). *Financing Global Health: Development Assistance and Country Spending in Economic Uncertainty.* Washington, DC: Institute of Health Metrics and Evaluation.

Maximizing Positive Synergies Academic Consortium (2009). *Interactions between Global Health Initiatives and Health Systems*. Geneva: WHO.

Ooms, G., Stuckler, D., Basu, S. and McKee, M. (2010). Financing the millennium development goals for health and beyond: Sustaining the 'big push'. *Globalisation and Health* **6**, 17.

Sepulveda, J., Carpenter, C., Curran, J., et al. (2007). *PEPFAR Implementation: Progress and Promise*. Washington, DC: The National Academies Press.

Shiffman, J. (2008). Has donor prioritisation displaced aid for other health issues? *Health Policy and Planning* **23**, 95–100.

Shridar, D. (2010). Seven challenges in international development assistance for health and ways forward. *Journal of Law, Medicine and Ethics Fall* **2010**, 2–12.

Stillman, K. and Bennett, S. (2005). *System-wide effects of the global fund: Interim findings from three country studies*. Bethesda, MD: The Partners for Health Reformplus, Abt Associates Inc.

Stuckler, D., Basu, S., Gilmore, A., et al. (2010). An evaluation of the International Monetary Fund's claims about public health. *International Journal of Health Services* **40**(2), 322–327.

World Health Organization Maximizing Positive Synergies Collaborative Group (2009). An assessment of interactions between global health initiatives and country health systems. *Lancet* **373**, 2137–2169.

Yu, D., Souteyrand, Y., Banda, M., Kaufman, J. and Perriëns, J. (2008). Investment in HIV/AIDS programs: does it help strengthen health systems in developing countries? *Globalization and Health* **4**, 8.

Global Public Goods and Health

R Smith, London School of Hygiene and Tropical Medicine, London, UK

Introduction

Public goods have, for centuries, been part of the economic analysis of government policy at the national level. This has included many goods associated with improving population health, such as water and sanitation. However, in an increasingly globalized world, health is an ever more international phenomenon. Each country's health affects, and is affected by, events and processes outside its own borders. The most obvious example of this is in communicable disease, where an outbreak such as sudden acute respirator syndrome (SARS) or pandemic influenza in one country very rapidly spreads and affects many others.

It is becoming clear in many areas that matters which were once confined to national policy are now issues of global impact and concern. This has been evidenced, for example, in dealing with environmental problems such as carbon emissions and global warming. These not only affect the nation involved in their production but also impact significantly on other nations; yet no one nation necessarily has the ability, or the incentive, to address the problem. Similarly, health improvement requires collective as well as individual action on an international as well as national level. Initiatives such as the Global Fund to Fight human immuno-deficiency virus (HIV)/aquired immune deficiency syndrome (AIDS), Tuberculosis, and Malaria reflect a growing awareness of this. However, initiating, organizing, and financing collective actions for health at the global level presents a challenge to existing international organizations. Recognition of this led initially to the development of the concept of Global Public Goods, and more recently the consideration of Global Public Goods for Health, as a framework for considering these issues of collective action at the international level.

What are Global Public Goods?

The global public good concept is an extension of the economic tradition of classifying goods and services according to where they stand along two axes – one measuring rivalry in consumption, the other measuring excludability – as illustrated in Table 1.

Pure private goods are those that are most used to dealing within day-to-day lives, and are defined as those goods (like a loaf of bread) that are diminished by use, and thus rival in consumption, and where individuals may be excluded from consuming them. At the opposite end of the spectrum are pure public goods, which are nonrival (not diminished by use) and nonexcludable (if the good is produced, it is freely available to all). For example, broadcast radio is nonrival (many can listen to it without preventing others from listening to it) and nonexcludable (it is difficult to exclude someone from receiving it). In between these extremes are 'impure' goods, such as 'club goods,' which have low rivalry but high excludability, and 'common pool goods,' which have low excludability but high rivalry. In these cases, exclusion may occur through geographic, monetary, or administrative prohibition, and some goods are rival relative to capacity (e.g., a sewage system with spare capacity is nonrival, but once at capacity, its use becomes rival).

One of the fundamentals of public economics is that the free market – the interplay of individual supply and demand decisions mediated through the price system – will result in the provision of less than the collectively optimal level of public goods. Thus, the nation state has a role to play either in producing the good directly (the traditional approach) or at least in arranging for its production by a private firm (the increasingly popular 'outsourcing' strategy).

Note that, importantly, a good need not be a pure public good to suffer from a collective action problem. Collective action problems also apply to private goods which have substantial positive externalities, as these too will be undersupplied (because externalities are not taken into account by private suppliers and consumers). For example, an individual secures only part of the benefit from his/her treatment for tuberculosis, as others benefit from the reduced risk of infection. However, it is only this private benefit that the individual will take into account when considering whether to

Table 1 Classification of goods by rivalry and excludability

Rivalry in consumption	*Excludability*		
	At negligible cost (high excludability)	*At moderate cost (moderate excludability)*	*At infinite cost (low excludability)*
No congestion (no rivalry)	(Impure) Public goods (e.g., books)	(Impure) Public goods (e.g., cable TV)	Pure public goods (e.g., clean air)
Congestion (moderate rivalry)	Club goods/local public goods (e.g., gyms)	Mixed public and club goods (e.g., toll road)	Common property resources (e.g., streets)
Infinite congestion (high rivalry)	Pure private goods (e.g., chocolate bars)	Natural resources, closed access (e.g., fish stocks)	Natural resources, open access (spring water)

 doi:10.1016/B978-0-12-375678-7.00623-4

seek treatment. Where the private benefit is less than the cost to the individual, they will not seek treatment, even though the population as a whole (including the individual sufferer) would be better off if the individual received treatment.

Thus, from a policy perspective it makes little sense to draw too categorical a distinction between private goods with large positive externalities and the pure public good case. In a sense, an intervention that would counter a nonpublic good-related collective action problem, so as to correct the under- or oversupply of positive or negative externalities, widely spread among the population, can itself be considered a public good. For example, providing infrastructure capable of delivering timely and effective treatment for tuberculosis, and the policies to provide an incentive for individuals to seek and complete treatment, may have the characteristics of public goods, even though the treatment of an individual is essentially a private good with positive externalities.

Turning to the global level, a reasonable functional definition of global public goods would be public goods that occur across a number of national boundaries, such that it is rational, from the perspective of a group of nations collectively, to produce for universal consumption, and for which it is irrational to exclude an individual nation from consuming, irrespective of whether that nation contributes to its financing. The key issue facing provision of these goods is how to ensure collective action in the absence of a 'global government' to directly finance and/or provide the public good.

For an interesting panel discussion of global public goods more generally, which includes the 2001 Nobel Prize winner for economics Joseph Stiglitz, http://www.youtube.com/watchv=2hmMWADaPJA

How do Global Public Goods Relate to Health?

As should be apparent from Table 1, 'health' itself is a private good, as are the majority of goods and services used to produce health. One person's (or one country's) health status is a private good in the sense that he/she (or it) is the primary beneficiary of it. To illustrate this, consider the parallel of a garden: if someone cultivates an attractive garden in front of his/her house, passersby will benefit from seeing it; but it remains a private good, the main beneficiary of which is the owner, who sees more of it and is able to spend time in it. An individual's health remains primarily of benefit to that individual, although there may be some (positive or negative) externalities resulting from it; typically exposure to communicable disease.

Further, in terms of the goods and services which are necessary to provide and sustain health, such as food, shelter, and use of curative health services, 'health' is often rival and excludable between individuals and nations. Nonetheless, there are two important externality aspects of health, both at the local level and across national borders, which may be amenable to conceptualizing as having global public good (GPG) properties: (1) prevention or containment of communicable disease and (2) wider economic externality effects (Box 1). However, there are several global public goods for health which are public goods yielding improvements in health globally. These include aspects of knowledge (and technology) production and dissemination, policy and regulatory regimes, and health systems (Box 2).

Box 1 Global public good aspects of health

Prevention or Containment of Communicable Disease

Preventing one person (country) from getting a communicable disease (or treating it successfully) not only benefits the individual concerned but also provides a benefit to others (countries) by reducing their risk of infection. Yet, although communicable disease control is nonrival in its effect (one person's lower risk of contracting a disease does not limit the benefits of that lower risk to others), its production requires excludable inputs, such as vaccination, clean water, or condoms, as well as nonexcludable inputs, such as knowledge of preventive interventions and best practice in treatment. In this sense, it is a 'club good' (nonrival but excludable), although its nonrival effect implies that even if it is feasible to exclude people, it may not be desirable as the marginal effects on the health of others may outweigh the marginal savings from exclusion. However, since not all communicable diseases are global, only the prevention or containment of some communicable diseases may be considered as global public goods: HIV/AIDS, tuberculosis, eradicable disease (e.g., polio), and antimicrobial resistant disease. Others such as malaria (a regional public good) or acute respiratory infection (population subgroups) are not global public goods.

Wider Economic Externality Effects

The economic effects of ill health on households may be considerable. Although these effects are essentially private, the cumulative effect on the economy of the resulting loss of production and income, and thus the potential gains from health improvements, may be substantial. For example, the difference in annual growth accounted for by life expectancy at birth between a typical developed and developing nation is approximately 1.6%. The close, mutual, relationship between poverty and disease – particularly communicable disease – has been recognized for generations. Not only does disease reduce the productivity and incomes of people and nations, as indicated, but also the resultant poverty impacts on health through its effects on nutrition, education, housing, and health care, creating a cycle of ill health and poverty which is hard to break.

The last of these may not be immediately apparent, as it is not a public good but what is termed an access good. These are private goods that are required such that a public good may be accessed. For instance, taking the example of broadcast radio earlier, to obtain this public good one requires a radio (which is excludable and rival) to access it. Thus, in many cases, public goods, such as disease eradication, require a minimum health system (e.g., access to vaccinations) to enable access to them (or, alternatively, to allow production of them). Such private goods may often be considered as if they were public goods to the extent that their provision is a vital element of provision of the public good itself.

Production and Finance

Clearly, global public goods need to be produced and financed, and the precise details of each of these will vary according to the specific issue at hand. For instance, production of disease eradication will require production to be locally

Box 2 Global public goods for health

The scope of potential global public goods that affect health is wide but can be broadly divided into those which address in-country health problems with cross-country externalities (primarily communicable disease control, but perhaps also noncommunicable disease control to the extent that it has economic effects) and those which address the cross-border transmission of factors influencing health risks (e.g., food safety, tobacco marketing, and international trade in narcotics). Within each of these categories, global public goods may then be classified into three broad areas.

1. *Knowledge and technologies*: Information per se, such as on health risks and treatment regimes, is a global public good. However, in practice, it may not be (e.g., control of communicable disease relies on countries to produce and to act on information, which requires an effective health infrastructure). Similarly, much of the technology for curative and preventive interventions is embodied in private goods such as pharmaceuticals and vaccines, and thus a club good.
2. *Policy and regulatory regimes*: The collective nature of policies makes them public goods. Regulatory regimes (e.g., for food and product safety or pharmaceuticals) are 'club goods,' as groups can be included or excluded by a regulation, but once a regulation exists, it can apply to one or many.
3. *Health systems, as an access good*: Many global public good aspects of health depend on the existence of a functioning health system and so they are so integral that they may thus be treated as if they were themselves global public goods.

based in the distribution and administration of vaccines, but may be financed through a variety of organizations and with different mechanisms, from local health services to Non-Governmental Organization (NGOs) to private companies, through gifting of vaccines, provision of local health service personnel, and international surveillance. In this respect, the example of polio eradication is provided in **Box 3**.

The core issue in provision and finance is that national public goods are dealt with by government intervention, through direct provision, taxes, subsidies, or regulation, but in the case of global public goods, the absence of a 'global government' means that the collective action problem becomes more complex with the increased number of players involved and the need for effective incentives for compliance. The main potential contributors to provision and/or finance are: (1) national governments; (2) international agencies (including philanthropic foundations and NGOs); and (3) commercial companies. However, these players' agendas (their preferences or priorities) do not necessarily coincide with each other. The more divergent these agendas are, the lower the chance of the good being produced. Impediments to international cooperation, and the role of international bodies in facilitating it, are, therefore, central to consideration of the provision and financing of GPG.

A significant constraint in global collective action is the ability of countries to pay according to the proportion of the benefits they receive from the good in question, as this undermines the political will to cooperate and limits effective participation. Even the creation of a legal duty does not ensure compliance, as this depends on having adequate resources to fulfill such obligations. Further, where countries with inadequate resources do participate in global programs, financial and human resources may be diverted from other essential activities, with possible adverse effects on health. The opportunity cost of these resources is far greater in developing than developed countries, creating tensions in securing global cooperation and reducing the net health benefits. Circumventing this problem requires that financial and other contributions reflect each country's ability to contribute, as well as its potential benefits. In practice, this means that financing needs to come predominantly from the developed world. However, it is important here to understand that this does not imply the use of overseas development assistance. Global public goods are not substitutes for aid but a complement to it: presenting an added rationale for international cooperation and assistance. Developed countries benefit from global public goods; yet because their provision is rooted in the national level, it is, therefore, in the self-interest of wealthy nations to assist poorer nations in contributing to the production of such goods. Thus, investment in poor countries is encouraged, not because they are poor per se, but to enable them to make their contribution to goods essential to developed countries.

The provision of global public goods depends on the ability to create arrangements that account for differing incentives and means of developed and developing countries. Thus, where developed countries have the incentives to produce the good and developing countries do not, but where the participation of the developing country is vital, developed countries will be required to fund the costs to developing countries of participating in production of the good. In contrast, when incentives exist for developing countries, but not for developed countries (where diseases are disproportionately incident in poor countries), developed countries might assist in providing incentives for the commercial sector ('push and pull' mechanisms such as subsidization for research, advance purchase commitments, and expansion of orphan drug laws) or facilitate market access. This brings us on to mechanisms for financing global public goods.

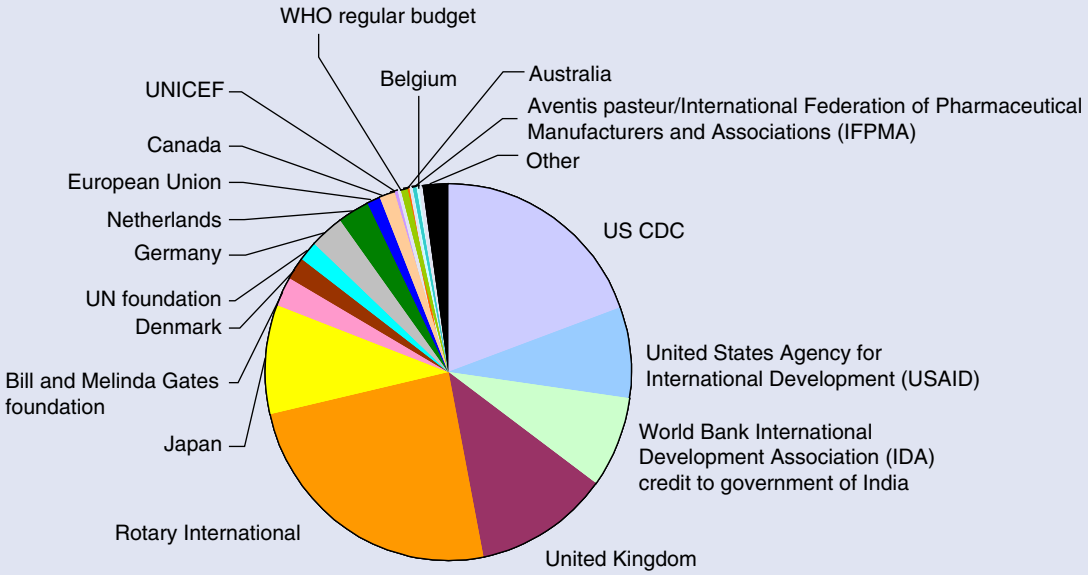
Here, voluntary contributions are the most straightforward option but are particularly prone to the free-rider problem as each country has an incentive to minimize its contribution. More formal coordinated contributions, negotiated or determined by an agreed formula, are commonly used to fund most international organizations (e.g., the World Health Organization (WHO)). Although limiting the 'free-rider' problem, each country has an incentive to negotiate the lowest possible contribution for itself (or the formula that will produce this result). Rewarding contributions with influence, to avoid this problem, skews power toward the richest countries (e.g., the international monetary fund (IMF) and World Bank); but without such incentives (or effective sanctions), countries have little incentive to pay their contributions in full (e.g., the US contributions to the United Nations (UN)). Global taxes, although theoretically the most efficient means for financing global public goods, face substantial opposition, limiting the prospects of securing funding from this source for the foreseeable future. More 'market'-based systems have been advocated, but as the USA's withdrawal from the carbon-trading system proposed in the Kyoto Agreement demonstrates, without effective enforcement mechanisms, the

Box 3 Global Polio Eradication Initiative

In the 1960s, two effective vaccines were licensed against polio, and by 1990 routine childhood immunization coverage against polio had risen to >70% worldwide, yet significant disparities in immunization coverage remained. In 1988, the World Health Assembly launched the Global Polio Eradication Initiative (GPEI). Everyone would be protected from polio (nonexcludable) and one person's protection will not reduce another's (nonrival). The problem was that the effort required to eradicate polio correlated inversely with income. In particular, the National Immunization Days required huge numbers of people and vehicles, and surveillance and laboratory work reporting standard data to the WHO regularly was also costly. So, how was it achieved?

Specific polio eradication activities were led, coordinated, and implemented by the governments of polio-infected countries but financed by a public–private partnership spearheaded by the WHO, Rotary International, the US communicable disease control (CDC), and United Nations Children's Fund (UNICEF). Rotary International especially played a central role through its 'PolioPlus' Program. The International Red Cross and Red Crescent Movement, The International Federation of the Red Cross, Médecins Sans Frontières, Save the Children Fund, World Vision, CARE, and the US-based NGO umbrella-organization CORE have also facilitated strategy implementation in the field. The United Nations Development Program, World Food Program, Office of the United Nations High Commissioner for Refugees, and others facilitated activities at the country level through the provision of transport, human resources, security, and communications. Civil society advocates, special ambassadors, business leaders, and celebrities from the arts, sciences, entertainment, and sports fields supported the GPEI, particularly in the areas of advocacy and communications.

Implementation of the GPEI required substantial in-kind and financial contributions from endemic and polio-free countries. Conservatively, polio-endemic countries are estimated to have contributed at least US$1.8 billion in volunteer time alone for polio eradication activities between 1988 and 2005, whereas external sources provided at least US$2.75 billion. External financing comes from a broad range of public and private sector sources (see figure below), channeled through multilateral funding through the WHO or UNICEF and direct bilateral funding to recipient countries, which allows the needs of both donors and recipient countries to be accommodated, although maximizing the efficient use of funds.



Overall, framing the GPEI as a global public good for health helped in understanding and presentation of the costs, financing and benefits of eradication, especially the emphasis on 'fair shares,' identification of the bearer of burden and opportunity cost, helping establish and sustain societal and political support.

free-rider problem remains. More recently, the constructive use of debt has been suggested, to allow the world to consume more goods that are global sooner and pay for them over a longer period. For certain diseases, the risks that they pose and the consequences of poverty that they perpetuate, debt (and hence loans) might make good sense. Buying time also allows the possibility that those countries not able today to help to pay for global public goods, borrow to do so in the future when their economies are more productive. The appropriateness of the precise mechanism chosen will depend on the specific good being considered.

Conclusion

The problem with public goods is that market mechanisms undersupply them. National governments usually provide finance and/or production. At a global level there is no world government. Thus global public goods require some means to ensure collective action to correct market failure at a global level. The advantage of the global public good concept in areas requiring global collective action is that it frames issues and objectives of policy – improving heath – in ways that make explicit the inputs needed (mix of public and private goods, domestic and international inputs, and incentives required) to produce and disseminate the final 'good.' Treating the final product as a 'good' in this way rather than a policy objective facilitates the analysis of who benefits and loses from its production, identifying (dis)incentives involved and thus facilitating the design of appropriate financing mechanisms.

The concept makes it clear that policy makers and their constituencies need to recognize interdependencies and the futility as well as the inefficiency of attempts to act unilaterally – porous borders have globalized health issues, and

international cooperation in health has become a matter of self interest.

See also: Pollution and Health

Further Reading

Commission on Macroeconomics and Health (2001). *Macroeconomics and health: Investing in health for economic development*. Geneva: World Health Organization.

Kaul, I. and Conceição, P. (2006). *The new public finance: Responding to global challenges*. New York: Oxford University Press.

Kaul, I., Conceição, P., Le Goulven, K. and Mendoza, R. (2003). *Providing global public goods: Managing globalization*. New York: Oxford University Press.

Kaul, I., Grunberg, I. and Stern, M. A. (1999). *Global public goods: International cooperation in the 21st century*. New York: Oxford University Press.

Sandler, T. (1997). *Global challenges: An approach to environmental, political and economic problems*, ch. 5. Cambridge, New York, and Melbourne: Cambridge University Press.

Smith, R. D. (2003). Global public goods and health. *Bulletin of the World Health Organization* **81**(7), 475 (editorial).

Smith, R. D., Beaglehole, R., Woodward, D. and Drager, N. (2003). *Global public goods for health: A health economic and public health perspective*. Oxford: Oxford University Press. Note that a set of accompanying slides and material for the above book can be found at http://www.who.int/trade/distance_learning/gpgh/en/index.html

Smith, R. D. and MacKellar, L. (2007). Global public goods and the global health agenda: Problems, priorities and potential. *Globalization and Health* **3**, 9, http://www.globalizationandhealth.com/content/3/1/9. doi:10.1186/1744-8603-3-9.http://www.globalizationandhealth.com/content/3/1/9

Smith, R. D., Thorsteinsdóttir, H., Daar, A., Gold, R. and Singer, P. (2004). Genomics knowledge and equity: A global public good's perspective of the patent system. *Bulletin of the World Health Organization* **82**(5), 385–389.

Smith, R. D., Woodward, D., Acharya, A., Beaglehole, R. and Drager, N. (2004). Communicable disease control: A 'Global Public Good' perspective. *Health Policy and Planning* **19**(5), 271–278.

Tobin, J. (1978). A proposal for monetary reform. *Eastern Economic Journal* **4**(3–4), 153–159.

Health and Health Care, Macroeconomics of

R Smith, London School of Hygiene and Tropical Medicine, London, UK

Introduction

Public expenditure targets, inflation, tax policy, and exchange rates, among other factors, will have effects on the provision of health care and the health status of the population. For instance, national income and fiscal targets will constrain how much a government can spend on health care, the exchange rate will be a factor determining the cost of vaccines and drugs, and tax policies relating to tobacco, alcohol, and 'fast food' will influence people's demand for these products and ultimately their health. Conversely, of course, the health of a population can significantly influence macroeconomics, affecting a country's rate of economic growth for example.

Macroeconomics, which encompasses these and other factors, is thus increasingly important for health and health care, especially as economies become more integrated in international trade and financial systems. This article outlines the key concepts within macroeconomics, and their application with respect to health and health care.

What is 'Macroeconomics'?

Economics is broadly divided into microeconomics and macroeconomics. Microeconomics is essentially concerned with choices and activities at the individual or firm level. It is concerned with what goods firms decide to produce and what goods households decide to consume. The interaction of households and firms takes place within a market, where price movements seek to equate demand and supply. Typically these markets are combined to form what are termed 'sectors,' such as agriculture, manufacturing, or health care. Together the interaction of these sectors comprises 'the economy.' Macroeconomics is then concerned with choice and activities across a number of these markets and sectors, and thus 'the economy' as a whole. In doing so, a whole set of terminology different to microeconomics is found, the main ones outlined in the glossary in Box 1.

International Trade

An important element of macroeconomics is international trade. According to the 'law of comparative advantage,' free trade (i.e., exchange of goods) between countries encourages countries to produce the goods that they are best placed to produce compared with other countries. A comparative advantage exists when an individual, firm, or country can produce a good or service with less forgone output (opportunity cost) than another. This differs subtly from 'absolute advantage'; for instance, where a country with lots of sunshine and wide open spaces could be seen to have an absolute advantage in agriculture compared to a country with little sunshine and mountains. Thus, call centers are increasingly located in countries such as India, not because their location there involves fewer inputs for any given number of calls or because wages are lower than elsewhere (which would confer an absolute advantage), but because the lost output from using people in this way rather than another way is smaller than it would be in, say, most European countries or North America. Conversely, research-based industries, like innovative pharmaceutical firms, are located mainly in high-income countries despite their relatively high wage levels because they too have a comparative advantage. Clearly, some countries may have an absolute advantage in producing nearly everything, but it is impossible for them to have a comparative advantage in everything. Conversely, some countries have an absolute advantage in virtually nothing, but they too necessarily have a comparative advantage in something. Given certain assumptions, total world production will therefore increase, and consumption possibilities increase, if countries specialize according to their comparative advantage and trade these goods with each other. Those countries that engage in trade will therefore see increasing gross domestic product (GDP), a wider selection of available goods and services, higher employment, and higher government revenues (due to higher income).

The problem of course is that, in practice, many countries create barriers to trade to 'protect' domestic industries, including tariffs, import restrictions, and bans. The effect of such protection is that it enables countries to continue to produce goods in which they have no comparative advantage, but at the same time discourages those countries who do actually hold the comparative advantage in such products. Why would a country do this? Typically this is specific political lobbying by an industry/sector or relates to an area deemed important for national security. However, the period since World War II has seen significant initiatives targeted to increase free trade, and has witnessed unprecedented increases in global trade activity.

How Does Macroeconomics Relate to Health and Health Care?

In this article, the term macroeconomics is used to refer to consideration of issues that fall outside of the health (care) sector. Thus it is not concerned with the inner workings of the health sector – such as how doctors are paid, or the cost-effectiveness of alternative screening programs – but the wider interactions between health and economy, health versus other sectors, and trade impacts on health. In this respect, there are a range of proximal and distal linkages between macroeconomics and health; illustrated in Figure 1. The lower half of the figure represents the individual country under

Box 1 Glossary

Appreciate

When a currency is rising relative to other currencies, it is appreciating in value.

Balance of payments (BOP)

Measures currency flows between countries. Payments are usually measured in the currency of the country that is paying. Payments made to other countries are seen as debits (e.g., imports) and payments received from other countries are seen as credits (e.g., exports). So an important indicator of a country's performance in international trade and investment is the level of surplus or deficit in their balance of payments.

Constant dollars

Constant dollars or currency correspond to values that have been adjusted for inflation and so reflect their 'real' or actual purchasing power as perceived from some base date.

Current dollars

Current dollars or currency refer to the actual dollars spent, with no adjustment for inflation.

Depreciation

When a currency is falling relative to other currencies, it is depreciating in value.

Depression

A sustained, long-term, downturn in economic activity – more severe than a recession, often judged as a 10% decrease in real Gross Domestic Product.

Economic growth

A positive change in the level of production of goods and services by a country, usually measured annually.

Exchange Rates

Exchange rates tell you how much one country's money is worth in another country's currency. If the value of a currency is going down relative to another, it is depreciating; if it is rising relative to other currencies, it is said to appreciate in value. Fluctuations in exchange rates are very important as every country imports and exports goods and services.

Fiscal policy

Policies introduced by the government to influence the economy through taxes and government spending.

Gross Domestic Product

GDP is the total expenditure by residents and foreigners on domestically produced goods and services in a year. It is the main indicator used to measure the size or output of an economy.

Gross national income

Gross National Income (GNI) measures the economic activities undertaken by residents and firms of that country regardless of where they take place. GNI is GDP plus income earned by its residents from abroad minus income earned in that country by residents of other countries abroad.

Inflation

General rise in prices over time. This means that money loses its value (purchasing power) through time.

Monetary policy

Policies by the government of adjusting interest rates and the amount of money in circulation.

Price index

A price index is created by selecting a bundle of goods and services according to the purpose of the index. Their prices are collected in a base year and compared with prices of the same bundle in another year. The overall price change of the goods in the bundle measures inflation. The price index is set at 100 for the base year and subsequent changes in prices are compared with this base year.

Purchasing power parity

An exchange rate that equates the prices of a basket of identical traded goods and services in different countries.

Recession

A downturn in the rate of economic activity, with real GDP falling in two successive quarters.

consideration, and the upper half the aspects of the international system. The arrows between the various components indicate the major linkages. This is a deliberately simplified picture to provide a concise and understandable frame of reference.

Taking the lower half of the figure first, what may be termed as the 'standard' influences on health are illustrated. These include risk factors, representing genetic predisposition to disease, environmental influences, and infectious disease. Next is the household, which represents factors associated with how people behave and, crucially, invest in their health. There is then the health sector, which comprises those goods

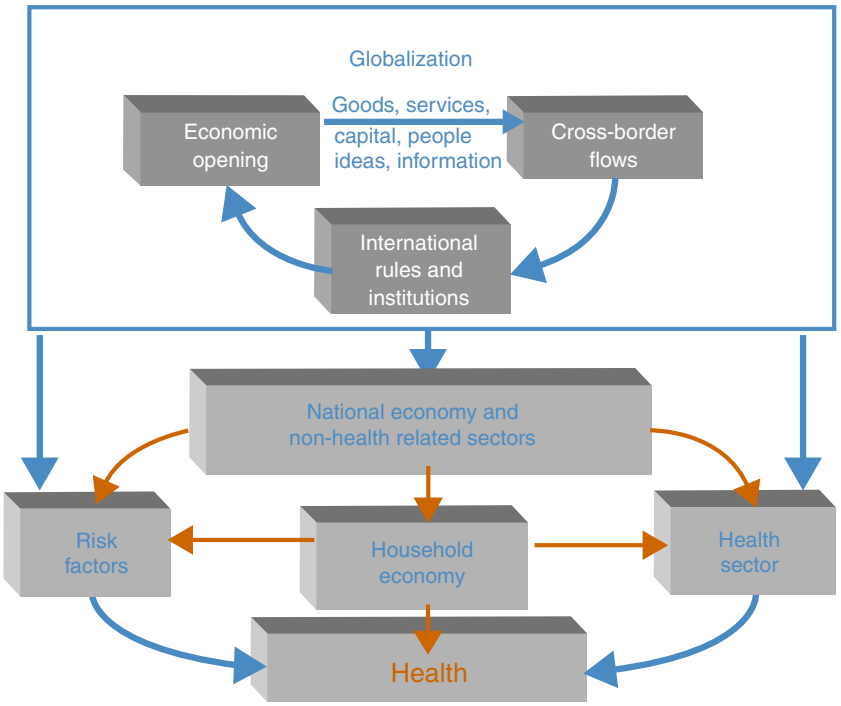


Figure 1 Major elements and linkages between macroeconomics and health.

and services consumed principally to improve health status. Finally, encompassing all these, there is the national economy, representing the metainfluences of government structures and other sectors.

In the upper half of the figure, the influences of factors that are usually outside national government jurisdictions are illustrated. For example, there is a wide variety of international influences directly upon risk factors for health, including an increased exposure to infectious disease through cross-border transmission of communicable diseases, marketing of unhealthy products and behaviors, and environmental degradation. Increased interaction in the global economic system will also affect health through influences upon the national economy and wealth. It is well established, for instance, that economic prosperity is 'generally' positively associated with increased life expectancy. Finally, health care will be affected through the direct provision and distribution of health-related goods, services, and people, such as access to pharmaceutical products, health-related knowledge and technology (e.g., new genomic developments), and the movement of patients and professionals. Also note that in this upper half of the figure, the importance of international legal and political frameworks that underpin much of these activities, such as bilateral, regional and multilateral trade agreements is seen.

In terms of linkages between these influences, increased macroeconomic trade will bring associated changes in risk factors for disease. These will include both communicable diseases, as trade encourages people and goods to cross borders, and noncommunicable diseases, as changes in the patterns of food consumption, for instance, are influenced by changes in income and industry advertising. Increased macrolevel interaction will also impact upon the domestic economy through changes in income and the distribution of that income, as well as influencing tax receipts. This will influence the household economy and also the ability of the government to be engaged in public finance and/or provision of health care. Finally, there will be direct interactions in terms of health-related goods and services, such as pharmaceuticals and associated technologies, health care workers, and patients. Let us explore these in a little more detail.

Macroeconomics and the Household

Macroeconomic policy is concerned with economic growth – increasing levels of GDP – as higher GDP leads to greater opportunities to consume which will, *ceteris paribus*, improve health (although it may not!). The relevant factors in this relationship are improved nutrition, sanitation, water, and education. In this respect, engaging in global macroeconomic integration – or international trade – is a key factor leading to economic growth through specialization. However, although trade liberalization may be poverty-alleviating in the long run, at least in the short term it is often the adverse consequences, particularly to the most poor, that are observed (e.g., increased cost of living, development of urban slums, chronic disease, pollution, and exploitative and unsafe work conditions) and lead to significant ill-health.

One of the criticisms of conventional macroeconomic approaches is the inadequate attention paid to distributional impacts – most are generally based on the aggregate indicators such as 'total' income, trade volume, employment, etc. This reflects a focus on growth and efficiency over equity. Thus, although trade liberalization may be advantageous, the crucial

factor in how advantageous and to who depends on how countries manage the process of integrating into the global economies. For example, employment creation through economic growth is often also accompanied by job destruction as labor moves from one sector or industry to another. In the absence of social safety nets, not only does such economic insecurity potentially push people into poverty, but it can also impact on health through the stress caused by economic and social dislocation.

Another important aspect of macroeconomic growth and health is that of the stability of the growth. Economic instability results in volatile markets, increased frequency of external shocks, and increased impact of such shocks. These translate into economic insecurity for an individual, which is closely linked to increased stress-related illness. It will also affect the adequacy of financial planning for ill-health by the household and the (public and private) health sector, and generate investor reluctance (including within the health sector itself).

Economic stability is affected, among other things, by the proportion of income/growth dependent on trade, with the general view that trade liberalization, especially in financial services and in the movement of capital, results in volatile markets. Of course, being an open economy does not automatically lead to economic instability/shocks – it is smaller, often developing countries, where trade contributes a much higher share of GDP that are more vulnerable as they rely more on imports and exports.

Macroeconomics and Risk Factors for Disease

It is well documented that there are many 'social determinants of health,' which refer to the general conditions in which people live and work and which influence their ability to lead healthy lives. These include factors such as employment, nutrition, environmental conditions, and education. These 'social determinants' contribute to the risk of different diseases and are often seen to differ in their role in influencing communicable and noncommunicable diseases.

The contribution of macroeconomics to the spread of communicable diseases is made in two ways. First, the overall environment in which people live (concerned with pollution, sanitation, etc.) is determined – in large part – by their income and wealth. Second, the increased international movement of people, animals, and goods associated with increased trade will affect the movement of disease. This is illustrated well by the example of SARS and other areas.

Perhaps less obvious is the relationship between macroeconomic activity and noncommunicable disease. Although macroeconomic growth can be beneficial when it leads to an expansion in the consumption of the goods that improve health, such as clean water, safe food, and education; it also facilitates the increased consumption of goods which may be harmful or hazardous to health, which may be termed 'bads.' Trade liberalization will reduce the price of imported 'bads' through reduced tariff and nontariff barriers, and increase the marketing of 'bads,' such as tobacco, alcohol, and 'fast food.' In the case of alcohol and tobacco, the development of regional trade agreements have helped to significantly reduce barriers to trade in tobacco and alcohol products, by breaking up the hitherto protected markets, contributing to enhanced consumption.

In terms of food-related products, increased macroeconomic integration will affect the entire food supply chain (levels of food imports and exports, foreign direct investment in the agro-food industry, and the harmonization of regulations that affect food), which subsequently affects what is available at what price, with what level of safety, and how it is marketed. For example, in what is termed the 'nutrition transition,' populations in developing countries are shifting away from diets high in cereals and complex carbohydrates, to high-calorie, nutrient-poor diets high in fats, sweeteners, and processed foods. Increased trade liberalization is one driver of the nutrition transition because it has had the effects of increasing the availability and lowering the prices of foods associated with the growth of diet-related chronic diseases, as well as increasing the amount of advertising of high-calorie foods worldwide. Furthermore, trade and economic development encourages the use of labor-replacing technologies, such as cars, and creates greater leisure time, both of which in turn can be seen to encourage more sedentary lifestyles.

Macroeconomics and the Health Sector

Perhaps the most visible link between macroeconomics and health is at the overall level of health care spending. Most nations, rich or poor, face the problem of rising health care costs and confront two basic questions: How to finance this rising burden and how to contain the pressures for health expenditure growth. Here, the critical issues relate to government-funded health care, where the ability to finance and/or provide public services is determined by tax receipts. Tax income is broadly dichotomized into taxes that are 'easy to collect' (such as import tariffs) to those that are 'hard to collect' (such as consumption taxes, income tax, and value added tax). Tariff revenues are a very important source of public revenues in many developing countries.

Trade liberalization, by its nature reduces the proportion of government income from 'easy to collect' sources. Although theoretically, governments should be able to shift tax bases from tariffs to domestic taxes, such as sales or income taxes, in practice, developing countries, especially low-income countries, find this difficult, especially because of the informal nature of their economies with large subsistence sectors. Low-income countries are usually able to recover only approximately 30% of the lost tariff revenues resulting in a decline of government income available to pursue public policies, be it through health care, education, water, sanitation, or a social safety net.

The exchange rate is also a key determinant of the relative prices of imported and domestically produced goods and services. For many countries, products such as pharmaceuticals, but also various elements of other technologies, such as computer equipment, surgical tools, and even lightbulbs, used to provide health care are imported. Changes in the exchange rate brought about by macroeconomic developments may therefore see the price, and hence cost, of health care

increase or decrease. Conversely, changes in demand for domestically produced goods from overseas importers may see the price of those goods domestically change in response (e.g., increased foreign demand may push up local prices). Increased linkage between economies at the macrolevel thus generates greater levels of exogenous (i.e., beyond the domestic health sector control) influences over prices, and hence cost of health care.

Finally, the health sector is increasingly involved in the direct trade of health-related goods and services. For instance, spending on pharmaceuticals represents a significant portion of health expenditure in all countries. Pharmaceuticals are also the single most important health-related product traded, comprising approximately 55% of all health-related trade by value (the share of the next most significant health-related goods traded, small devices and equipment, is $<20\%$). The market is highly concentrated, with North America, Europe, and Japan accounting for approximately 75% of sales (by value). Overall, high-income countries produce and export high-value patented pharmaceuticals and low- and middle-income countries import these products; although some produce and export low-value generic products. This leads to many developing countries experiencing a trade deficit in modern medicines, which often fuels an overall health sector deficit.

Trade in health capital and services has also expanded greatly in the last decade, in large part due to improvements in information and communication technology. These improvements have contributed, for instance, to the remote provision of health services from one country to another, known as 'e-health.' Examples of services provided include diagnostics, radiology, laboratory testing, remote surgery, and teleconsultation.

Another type of trade in health services arises from the consumption of health services abroad. This is also known as 'health tourism' and it entails people choosing to go to another country to obtain health care treatment. This attracts approximately four million patients each year, with the global market being estimated to be US$ 40–60 billion.

As liberalization increases and migration becomes easier, the movement of people across borders also increases. As a result, many health professionals choose to leave their home countries for richer, more developed ones. This is the case for doctors, nurses, pharmacists, physician assistants, dentists, and clinical laboratory technicians. It is estimated that in the UK, the total number of foreign doctors increased from 20 923 in 1970 to 69 813 in 2003. These figures may not seem that significant, but they often represent a large share of a country's total doctors. In Ghana, for example, the number of doctors leaving accounts for 30% of the total number of doctors.

Conclusion

Health is essential not only for human development, but also for economic development. Economic development also significantly influences health. This reciprocity means that activities at the macrolevel are increasingly important to population health, and the provision of health care. The growing interconnectedness between countries especially through greater trade and trade liberalization means that health sectors are more vulnerable to shocks from events that are happening around the world. It is therefore of critical importance that those concerned with health and health care have an understanding of the core issues; further articles in this volume are therefore highly recommended.

See also: Education and Health in Developing Economies. Emerging Infections, the International Health Regulations, and Macro-Economy. Global Health Initiatives and Financing for Health. HIV/AIDS, Macroeconomic Effect of. International E-Health and National Health Care Systems. International Movement of Capital in Health Services. International Trade in Health Services and Health Impacts. International Trade in Health Workers. Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity. Macroeconomic Dynamics of Health: Lags and Variability in Mortality, Employment, and Spending. Macroeconomic Effect of Infectious Disease Outbreaks. Macroeconomy and Health. Medical Tourism. Noncommunicable Disease: The Case of Mental Health, Macroeconomic Effect of. Nutrition, Health, and Economic Performance. Pharmaceuticals and National Health Systems. What Is the Impact of Health on Economic Growth – and of Growth on Health?

Further Reading

Bloom, D. and Canning, D. (2000). The health and wealth of nations. *Science* **287**(5456), 1207–1209.

Blouin, C., Chopra, M. and van der Hoeven, R. (2009). Trade and social determinants of health. *The Lancet* **373**(9662), 502–507.

Blouin, C., Drager, N. and Smith, R. D. (eds.) (2006). *International trade in health services and the GATS: Current issues and debates*. World Bank.

Hsiao W. and Heller, P. S. (2007). What Should Macroeconomists Know about Health Care Policy? IMF Working Paper WP/07/13. Available at: http://www.imf.org/external/pubs/cat/longres.cfmsk=20103.0 (accessed 27.02.07).

Pritchett, L. and Summers, L. H. (1996). Wealthier is healthier. *The Journal of Human Resources* **XXXI**, 841–868.

Sachs, J. (2001). Macroeconomics and health: Investing in health for economic development. *Report of the Commission on Macroeconomics and Health*, Geneva: World Health Organization. Available at: http://www.cmhealth.org/.

Smith, J. (1999). Healthy bodies and thick wallets: The dual relationship between health and economic status. *Journal of Economic Perspective* **13**(2), 143–166.

Smith, R. D. (2012). Why a macro-economic perspective is critical to the prevention of non-communicable disease. *Science* **337**, 1501–1503, Available at: http://www.sciencemag.org/cgi/content/short/337/6101/1501.

Smith, R. D., Chanda, R. and Tangcharoensathien, V. (2009). Trade in health-related services. *The Lancet* **373**, 593–601.

Smith, R. D. and Correa, C. (2009). Trade, trips, and pharmaceuticals. *The Lancet* **373**, 684–691.

Smith, R. D. and Lee, K. (2009). Trade and health: An agenda for action. *The Lancet* **373**, 768–773.

Relevant Websites

http://www.bbc.co.uk/news/business11177214
British Broadcasting Corporation.

http://www.economist.com/node/12637080story_id=E1_TNGPSDRD
The Economist.
http://www.guardian.co.uk/world/2008/apr/29/philippines
The Guardian.
http://www.sundaytimes.lk/070527/FinancialTimes/ft306.html
The Sunday Times.
http://www.who.int/mediacentre/factsheets/fs301/en/index.html
The World Health Organization.
http://www.wto.org/library/flashvideo/video_e.htmid=6
The World Trade Organization.

Health and Health Care, Need for

G Wester, McGill University, Montréal, QC, Canada
J Wolff, University College London, London, UK

Glossary

Capacity to benefit It refers to a particular definition of 'need' (for treatment), whereby a patient only has a need for treatment where his or her health will improve as a result of that treatment.

Cost-effectiveness A measure of the cost per desired outcome or effect of an intervention or course of action. Whether a given intervention is considered cost-effective typically depends on how it compares to other relevant alternatives with similar outcomes; the intervention with the lowest cost per desired outcome is the most cost-effective intervention.

Fair innings It represents the idea that we are all entitled to a fair chance in enjoying a lifespan of ordinary length.

Health equity It refers to the study of differences in the quality of health and health care across different populations. It relates in general to ethical judgments about the *fairness* of the distribution of such things as income and wealth, *cost* and *benefit, access* to health services, exposure to health-threatening hazards and so on.

Life expectancy The expected number of years a person or population may anticipate to live, at birth or at any given age.

Normal functioning The correct working of a person's bodily parts (e.g. a limb or organ), functions (e.g. breathing, digesting) or structures (e.g. teeth, bones).

Opportunity range The array of activities and projects available to a person, referring both to the variety and quantity of possible undertakings.

Rationing The rationing of health care refers to the denial of a treatment to a patient, or a class of patients, who would have benefited from that treatment. The usual ground for such a decision by the patient is that the price has been judged to be too high relative to their expected benefit. The usual ground under public or private insurance is the high cost of treatment relative to its expected benefit as judged by third party payers.

Severity of health state An evaluation of how bad that health state is for the person.

Introduction

Any society must come to a decision concerning the allocation of health resources, of which access to medical services, or health care, is the clearest focal point. For many theorists and ordinary citizens health care services are a 'special' type of good that should not be distributed on the market-based principle of ability to pay. Rather, it is often said, health care should be distributed on the basis of 'need'; if there was ever a place for Marx's dictum 'each according to their need' it would seem that health care would be a good candidate. Bernard Williams (1973, p. 240), for example, suggested that "the proper ground of distribution of medical care is ill health: this is a necessary truth." Of course not all societies have organized themselves entirely on this basis, but virtually all countries include a significant element of distribution of health care resources on the basis of some notion of need, whether as the main criterion for allocation, as in most European countries, or in services for the elderly, the poor, and the military personnel, as in the US. But if health care is to be distributed according to need it is necessary to explain what a need for health care means.

It would seem that, because the purpose of health care, broadly speaking, is to promote health, the need for health care must be derived from the need for health. Therefore, one ought to start with a prior question, what is the need for health?

Here it is argued that 'distribution according to need' names a general approach to health policy as opposed to distribution on the basis of ability to pay, rather than a specific principle of distribution. One reason for this claim is that all of the most prominent candidates for specifying a principle of distribution of health resources according to need face difficulties. Accordingly, a policy maker wishing to allocate resources according to health need will be compelled to balance a number of need-related considerations, among other relevant concerns, rather than follow a specific principle of distribution.

The Concept of Health

To discuss different ways in which one can be said to need either health or health care, it is necessary, first, to clarify what is meant by health. But what health is continues to be highly contested. Nevertheless, without claiming to have resolved any of the difficult questions on which these debates center, it is possible to give a rough outline of a concept of health for the purposes of this discussion.

Consider, first, two well-known but rather different definitions of health. According to Christopher Boorse's (1977) definition, health is the absence of disease. How disease, in turn, is defined is one of the most important parts of Boorse's account of health and merits much more discussion than can be accommodated here, but suffice it to say that disease, in his view, is a deviation from the 'normal' functioning of certain parts and processes of the organism. Even disabilities and injuries would fall within the scope of this definition of disease, and as such it leaves a much narrower range left for health compared to how it is ordinarily understood.

In contrast to this definition, the World Health Organization (WHO) adopts a much wider definition of health, according to which health is "a state of complete physical, mental and social well being and not merely the absence of disease or infirmity" (1946).

Although it seems right that health should be closely related to well being, the WHO's definition goes too far: in this view, health problematically appears to be indistinguishable from well being or happiness. Nevertheless, this definition draws our attention to a different aspect of health beyond Boorse's definition in terms of the absence of disease: that of 'positive' health achievement. One can imagine other such positive health achievements, for example, athleticism or living a healthy lifestyle. Furthermore, certain aspects of health such as physique, physical strength, or endurance, can today be enhanced through various drugs and procedures – such enhancements could be said to constitute improvements to one's health, regardless of their impact on the presence or absence of disease.

It cannot be resolved here which of these two, or indeed of the many other definitions that have been proposed, represents the most appropriate account of what health is. However, a disease model of health seems more appropriate in the context of a discussion of health need. Even if it is allowed that 'merely' being free of disease is not the 'best' level of health one can achieve, and that there are other health states that are superior, it is more difficult to see the 'need' for this latter form of health achievement. It might be proposed, then, that the need for health is best captured in terms of the need to be (reasonably) free of disease. For our purposes, the disease concept is narrowed down further, to encompass only such deviations from normal functioning that are harmful to the person.

But it is also necessary to think about the question of time span. The definition of health given so far is silent on the question of how extended a period should be considered. It is clear that one may be afflicted by disease at any given time in one's life. Moreover, the duration of any particular disease can vary greatly; some are short-lasted or can be cured, others are chronic. Clearly, the duration of any disease will matter greatly to how significant a departure from health one thinks that disease state represents. However, health is also a prerequisite for life itself – without health there is no life. So one could also conceive of health in terms of the time passed before the total loss of health, death, occurs.

Thus, time and duration seem of central relevance when health is discussed. One may therefore ask whether duration or longevity should be included in the concept of health. Is the duration of health – the length of life – a dimension of health? It would then follow that a shorter life would be a less healthy life. That would be true even if life had been lived 'in full health,' completely free of disease, in every moment up until the point of death. Alternatively, it could be said that length of life is simply health combined with duration. In that case, the length of life would not affect how healthy one would consider a particular life to be.

For our purposes, health is defined as including duration. That means that not only the absence of disease, but also a lifespan of a certain length, constitute the baseline of health achievement against which health need is measured.

The Logic of Need

It is often argued that 'need' is a three-element relation: in case of human need x – a person – needs y – an object – in order to z – to achieve a purpose or goal. In this framework it is clear that one question of health need is what is needed to achieve health (health as z). Yet one can also ask what may be a logically prior question: What is health needed for (health as y)? This prior question will be considered first.

At least two central dimensions of one's quality of life where one's health will have a considerable impact can be identified. The first of these dimensions is well being. Disease is often accompanied or constituted by various forms of suffering such as pain, nausea, 'feeling ill,' or feelings of anxiety or depression, all of which have a very direct and to varying degrees negative impact on our immediate physical and mental well being. The second dimension is the ability to engage in ordinary human activities. Norman Daniels has discussed this in terms of the importance of 'normal species functioning,' a concept adapted from Boorse's framework, for enjoying a normal opportunity range. The concept of normal species functioning is less clear than can be wished for, but at least a few relatively uncontroversial examples of normal functioning, come to mind such as having all major limbs intact, basic mobility, and being able to see and hear. These and other functionings will clearly be important for the pursuit of a wide variety of goals and projects. Many health conditions will be detrimental to or involve the loss of such functionings, and will hence negatively affect our opportunity range.

Nevertheless, the idea that health is needed for opportunity is not without difficulties. Consider, for example, the extent to which a condition such as paraplegia would affect one's opportunity range. It has been pointed out that, a person living in a poor rural village with only dirt roads is likely to experience paraplegia as a much more disabling condition than a person living in a wealthy, urban environment with a well-developed infrastructure. In other words, the extent to which limited mobility or other functional impairments will restrict one's opportunity range also depends on the nature and quality of one's social and material environment, and not just on the health condition itself.

But even individuals living in the same environment may be affected very differently by the same health condition depending on their own particular circumstances, such as their resilience, ability to adapt, social support network, or their preferences. The level of health achievement that is needed in order to enjoy a reasonable range of opportunities will clearly vary across such individual circumstances as well as the social, cultural, and historical context.

Furthermore, longevity generally tends to be valued, and it is not uncommon to think that a certain length of life is a central aspect of a good life. It is not immediately clear exactly what it is about a shorter life that is unfortunate; after all, a premature death does not in itself, retrospectively as it were, alter the quality of life lived up to the point of the onset of death or the events that led to death (though that is not to deny that having advance knowledge of one's own to be shortened lifespan is likely to affect one's quality of life in various ways). But perhaps one could say that a shorter life is a life with less opportunity, both in terms of variety and the total 'amount' of opportunity. This

diminished range of opportunities due to premature death would not affect the individual in the same way as diminished opportunities due to loss of functionings – perhaps it is not even quite correct to say that the diminished range of opportunities in the former case really 'affects' the individual's lived life as such – but the loss of opportunity still represents an unfulfilled potential and, therefore, a shortfall.

Returning to the idea of need as a three-element relation (x needs y in order to z), it was noted that in addition to the question discussed in the previous section – what is health needed for? – one can ask 'what is needed to achieve health?' This is the central question of 'health need,' which will be considered next.

Health care appears to be the most obvious candidate for what is needed to achieve health. After all, health care is a means to improving health, and thus it would seem that a need for health is simultaneously a need for health care. However, not all health needs indicate a need for health care. Ordinarily, other basic needs such as food, water, sanitation, and shelter must be met as a minimal precondition for health; many health needs arise as the result of a failure to meet these other needs. In such cases, although health care might be necessary for short-term intervention, ensuring that these basic needs are met will clearly be more effective for overcoming population health needs in the longer term. Even in developed societies where basic needs are mostly catered for, it is argued that a level of health need arises as a result of poor quality housing, material insecurity, working conditions, and social exclusion. In many cases, targeting such 'upstream' causes of disease will be a better strategy for reducing health need overall.

A need for health, then, cannot be identified with a need for health care; only some health needs are at the same time health care needs. The concept of health care need will be considered once more in the last part of the article, but first the relationship between a shortfall in health and the need for health will be considered in more detail.

The Health Baseline

For the purposes of this discussion, health is conceived as the absence of harmful disease (understood very broadly). But it is also noted that what is to count as 'harmful disease' can vary culturally and individually. It is also suggested that longevity should be seen as a dimension of health. The notions of the absence of disease, and of living to a certain age, function not only as conceptions of health, but can also be conceived of as a particular baseline of health, against which shortfalls in health can be measured. Thus, premature death and the presence of disease both in different ways represent shortfalls in health achievement.

This baseline of health has a double function. On the one hand it provides an account of what it is to achieve health (as a means to a life of good experience and opportunity). On the other it provides a standard by which other things, such as health care, can be judged as meeting health need or not. The baseline of health, therefore, is central to the concept of need for health and health care. The question of what this baseline of health should be will be considered next.

On the most expansive conception of health need, the highest attainable health would be adopted as the baseline against which health need is measured. Consider the case of life expectancy. The life expectancy at birth in Japan, which is one of the highest in the world at nearly 84 years (CIA: The World Factbook, 2012), is usually used as the standard for the highest attainable life expectancy. Accordingly, if this life expectancy is adopted as the relevant baseline, any shorter life expectancy represents a health need. However, one might be skeptical of the idea that any shortfall from this very high standard of health is appropriately characterized as a health need. The UK, for example, has a slightly lower life expectancy at birth than Japan at approximately 80 years (CIA Factbook). But would one thereby say that the UK has a health need? This seems debatable.

One possible argument is that a shortfall in health is only a health need if it reflects a genuine possibility for health gain. But it is not clear that the highest known life expectancy attained by some is attainable by all. This will depend on what factors determine longevity and the extent to which these factors are within the scope of human control. Perhaps longevity is partly genetically determined. Other determinants, such as diet and lifestyle, are in principle within our control, but in practice it is hard to imagine a government imposing a particular diet on its citizens. One can see why one might think that only cases where there is a genuine possibility for improving health should be considered a health need: after all, to say that there exists a need seems to imply that something ought to be done. And to say that something ought to be done in turn seems to imply that something can be done – or so proclaims that familiar Kantian principle.

This issue can be set aside for now. Instead, consider a different reason to be skeptical that the UK has a health need in this case. One could argue that the highest attainable health is simply the wrong standard of health against which to compare our own health achievement for the purpose of identifying health need. Just as athleticism or other forms of positive health achievements go beyond what one would ordinarily say is needed, this ideal standard also seems to exceed what is required. Reserving the term 'need' for more substantial shortfalls in health seems more intuitive.

This point can be accommodated if a more modest level of health is adopted as the relevant baseline, for example, a level of health that it is reasonable or realistic to expect to attain. Alan Williams has expressed a related view with respect to length of life, arguing that 'we are each entitled to a certain level of achievement in the game of life,' and that anyone exceeding this level, which he refers to as a 'fair innings,' 'has no reason to complain when their time runs out' (Williams, 1998, p. 319). It is possible to extend and apply this concept of a 'fair innings' to the standard of health; the idea is that because it is clearly both possible and desirable to improve health beyond this level, a person or a population that has reached this standard of health has attained a fair or sufficient level of health, and therefore does not have a need for health.

Although it remains true that there is a sense that someone who has lived beyond the age of the 'fair innings' could understandably still claim to have a need for health, just as a wealthy person could claim a need for more money, there is a sense, in both cases, in which one could say that their needs

have been met, and what they claim to need is a form of luxury or excess. On this account, need is assimilated to something like basic need. It is true that one can have further needs even when basic needs have been met, but for political purposes it could be that only basic needs call for action.

It may be that this notion of a fair innings of health does not lend itself equally well to all dimensions of health or all levels of analysis; or perhaps one must approach the notion of sufficiency differently with regard to such different dimensions of health rather than speak of sufficiency of health overall. For example, perhaps only moderate levels of pain will be accepted as 'within' our standard for being 'sufficiently' pain free, whereas our standard for a sufficient length of life could be significantly lower than the known human potential; and having achieved sufficiency in one dimension of health may not imply sufficiency in a different dimension. Clearly, the notion of a fair innings of health requires more work. Nevertheless, one can make sense of the idea that a shortfall from or failure to achieve the highest attainable level of health does not have to indicate a health need.

If this idea of a fair innings of health is accepted, how should one go about determining what level of health it would be reasonable to expect to attain? Health outcomes are partly determined by one's social, material, and economic environment. The quality and nature of this environment in turn depend on a society's level of affluence and on how its resources are distributed. The question of what level of health it is reasonable to expect to attain can only be answered with reference to these further substantive issues, and as such is hardly normatively neutral.

On the global level, there are enormous inequalities in material standards of living. Hundreds of millions of people live in extreme poverty lacking adequate nutrition, clean drinking water, sanitation, and access to basic health care. Whereas these levels of extreme poverty are avoidable, it is perhaps less clear what level of material living conditions would be generally attainable if global resources were distributed more fairly. The standard of living is not the only important determinant of health, and health achievement is unlikely to improve exponentially with improvements in the material standard of living; nevertheless, the realistically attainable standard of living is likely to impose some constraints on the level of health one can reasonably expect to attain. For example, it seems dubious that the exceptionally high standard of living found in Monaco, where citizens are generally extremely wealthy, is attainable for all. Life expectancy at birth here is the highest in the world at nearly 90 years (CIA: The World Factbook, 2012) – but insofar as this health achievement is a result of their wealth and high standard of living, it is not realistically attainable for the world's population as a whole.

The question of what standard of living will be generally attainable aside, within any society there will be other important decisions to be made about how much priority should be given to the promotion of health over other things that are valued. Such questions of priority are likely to arise in many different contexts, but one can illustrate the point by considering the case of reducing or eliminating health risks. How much effort should be expended on this task? Some interventions furthering this objective could have prohibitive costs in other areas of life. For example, road accidents, being one of the top 10 causes of death worldwide (WHO (World Health Organization), 2008), constitute a severe health risk. However, even if banning the use of cars altogether were to improve our health overall, there are obvious reasons why it is neither desirable nor practicable to go through with such a proposal.

The answer to the question of what level of health it is reasonable to expect to attain will depend on other normative judgments, such as 'what is a fair distribution of resources?' and 'how important is health compared to other dimensions of quality of life?' Depending on what answers are given to these and related questions, one will have different ideas about the appropriate baseline of health achievement against which shortfalls in health should be measured, and therefore, about what counts as a health need.

Three Concepts of Health Care Need

Next, consider the question of need for health care. It was established that health care is not always needed to achieve health. But it is necessary to look at the relationship between need for health and need for health care in more detail. Here, three concepts of health care need which each limits the concept of need in different ways are considered: presence of disease, capacity to benefit, and cost-effectiveness of treatment.

The idea that the presence of disease equals a need for health care is very straightforward: if a person is sick or injured, it seems natural to say that he or she is in need of health care. However, not all diseases can be treated or cured. Although one could still consider such cases to be health needs, it is perhaps less clear whether one can say that there is a need for health care in these cases. Arguably, it seems strange to say that there is a need for health care if no health care exists, or if health care provision is at such a primitive or underdeveloped level that it would be harmful rather than beneficial. Many medical practices common in the past are now known to be either inefficient or in fact harmful, such as lobotomy or bloodletting; it cannot be said that there was ever a genuine need for such services.

However, it seems more appropriate to say that such cases represent a need for health care in general, even if there is no specific treatment available at a given time that would be of benefit. Furthermore, one can point to examples where it might be said that effective health care 'ought' to have been available. For example, not much effort has been spent on developing modern effective treatments for a group of debilitating diseases often referred to as 'neglected tropical diseases.' This group of diseases primarily affects poor populations in the developing world, and has typically received little attention from the pharmaceutical industry; there is reason to believe that more funding and research could lead to significant improvement in treatment options. In cases such as these, it also seems right to say that there is a need for health care, even if currently no specific treatment exists.

In other cases, treatment is available, but for different reasons a particular individual may be unable to benefit from the treatment. For example, a treatment may be contraindicated for patients outside a particular age bracket, patients with other, preexisting health conditions, and so on. These

patients would not benefit from the treatment in question. It therefore seems somewhat counterintuitive to say that these patients 'need' this particular treatment.

For reasons such as these, some would reject the proposal that the presence of disease itself is sufficient for there to be a need for health care. That brings us to our second proposed definition of need for health care, as 'capacity to benefit (from treatment).' This definition is often favored by health economists. According to this view, a patient is only in need of a given treatment if the patient can benefit from that treatment. Thus, on this view the patients in the examples above could not be considered to be in need of that particular health care treatment.

In many ways this definition of health care need is intuitive. At the same time, narrowing down the concept of health care need in this way does not seem to take anything away from our sense that something ought to be done. As has already been suggested, it seems important to distinguish between the need for a particular treatment or intervention (or the lack thereof), and a more general need for health care. Furthermore, the reasons why a given treatment will not be effective also seem to matter to our judgment. In some cases, for example, the treatment being effective is contingent on the patient complying with certain behavioral requirements, for example, quitting smoking or losing weight. In this case, it seems somewhat more intuitive to say that the patient needs the treatment, even if he or she is failing to comply with the requirements in question. Alternatively, imagine that the effectiveness of a treatment was contingent on the patient being well nourished before the start of the treatment. In cases where lack of resources meant patients were inadequately nourished, it also seems incorrect to say that the patient had no need of the health care treatment in question.

The ability to offer decent health care may also be limited by resource shortage and competing needs. Many countries limit the availability of health care in accordance with the cost-effectiveness of the various treatments or interventions. Sometimes certain treatments will not be offered, even if they can improve a patient's health, because the cost is considered too high relative to the health benefits it would yield. Our third proposed definition of health care need incorporates considerations of cost-effectiveness, such that a patient is considered to be in need of a given treatment only if that patient will benefit from that treatment, and that treatment is considered to be cost-effective. Thus, a patient does not need a given treatment if that treatment is too expensive or yielding too little health benefit to be cost-effective, even if the patient could benefit from the treatment.

Some very expensive and cost-ineffective cancer drugs for advanced stage disease are sometimes excluded on the grounds that they are not cost-effective. In cases where the cancer cannot be cured, treatment may nevertheless give the patient a few more months of life. In the UK there have been cases where these drugs were not offered through the National Health Service because they were deemed of too limited benefit to justify their very high cost. How many patients can avail of a given treatment can affect the price and hence the cost-effectiveness of that treatment. The so-called orphan drugs is a relevant example here. Orphan drugs are drugs for very rare conditions. If a condition is rare, market demand for the drug will be expected to be low, and it will be difficult for a pharmaceutical company to sell enough drugs to cover the expenditure involved in the research and development of the drug. Therefore, the price of such drugs is often very high, and they will rarely be cost-effective.

In this definition of health care need, the extent of need in a population will be relative to the society's level of affluence. That leads to the interesting implication that as a society becomes wealthier, and can afford to relax the cost-effectiveness constraints, all else being equal, the total need for health care would in fact increase. Although it may seem a surprising result that the need for health care increases in accordance with a society's increased wealth, this view also captures something of importance. For example, in a wealthy society, crooked teeth could be considered a need for dental services. But in a very poor society, the correction of crooked teeth would rather be considered a luxury than an actual need. Something seems right about this judgment. It is possible that what should be considered a need could be somewhat relative. Our sentiments will vary to some extent depending on what it is perceived as 'reasonable' to expect to achieve in a given context with the given level of resources. This echoes the arguments put forward earlier in the discussion about what level of health would constitute an appropriate baseline for measuring health need. Although it is relevant to know what the highest attainable standard of health is, one also ought to consider what kinds of conditions – including the level of provision of health care – will be necessary in order to reach this level of health, and the costs of bringing about such conditions.

Health Care Rationing and the Ranking of Health Care Needs

There is something to be said for each of the proposed definitions of health care need that have been considered so far. But going back to the initial observation that the concept of need is often perceived as the most appropriate guiding principle for the distribution of health (care) resources, one may ask, what would a principle of distributing health care according to need look like on each of these three concepts of health care need?

For the purposes of this discussion, it will be assumed that not all health care needs can be met. How are needs ranked, according to each of the definitions of need? As will become clear each candidate will have different implications for which needs are the greater needs. Assuming that greater needs should be given priority over lesser needs, each definition will imply different strategies for rationing health care resources. Although it is not possible to go into detail for each of the concepts here, a few examples will be pointed out that demonstrate that distribution of resources on the basis of any of these concepts of need on its own will have distributive consequences that are unsatisfactory.

The first definition of health care need that was identified was health care need as the presence of disease. How would needs be ranked on this definition? It is useful to distinguish between a severe and an urgent health state, where severity reflects how poor a health state is, and urgency reflects the

imminence of death. For simplicity, the questions of urgency will be put aside here. If one focuses merely on the severity of a health state, then the greater the health need (i.e., the worse the health state), the greater the need for health care.

Although it seems intuitive that those with the greatest health needs should also have the greatest needs for health care, it is unreasonable to give absolute priority to those with the worst health. The need for treatment can in principle be infinite; one can imagine cases where a health condition is very severe, and incurable, but where medical treatment can nevertheless be of (ever so slight) benefit. In such cases, there is potentially no limit to the amount of health care resources that could be spent in order to improve health, but without fully satisfying and hence eliminating the need. Therefore, a need for health care would remain, no matter how much health care is provided. This is the well-known problem of the bottomless pit. And the bottomless pit problem aside, some increments in health – for example, going from near-complete immobility except being able to wiggle one toe, to being able to wiggle two toes – may simply be too small to be a worthwhile expenditure. But ranking needs for health care entirely on the basis of the severity of the health state cannot accommodate such judgments.

Our second proposed definition of health care need, as capacity to benefit from treatment, avoids this problem. According to this view, need is synonymous with potential for gain; thus, the greater the potential for gain, the greater the need. Naturally, health states that are close to full health do not represent great potential for gain, and thus patients who are not very sick will not be considered to have a great need for health care. Here, the second definition is in agreement with the first definition. But patients who are very sick will only be considered to have a great need for care if effective treatment that can significantly improve the patient's health is available. Considering the example above, it is clear why this definition is so appealing: if there is not much that health care can do, the need for health care is deemed minimal.

However, ranking needs on the basis of who can benefit the most can also be problematic. Consider the following example: Imagine two patients who both need a kidney transplant, but only one kidney is available. Patient A is 30 years old and expected to live for another 40 years after the transplant, whereas patient B is 40 years old and expected to live another 20 years. In this case, allocating the kidney to patient A will yield the greatest health benefit, and therefore patient A is considered to have the greater need. But at least some would object to distinguishing between and ranking the needs of these two patients in this manner; after all, patient B also stands to gain significantly from the kidney transplant. Furthermore, consider a different example: as before, one must decide which of two patients should be allocated a kidney transplant. But in this case, patient C will attain full health after the transplant, whereas patient D will only attain a lower level of health, because this patient also has a permanent disability (which is unrelated to the kidney disease). Say that, on a scale from 0 to 1, where 0 is being dead, and 1 is full health, the kidney disease is rated at 0.3. Without the treatment, both patients have 0.3 in health; although patient D also has a disability, in this case the disability does not 'add' to the severity of the overall health state (this would be true if, e.g., the kidney disease causes you to be constantly hooked up to a dialysis machine, in which case a disability like paraplegia would not add further disadvantage to the overall health state). If paraplegia is rated at 0.7, then patient D would only gain 0.4 (i.e., an increase in health from 0.3 to 0.7) in health as a result of the kidney transplant, whereas patient C would gain 0.7 (i.e., an increase from 0.3 to 1.0). The most effective use of the health resources in this example according to a health maximizing principle would be to allocate the kidney to patient C. This implication is a particularly controversial outcome of ranking needs on the basis of maximizing health benefits.

Finally, these two cases aside, this principle of ranking needs cannot help us distinguish between different health states that have equal potential for health benefit. That is, a patient whose health can be improved from 0.2 to 0.6 will be considered as equally needy as the patient whose health can be improved from 0.5 to 0.9. But here, the severity of the health state would seem to be a relevant consideration for determining which patient has the greatest need for health care; it does not seem right to rank these two patients as having an equally great need for health care, even if their prospective health gain is of the same magnitude.

The last of the proposed definitions of health care need defined need as cost-effectiveness of treatment. That means that the ranking of a need will depend on how much a patient's health will benefit from treatment relative to the cost of that treatment. Even small health gains can be cost-effective, as long as the cost of the intervention is very low. One reason for ranking needs in this manner would be to get as much health as possible with our scarce resources; the money that can be saved by choosing more cost-effective treatments can in turn be used to pay for further treatments. As such there is an overlap with the previous definition of health care needs, which ranking of needs also pushes us to maximize health outcomes.

However, this approach can lead to many small and relatively trivial health gains being ranked as of higher priority than much more substantial health gains. This is exactly what happened in Oregon with the introduction of the Oregon Health Plan in 1990. The prioritization of health care services offered through the Medicaid health plan came about as a result of seeking to extending care to a greater number of people. But in order to achieve this, the system had to be made more cost-effective. An expert group, the Health Services Commission, compiled a list of prioritized health services, on the basis of, amongst other things, the relative cost-effectiveness of different services. The first version of the list ranked tooth-capping as of higher priority than life-saving appendectomy – a controversial result which has been the subject of much commentary and discussion since. Although cost-effectiveness is an important consideration, it seems that focusing solely on this aspect would miss other important considerations.

The discussion thus far has covered the ranking of health care needs in accordance with three different principles: severity of health state, capacity to benefit, and cost-effectiveness of treatment. The discussion has shown that a ranking of needs on the basis of any one of these considerations on its own is unsatisfactory. There is merit to all three of the

considerations, and all of them ought to be taken into account in deliberations on how to allocate our health care resources. Indeed, in practice, most countries will give all three kinds of considerations weight when allocating health care resources. Sometimes those with the worst health will be prioritized, even if a treatment only provides a minor health benefit. Other times it seems important to provide a treatment even if it is not cost-effective to do so. For example, governments sometimes do provide orphan drugs, even if they are not cost-effective.

The question of how such different and often conflicting considerations should be weighed against each other is a task for another day; here, it should simply be noted that such issues cannot be resolved by a stipulative definition of the concept of need. Defining need for health care in terms of one of these considerations does not thereby undermine the force of any of the other considerations.

The question of need aside, there are several other considerations too that are relevant to the distribution of scarce health care resources. Health equity – which is itself an ideal that could be interpreted in many different ways – is one such consideration. For reasons of equity, it might be decided to disregard capacity to benefit or the cost-effectiveness of treatment, and treat like health states alike, regardless of, for example, age or preexisting disabilities. Health care resources can also be rationed with the use of waiting lists or lotteries, without giving particular groups or individual patients greater or lower priority as such. Alternatively, one could interpret health equity as requiring us to reduce inequalities in health outcomes; that could be a reason to prioritize treatment for a patient with worse health, even if the treatment is expensive and of only modest benefit. Desert could be another relevant consideration – it might be decided to give priority in health care to groups that have taken significant risks for the sake of the country, such as firefighters and military personnel.

Conclusion

The notions of need for health and need for health care are clearly important at what one might think of as high-level strategy for resource allocation. If the government announces that it will distribute access to health services on the basis of need, it is clear that it has rejected market-based pricing for services, and will allocate its services according to something like the burden of illness or disease. Yet there is a limit to how much can be done with the concept of need alone. It is not plausible that a health service should allocate services purely on the basis of need. More importantly, however, as the discussion has shown, the concept of need is neither self-evidently clear nor normatively neutral. Defining the concept of need already requires us to take a stand on complex moral questions; one cannot cut through these difficult issues simply by referring to need. 'Distribution on the basis of need' is the name of a social program rather than a principle of distribution, and many different detailed principles of allocation are broadly consistent with a needs-based approach.

See also: Cost–Value Analysis. Disability-Adjusted Life Years. Efficiency and Equity in Health: Philosophical Considerations. Efficiency in Health Care, Concepts of. Health and Its Value: Overview. Quality-Adjusted Life-Years. Welfarism and Extra-Welfarism. What Is the Impact of Health on Economic Growth – and of Growth on Health?

References

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science* **44**, 542–573.

CIA: The World Factbook (2012). Available at: https://www.cia.gov/library/publications/the-world-factbook/index.html (accessed 12.02.13).

WHO (World Health Organization) (2008). *The top ten causes of death, fact sheet 310.* Geneva: WHO. Available at: http://www.who.int/mediacentre/factsheets/fs310/en/index.html (accessed 12.02.13).

Williams, A. (1998). If we are going to get a fair innings, someone will need to keep the score!. In Barer, M. L., Getzen, T. E. and Stoddart, G. L. (eds.) *Health, health care and health economics*, pp. 319–330. New York: Wiley.

Williams, B. (1973). *The idea of equality. Problems of the self.* Cambridge: Cambridge University Press.

Further Reading

Anand, S., Peter, F. and Sen, A. (eds) *Public health, ethics, and equity.* UK: Oxford University Press.

Daniels, N. (2008). *What is the special moral importance of health? Just health: Meeting health needs fairly*, ch. 2, pp 29–78. Cambridge: Cambridge University Press.

Dworkin, R. (2000). *Justice and the high cost of health. Sovereign virtue: The theory and practice of equality*, ch. 8, pp 307–319. USA: Harvard University Press.

Kingma, E. (2007). What is it to be healthy? *Analysis* **67**(294), 128–133.

Marmot, M. (2006). Health in an unequal world: Social circumstances, biology and disease. *Clinical Medicine* **6**, 559–572.

Nordenfelt, L. (2007). The concepts of health and illness revisited. *Medicine, Health Care and Philosophy* **10**(1), 5–10.

WHO (World Health Organization) (1946). Preamble to the constitution of the World Health Organization as adopted by the international health conference, New York, 19–22 June, 1946; signed on 22 July 1946 by the representatives of 61 States and entered into force on 7 April 1948, Official Records of the World Health Organization 2.

Health and Its Value: Overview

E Nord, Norwegian Institute of Public Health, Oslo, Norway, and The University of Oslo, Oslo, Norway

Glossary

Capabilities The set of all possible physical and social functionings for a person.

Cardinal Cardinal measurement in economics has a characteristic that sequences of numbers attached to entities such as health or utility are equivalent measures if they can be related by a simple linear equation such as $X = \bar{a} + bY$. For example, temperature is cardinally measured by either Fahrenheit or Celsius, which are related by the equation $F = 32 + 0.8\ ^{\circ}C$. Distance is commonly measured by ratio scales like miles and kilometers, where the equation is $K/M = 0.6214$.

Decision utility The utility of an option prior to experiencing, consuming it, etc.

Disability-adjusted life-year (DALY) A measure of the burden of disability-causing disease and injury. Age-specific expected life-years are adjusted for expected loss of healthy life during those years, yielding measures of states of health or, when two streams of DALYs are compared, potential health gain or loss by changing from one health care or social intervention to another.

Ex ante A Latin tag meaning a variable as it was before a decision or an event, sometimes used to mean the planned value of a choice variable as in *'ex ante* saving'.

Ex post A Latin tag indicating the value of a variable after a decision or an event, sometimes used to denote the outcome value of a variable as in *'ex post* saving'.

Experienced utility The experienced utility of an episode of illness or wellness is derived experimentally from real-time measures of the attributes of the experience for the subject at the time of the experience.

Perspective The viewpoint adopted for the purposes of an economic appraisal (cost–effectiveness, cost–utility analysis, etc.) that defines the scope and character of the costs and benefits to be examined, as well as other critical features, which may be social value-judgmental in nature, such as the discount rate.

Quality-adjusted Life-year A generic measure of health-related quality of life that takes into account both the quantity and the quality of life generated by interventions.

Well-being An idea related to utility but to be distinguished from health-related quality of life and the inherent 'worth' of people.

What Is Health?

Health is a multidimensional concept. According to a simple definition, one has more health, the more free one is of disease and disability, including being free of diseases at early asymptomatic stages (e.g., high blood pressure and young tumors). In health economics, the concept of health usually refers to observable characteristics such as (1) functionality of bodily organs, (2) ability to move about and do normal activities of daily living, (3) freedom of symptoms in terms of physical discomfort – for example, pain or nausea, and (4) freedom of clinical psychological problems like anxiety disorder, depression, and psychosis.

Health can be viewed as an entity at a given point in time or as an aggregate over a given time period. A person's level of health at a given point in time may be perceived and described in verbal and/or numerical terms along some or all of the above dimensions of symptoms and functioning. This yields a health profile for that person. A number of standardized questionnaires and descriptive systems are available for establishing the health profile of patients. Some of these are disease specific, others are generic, for example, the Sickness Impact Profile and SF-36. Some of the generic ones yield overall index scores that are used in economic evaluation, see below.

Health over a given time period may be understood as an aggregate of the person's health at different stages of that period. If the time period is the future, the aggregate is expected future health and much the same concept as prognosis. If the time period is the whole life, the aggregate is called lifetime health. Both expected future health and lifetime health include longevity (life expectancy).

A description of a person's health over time on one or more dimensions of functioning and symptoms is called a health scenario.

The health of individuals may be used to estimate average, median, or typical health in groups of people, for instance, in a diagnostic group, an age group, a local community or a whole nation. All are examples of estimation of population health.

Both health profiles and scenarios are descriptive entities. They build on measurements of individuals' performance on specific health dimensions, for example, blood pressure, degree of hearing, number of meters one is able to walk without help, score on a pain scale, or score on a depression scale.

The Value of Health

Health profiles and scenarios can be valued. This means judging how good or bad, or how desirable or undesirable they are – all things considered – compared to other possible profiles and scenarios.

 doi:10.1016/B978-0-12-375678-7.00501-0

It is possible to see health as valuable per se, for instance, by regarding good health as something that is the will of God or consistent with a 'natural order.' This would be a deontological view. In health economics, the perspective on valuation is mainly consequentialist: The value of health derives from its positive consequences – or from avoiding the negative consequences of illness.

Consequences of health are of different kinds and may be judged from the viewpoint of different stakeholders.

From individual's personal viewpoint, good health enhances quality of life. This applies both at a subjective and emotional level – in terms of feelings of well-being – and more objectively in terms of capabilities for doing different things and thus opportunities for enjoying a rich life. These are all aspects of health-related quality of life. Good health also enhances longevity and personal income. The personal value of health lies in all these potential consequences.

But individuals' health (or health deficits) may also have consequences for others. Family members may be affected by a person's illness in various ways. Society as a whole may lose production and income as a result of absence from work caused by illness. And communicable disease in one person is potentially harmful to other persons. In short, health has societal value over and above personal value to the individual.

Measuring the Value of Health

In health economics, much attention has been devoted to the value of health for production, i.e., to economic valuation of health from a societal perspective. Key issues in this regard are production losses caused by sick leave and disability and the importance of population health for economic growth.

In personal valuation of health, one main theme is how much individuals are willing to pay out of pocket for improvements in health and for reductions in risks of health losses. Results of research in this area are used as inputs in monetary cost–benefit analyses of health programs.

Another main theme in personal valuation is how highly individuals value life in different states of illness compared to living in full health. In health economics, this is referred to as measurement of health-related quality of life. The quality of life associated with any given health state is expressed as a score on a scale running from zero (corresponding to a state as bad as being dead) to unity (corresponding to being in full health). Alternatively, the scale can be reversed in order to focus on the severity of a state of illness or disability rather than its positive quality. Severity is then expressed as a score running from zero (corresponding to 'no problem') to unity (corresponding to as bad as being dead).

Two different kinds of judgment of health-related quality of life need to be kept apart. One is judgments of own situation made by people with illness or disability. This is often referred to as ex post judgments (judgments made after experience with the illness or disability in question). The other is judgments in samples of the general population of health states that are presented to the subjects as states they might be in. This is often referred to as ex ante judgments (judgments mostly made before experience).

In both approaches, valuations may be elicited at different levels of measurement. Ordinal valuations are verbal reports or crude ratings that allow investigators to rank different health states with respect to value, without saying how much better one state is than another. Cardinal reports allow investigators to compare differences between health states more accurately and say that one difference seems to be X times more valuable than another one.

In health economics, judgments of health-related quality of life at a cardinal level are often referred to as judgments of individual utility. Utility measured as ex ante judgments (in general populations) is called decision utility, whereas utility measured as ex post judgments (in patients and disabled people) is called experience utility.

Research on ex post judgments of health has mainly been conducted by clinicians (physicians, nurses, and others) and by social scientists working more generally with quality-of-life issues. In this research tradition, focus has been on functioning and well-being measured mainly at an ordinal level. But there are also studies of patients' and disabled people's cardinal valuations of the states they are in.

In health economics, research on health-related quality of life has focused mainly on ex ante judgments in general populations. Here, the ambition has been to obtain data with cardinal level measurement properties. For this purpose, various specialized preference elicitation techniques have been developed. Furthermore, various so-called multiattribute utility instruments have been developed that allow investigators to first establish health profiles for patients in question and then translate the profiles into single index estimates of the overall personal value – utility – of the profiles.

The exact interpretation of utility scores for health states is open to debate. On the one hand, they may be understood as the level of personal welfare (subjective well-being, happiness) that individuals derive (or expect to derive) from different states. This interpretation relates to welfare economic theory and is called welfarist. On the other hand, they may be understood as valuations of health itself as judged by some wider criteria that include objective capabilities and levels of functioning. This interpretation is called extra-welfarist.

Utility scores for health states may be multiplied by time spent in the states in question to estimate the aggregate value of health over time for individuals or groups of individuals (including whole populations). The unit of valuation is then 1 year in full health for one individual. This unit is called a quality-adjusted life-year (QALY). Any health scenario may thus be assigned an overall utility in terms of a certain number of QALYs. Similarly, severity scores for health states may be used to estimate the value of aggregate health losses over time.

Health interventions may lead to health benefits both in the present and sometime in the future. Depending on the perspective of the analysis, the value of distant benefits may be considered to be less than the value of benefits that are close in time.

The Utility and Value of Health Care

In health economics, the utility of an intervention for an individual is conventionally estimated by (1) using decision

utilities or severity scores to calculate QALYs or disability adjusted life years (DALYs) and (2) computing the difference between the individual's post- and preintervention health scenario in terms of QALYs or DALYs. The utility of a program for a group of persons is estimated as a sum of the QALY (or DALY) gains of the individuals involved.

Utility estimated in this way is not necessarily the same as the value of care. By their distance from unity, decision utility scores reflect the loss of value – or the 'disutility' – that people in the general population, who mostly are in quite good health, associate with different kinds of health problems. But when people fall ill, their reference points may change. Their valuation of care may then depend on the extent to which the best possible is being done for them, even if they cannot be restored to full health. This source of value is not incorporated in decision utility judgments of health states. Furthermore, societal decision makers' valuations of care for different groups of patients may be affected by various concerns for fairness, for instance, special concerns for the worse off. In sum, there is a difference between expressing health benefits in terms of QALYs or DALYs and valuing care more completely. The importance of the issue is most easily seen in the context of life-saving medicine. All other things being equal, life is better in good health than in less good health. But it does not follow that a person in good health values life itself more – i.e., has a stronger interest in continued life – than a person in less good health. It also does not follow that society as a whole values protection of the former person's life higher than protection of the latter person's life. The value of life itself is not the same as the valuation of health.

See also: Cost–Value Analysis. Disability-Adjusted Life Years. Dominance and the Measurement of Inequality. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Measuring Health Inequalities Using the Concentration Index Approach. Measuring Vertical Inequity in the Delivery of Healthcare. Multiattribute Utility Instruments and Their Use. Quality-Adjusted Life-Years. Welfarism and Extra-Welfarism. Willingness to Pay for Health

Health Care Demand, Empirical Determinants of

SH Zuvekas, Agency for Healthcare Research and Quality, Rockville, MD, USA

Introduction

Economic theory provides a powerful but incomplete guide to the empirical determinants of health care demand. Health economists generally assume that the demand for health care derives from a demand for health. We consume health services as either an investment in our future health, to cope with chronic illness, or recover from acute illnesses or accidents. Rarely medical care is availed simply because it is enjoyed, but we seek it for our health. Health then, along with price (or its proxy health insurance), income, and consumer preferences, play the main role in formal economic models of consumer decision making about health care.

Yet, this theory takes us only so far. Although central to demand, economic theory is agnostic as to how individuals form preferences. It is suspected that individuals vary in attitudes and preferences toward risk, willingness to trade-off better health tomorrow for increased consumption today, and likes and dislikes. As a result, people with the exact same economic resources may respond differently to the same medical circumstances. However, these individual preferences are almost never directly observed – they are difficult to measure outside of controlled experiments. Health care prices and income too are almost never observed in the way we would like.

However, empirical studies consistently demonstrate that a wide range of sociodemographic characteristics including age, ethnicity, sex, and education are strongly correlated with health care use. These rarely appear directly in the theoretical models. Sociodemographic characteristics are intended to be included in empirical models because it is believed that they are correlated with otherwise unmeasured preferences toward health or capture dimensions of a person's health (e.g., age), or both. However, interpretation is difficult because these proxy measures are often confounded with so many different unobserved aspects of individuals and their environment.

Economic theory also posits that health care demand is jointly determined with supply. In practice, empirical models of consumer demand are almost never jointly estimated with models of supply because of data limitations. Instead, these empirical models assume that observed health care use is equal to consumer demand (technically, short-run supply is assumed to be perfectly elastic at the margin). Researchers sometimes add measures of provider supply and other market characteristics on an ad hoc basis to individual characteristics in empirical models of demand. However, interpretation here is problematic. For example, physicians tend to locate in areas with high demand. Thus, a measure of supply, like physicians per capita, will tend to reflect back demand rather than being a causal determinant of demand.

How then do we decide which determinants to include in our empirical models of health care demand? And how do we interpret them? The purpose of this article is to provide guidance to both questions. The discussion begins by introducing some general rules of thumb. Although theory is by no means definitive, economic principles can still be appealed to in understanding the relationships among the theoretical and proxy determinants. Statistical principles also play a role. Overall, competing concerns about usefulness of particular variables as predictors of health care use and the potential biases they introduce must be confronted. A brief survey of the recent literature is next provided to give a flavor of the range of determinants commonly included in recent empirical studies of demand. Finally, a representative empirical example of health care demand to more systematically illustrate the selection, use, and interpretation of empirical determinants has been developed. Because price and income are covered well in separate articles, focus will be on the primary demographic, social, and above all, health characteristics that determine health care demand from the consumer point of view.

Some Rules of Thumb

In its strictest sense, 'determinant' implies causation. Causal interpretation of this theory based observable determinants of health care demand, price and/or insurance, income, and health status is threatened by what is termed as endogeneity bias because they are jointly determined with health care use. First and foremost, the concern is about bias due to adverse selection – those with a greater need or preference for treatment will be more likely to purchase or enroll in insurance coverage. To the extent that health and preferences are unmeasured, their role will be misattributed as determinants thereby overestimating the effect of insurance. As such, this form of endogeneity bias can also be thought of as omitted variable bias. Omission of relevant variables, in this case unmeasured preferences and health status, biases all other variables correlated with it. Endogeneity bias also arises from reverse causality. For example, health care ideally improves health, so that if health status is measured after care is received, the effect of health on demand is underestimated.

Even without reverse causality, postdiction bias might arise when observing health status after treatment occurs. For example, a health condition might develop after an unrelated visit to a doctor, but an empirical model including this post-visit condition will incorrectly attribute some of the reasons for the visit to this condition. One way to minimize postdiction bias is to use the earliest measurement of health status (or other determinant) possible, but this risks an opposite measurement error problem.

Most other determinants included in empirical demand models serve primarily as proxies for unmeasured aspects of individual's health or preferences toward health and health care. Use of proxies is a valid method for including what we think are important unmeasured determinants of demand. But care must be exercised both in the choice of proxies and in

their interpretation because of the obvious omitted variables bias issues that arise, as well as the potential for reverse causality.

Deciding which determinants to include in empirical demand models and then specifying how they are used and interpreted requires balancing often competing objectives and biases. Here a few general guidelines, rules, or thumb for selecting and interpreting empirical determinants have been provided. It is emphasized that these are not hard and fast rules. Reasonable researchers may differ in their beliefs about biases and as a result, make different decisions. Bias generally arises from something unobserved about individuals or their actions making it almost impossible to quantify the true extent of any particular bias.

Rule 1: Include Theoretically Important Demand Determinants Where Possible

The theoretical models of health demand states that price (or its proxy, health insurance), income, and health are primary drivers of health care demand. Therefore it should be sought to use them wherever possible. Preferences, the other main theoretical determinant, are generally unobserved and proxies must be relied on (Rules 3 and 4).

Rule 2: Minimize Bias in Choice Variables

Also, the theoretical models of health demand states that the theoretical determinants are jointly determined with health care use and thus potentially endogenous. Five options are there:

1. Less endogenous versions of determinants should be used. For example, prior year observations of a potentially endogenous variable such as health or health insurance should be used.
2. The amount of endogeneity due to omitted variables bias should be reduced. For example, including better measures of health status can reduce the bias in the effect of health insurance.
3. Econometric techniques to reduce or eliminate endogeneity bias should be used.
4. Potentially endogenous variables are to be used if (1) and (2) are not sufficient. However, interpreting the results is important.
5. Endogenous determinants should be dropped as the last resort.

The choice between (4) and (5) weighs the endogeneity bias of including a determinant against the omitted variables bias introduced by omitting it.

Rule 3: Include Exogenous Proxies

Age, race, ethnicity, and almost always, sex, are thought to be fixed or exogenous characteristics of an individual. That is, they do not depend on our choices of health care use or other determinants and they are not subject to reverse causality. Thus, they serve as excellent proxies for unmeasured health, particularly age, and also for unmeasured preferences. However, as proxies correlated with multiple omitted characteristics of individuals, they generally cannot be interpreted as causal determinants.

For other potential proxies for health and health preferences, it is a matter of degree. What matters most is how these other potential determinants are related to our own choices about health and health care, and the extent that reverse causality is an issue. For example, education clearly depends on individual choices. However, in most contexts we can still treat it as fixed. For example, the choices a 75-year old made about their education 50 or 60 years ago are unlikely to have been closely related to their health and health care use today. However, poor childhood health or a catastrophic illness in late adolescence or early adulthood such as schizophrenia or Crohn's disease could easily affect educational success.

Rule 4: Balance Competing Concerns with Potentially Endogenous Proxies

For demographic or other candidate determinants that are not fixed and subject to bias, several competing concerns must be considered in deciding whether and how to use them: (1) importance as either a direct determinant or proxy determinant of demand; (2) extent of potential endogeneity and/or reverse causality bias; and (3) extent of the omitted variable bias created by excluding the determinant. A potential determinant with uncertainty about the connection to individual decisions about health care use and a high potential for bias probably is not a good choice.

Empirical Determinants of Health Care Demand: A Survey of Current Practices

Here the health status, economic, and socioeconomic determinants included in recent empirical studies of the demand for health care are surveyed. The survey includes 98 empirical studies published over the 12-year period 2000–2011 in the *Journal of Health Economics* and *Health Economics* that estimate the demand or use of health care services using individual or household level data derived primarily from household surveys. A few studies based strictly on claims or administrative data have been excluded because of the limited information about individuals.

The survey is not meant to be exhaustive. However, the 98 articles are broadly representative of empirical studies published in economics, health services research, and medical journals. They are based on a number of different household surveys and cover a broad range of low, middle, and high-income countries. None attempt to estimate a full structural model of all the joint choices that the theoretical health care demand models describe. Together, they give a sense of the range of determinants typically included in health care demand models and how they are used.

Few studies provide explicit rationales for each determinant. Most divide determinants into 'need' variables, measures of health and proxies such as age, and nonneed variables. In the context of economic models of consumer demand, these 'need' variables are simply inputs into an individual's

decisions. Others may think that an individual with a particular disease, say diabetes, needs treatment, but it is the individual that determines their own demand and whether they seek treatment. This issue of need and demand will be discussed later in an empirical illustration. A few studies appeal to an alternate framework in the selection of determinants, the Andersen–Aday behavioral model of health care use. The Andersen–Aday framework is less a formal behavioral model in the way economists use the term and more a catalog of characteristics correlated with health care use.

Economic Determinants

Table 1 provides a summary of the types of determinants included and how often they appear. Among the economic variables, consumer price appears in only 16 of the 98 demand models and even then is generally only partially observed. However, health insurance coverage is widely used as a proxy. Together, 64 of the 98 studies included price and/or health insurance coverage. Price and health insurance coverage were the clear focus of researchers' concerns with bias. Of the 64 studies including either of these determinants, 28 used either experimental data or econometric methods specifically designed to reduce bias. Rarely did researchers use econometric methods to specifically tackle bias in health (four studies) or other determinants.

Table 1 Determinants of health demand, frequency of use in survey of 98 recent empirical studies

Price	16/98
Health insurance	58/98
Time price	13/98
Income	89/98
Wealth/assets	19/98
Employment/main activity	
Employment status	38/98
Occupation and/or industry	18/98
Age	98/98
Life expectancy/time to death	5/98
Sex	97/98
Race and/or ethnicity	49/98
Immigration or citizenship status	16/98
Marital/partner status	69/98
Household size and/or composition	63/98
Educational attainment	86/98
Geographic indicators[a]	65/95
Trend[b]	42/45
Health status	91/98
Self-assessed health	60/98
Scale	18/98
Chronic conditions	65/98
Obesity/body mass index	10/98
Functional limitations and disability	42/98
Acute illness	23/98
Prior utilization	7/98
Other	21/98
Health behaviors (smoking, alcohol, drug, exercise, and diet)	31/98
Health beliefs and preferences	7/98
Health information	4/98
Environmental risks	3/98
Access to regular doctor	6/98
Supply side characteristics	
Physician supply	17/98
Distance to provider	8/98
Provider quality	8/98
Market characteristics	12/98

[a]Studies based on single location (city) excluded from denominator.
[b]Studies based on single cross-section excluded from denominator.
Source: Author's review of 98 empirical studies of health care use or health care demand based on survey data appearing in the *Journal of Health Economics* and *Health Economics* between 2000 and 2011.

Appealing to the economic notion that the price of consuming health care extends beyond direct out-of-pocket costs to time, 13 of the 98 studies included time price. A typical direct measure of time price multiplies a person's wage rate by the time they spend traveling to health care and wait time. In other cases, proxy measures such as travel time were used. Measures of income were almost always included (89 studies). In most cases, this was total family income divided by the square of the number of household members. This normalization assumes that the larger the family, the smaller the share of resources available to any one member. In a handful cases, wealth or household assets were used as a proxy for income. In 16 studies, wealth or assets were included in addition to income. The theory here is that consumers consider more than just current income.

Although widely available in surveys, employment status (38) and/or information about occupation or industry (18) were included in a minority of studies. These are choice variables that do not directly determine health as per the theoretical models of health demand. A major concern here is reverse causality. Poor health might lead to job loss, and thus employment-related variables will reflect some aspects of health. Reasons for including employment characteristics might be that certain industries and occupations carry greater health risks from accidents, exposure to hazardous materials, or stress. They might also proxy for preferences about health care. In the United States, industry, occupation, and firm size are also correlated with generosity of insurance and access to paid sick leave.

Health and Health-Related Determinants

Direct measures of individual health were included in 91 of the 98 studies. They were uniformly powerful predictors of use. Reflecting the multidimensional nature of health, a wide range of measures was used. The most common were chronic health conditions such as diabetes, asthma, and heart disease (65 studies). Some studies used counts of chronic conditions, others each individual condition. The next most common health status measure were variants of a single-item self-assessed health status scale (60) asking respondents to rate their health (e.g., excellent, good, fair, and poor). A number of studies (42) included measures of disability or functional limitations such as difficulty walking upstairs or lifting. These were more common in studies of older populations. Other studies included measures derived from longer health-rating scales (18), acute illness or symptoms such as fever (23), measures related to obesity (10), and a variety of other measures (21).

All measures of health status raise concerns about reverse causality. Choices of which measures to include are driven by a combination of availability and an individual researcher's beliefs about the trade-offs between relevance in determining demand and potential bias. Chronic conditions are appealing because their very nature makes them less prone to concerns that current health care use changes whether an individual has the condition or not. For example, once you have diabetes you always have diabetes; treatment manages symptoms. Bias is a greater concern with acute symptoms of illness explaining why they are less commonly used, even though it is believed that they drive much use. Acute illness was more commonly included in studies based in low-income countries, where the balance between bias and relevance may be different.

Prior health utilization is a strong predictor of current health care use, and seven studies appeared to include it for this reason. If we are simply interested in obtaining the best possible predictions of current health care use, this is fine. However, if we are interested in the extent to which health explains health care use, it is not fine. For example, if a person has consumed lots of health care in the previous year because of his/her diabetes, then the intensity would also continue the year ahead. Henceforth, the effect estimated for diabetes is diminished. By including last year's use, the estimate for diabetes is greatly diminished.

Seven studies included no measures of health status. In two of them, none were available. The remainder explicitly or implicitly excluded health because of potential endogeneity. The researchers are making the call that the bias introduced by including health is worse than the omitted variables bias created by excluding a powerful determinant. Most make the call the other way.

Another set of health-related determinants commonly included are smoking, alcohol and drug use, exercise, diet, and other health behaviors (31 studies). Many researchers exclude them because they are choices related to health. However, they are also often strongly related to use. For example, a history of smoking can lead to significant health problems today even if one does not currently smoke.

Access and Supply-Side Determinants

Six studies included whether a person has access to a regular medical provider. Although often available, most researchers omit this as a choice variable because of its clear endogeneity with health care use. A larger number of studies included local area physician, hospital or other provider supply (17), distance to provider (8), or market characteristics (12) such as managed care concentration. These are generally used as proxies for availability and access to providers. Some may also proxy for time price. Many researchers omit these because observed health care use is a simultaneous function of supply and demand (though rarely modeled this way) making supply-related variables endogenous.

Demographic Determinants

Among the demographic variables, age and sex appear universally. Five studies used life expectancy (or time to death) in addition to or in place of age, arguing that life expectancy is a stronger predictor of health care use. A number of studies included interaction terms between age and sex, allowing for the effects of age to vary with sex, and vice versa. A few ran separate models stratified by sex, allowing for the effects of all determinants to vary by sex. Measures of race, ethnicity, or cultural group were included in 49 of the studies. Studies without such measures tended to come from countries with more homogenous populations. Education status (number of years or degrees) was almost universal. Similarly, most included some combination of marital status and/or household composition (e.g., number of family members). More than two-thirds of the studies included geographic indicators such as locality and/or living in an urban or rural area.

Finally, almost all the studies that used data pooled across more than 1 year included some time or trend dimension in the model. With health care use generally increasing over time, it is important to capture overall shifts in health care demand.

Other Considerations

Table 1 describes the range of determinants included in empirical demand studies. In aggregate, some are used more regularly than others. The table does not capture that the studies reviewed varied considerably in how parsimonious or expansive the set of demand determinants included in each study were. This ranged from as few as six variables to as many as 87. In some cases, parsimony may be driven by computational demands. Some econometric methods are also fragile when including too many closely related variables. Often, though, individual researchers simply prefer more parsimonious models. If we are interested in only one or two determinants and a model with just a few variables captures these well, we may be okay omitting other potential determinants. However, if the omitted characteristics are correlated with these key determinants, our estimates will be biased.

Empirical Determinants of Health Care Demand: An Illustration

The important roles that various theoretically derived and proxy determinants play in empirical models are illustrated using an example drawn from the author's own work on the demand for mental health treatment. Specifically, treatment related to depression are examined. Depressive disorders are a group of chronic, but episodic diseases affecting millions of Americans. The effects that health, economic, and sociodemographic determinants have on the use of three treatment options for the treatment of depression are examined: nonspecialty visits (generally to primary care providers), specialty visits (psychiatrists, psychologists, and social workers), and antidepressant medications. Aside from obvious convenience, this example has been chosen to illustrate how empirical determinants can vary across different types of treatment. This depression treatment example also conveniently illustrates the difference between need for treatment and individual demand.

Being consistent with the literature, it is not attempted to jointly estimate demand and supply. The dependent variable in the models is observed utilization and is assumed to be equivalent to demand. The empirical example in other respects is simplified. First, using probit equations; it is only modeled whether a person used each type of treatment and not quantities. Second, it is not attempted to jointly model other aspects of consumer decision making (e.g., other goods and services, income, and employment). Third, the main estimates presented do not correct for potential endogeneity of health insurance, income, and health status.

The rules of thumb described above can help guide both selection of empirical determinants and their interpretation using this example.

Data

The data are drawn from the Medical Expenditure Panel Survey (MEPS), a large nationally representative household survey conducted annually in the United States since 1996 by the Agency for Healthcare Research and Quality. The MEPS contains a rich array of information on each household member's health care use and expenditures, health insurance coverage, employment and income, health status and health conditions, and other sociodemographic characteristics. The MEPS is widely used to model the demand for health and to plan and evaluate health policy reforms and changes.

The MEPS utilizes an overlapping panel design to represent the civilian noninstitutionalized population in each calendar year. Households are interviewed in-person for five rounds covering 2 full calendar years. The average recall period for these five rounds is approximately 5 months. Generally, one person responds for all members of the household. In-person interviews are supplemented with self-administered health questionnaires (SAQs) of every adult to assess health status and experiences of care that might not be reliably captured by proxy. Follow-back surveys of physicians, hospitals, home health agencies, and pharmacies are used to collect more detailed information on health care spending and prescription medications.

Current sample sizes for each panel are approximately 7500 households and 18 000 individuals. The analytic sample used here is drawn from the 2004–08 panels of the MEPS and includes 37 173 adults aged 18–64 with two observations each with complete information on treatment use, depression status, and other covariates.

Analyses

Table 2 presents means of the dependent variables and all demand determinants for the full sample and also stratified by an indicator for probable depression. Departing from standard practice, the specification of each demand determinant and its rationale for inclusion one by one with the results and interpretation from the empirical demand model are described.

Economic Determinants: Specification, Results, and Interpretation

Table 3 presents the empirical estimates of the effects of economic, health status, and sociodemographic determinants from three probit equations describing any nonspecialty, specialty, and antidepressant use. The table adds a fourth column, which computes the combined effect of each determinant on the use of any of the three types of treatment. To ease in the interpretation of magnitudes, marginal effects are presented instead of coefficient estimates. For binary indicators, the marginal effects represent the change in the expected probability of using treatment for that group compared to the omitted group. For example, the marginal effect of 0.022 females for nonspecialty services implies that women are 2.2% points more likely to have nonspecialty mental health visits than men. The overall mean use of nonspecialty treatment is 5.9%, men and women combined, so this represents a substantial differential. For continuous measures, the marginal effect represents the change in the probability of use for a one unit change. For example, each additional child in a household less than 6 years (marginal effect = 0.006) is associated with a 0.6% point decrease in the use of nonspecialty care.

Health insurance

Like many surveys, price is only observed in the MEPS among users. Deriving theoretically consistent prices suitable for demand estimation from these partial observations is conceivable but difficult. Health insurance is used as a proxy instead. The MEPS contains extensive insurance coverage information. For simplicity, a three category summary of insurance status provided on the MEPS public use file (INSCOV) is used. In the sample of adults aged 18–64 years, 23% were uninsured the entire calendar year, 64% had private insurance (mainly through employers or unions) for all or part of the year, and 13% had public insurance only, mainly Medicaid or Medicare (Table 2).

It is seen that private and especially public insurance are strongly correlated with treatment. For example, people with public insurance are 10.2% points more likely to use any type of treatment than people without insurance. Although we expect positive effects of insurance on use, there are reasons to believe the estimated magnitudes are too large. First and foremost is adverse selection. Second, public health insurance may proxy in part for unmeasured severity of depression because both Medicare and Medicaid, in part, serve as disability programs. The qualifying process itself, which includes clinician diagnoses, may differentiate between levels of depression in ways that move beyond a limited depression scale. Using first month insurance indicators instead of full-year insurance to minimize postdiction bias does little to magnitudes. However, when the model is reestimated explicitly accounting for the potential endogeneity of insurance, the estimated effects of public and private insurance drop by half (not shown).

Income

Income is included as a theoretically important determinant but discussion on its interpretation has been brief. Following common practice, the log of total family income is divided by the square root of the number of household members. Positive income effects are generally expected, but for antidepressant use, only a small effect is observed. Income may be confounded with unobserved depression severity and other

Table 2 Descriptive means, adults aged 18–64, 2004–09 pooled MEPS sample

	Full sample 100% (n = 74 346)	*Probable depression PHQ-2 ≥ 3 10.1% (n = 7526)*	*Below threshold PHQ-2 < 3 89.9% (n = 66 820)*
Any treatment use			
Any nonspecialty provider (0, 1)	0.059	0.193[c]	0.044
Any specialty provider (0, 1)	0.045	0.161[c]	0.032
Any antidepressant fills (0, 1)	0.110	0.324[c]	0.086
Any treatment (0, 1)	0.144	0.397[c]	0.116
Level of use conditional on use			
Number of nonspecialty visits	4.01	4.93[c]	3.56
Number of specialty visits	8.30	9.40[c]	7.68
Number of antidepressant fills	7.20	8.15[c]	6.80
Health insurance coverage			
Any private (0, 1)	0.64	0.41[c]	0.67
Public only (0, 1)	0.13	0.34[c]	0.11
Uninsured (omitted)	0.23	0.24[c]	0.22
Family income			
Log family income	10.09	9.40[c]	10.16
Physical health status			
Chronic conditions (0–11)	0.67	1.30[c]	0.60
SF-12 physical component summary	50.49	42.15[c]	51.43
Poor/fair physical health (0, 1)	0.20	0.55[c]	0.16
Mental health status			
Poor/fair mental health (0, 1)	0.11	0.46[c]	0.07
PHQ-2 score (0–6)	0.79	4.21[c]	0.40
Age			
19–29 (0, 1)	0.23	0.19[c]	0.23
30–44 (0, 1)	0.36	0.33[c]	0.36
45–54 (0, 1)	0.24	0.28[c]	0.24
55–64 (omitted)	0.17	0.20[c]	0.17
Sex			
Female (0, 1)	0.54	0.63[c]	0.54
Male (omitted)	0.46	0.37[c]	0.46
Race/ethnicity			
Hispanic (0, 1)	0.26	0.25	0.26
Black (0, 1)	0.17	0.22[c]	0.16
Other (0, 1)	0.06	0.05[c]	0.06
White (omitted)	0.51	0.48[c]	0.52
Marital status			
Not currently married (omitted)	0.43	0.56[c]	0.42
Married (0, 1)	0.57	0.44[c]	0.58
Household composition			
0–5 Years old	0.34	0.30[c]	0.34
6–17 Years old	0.67	0.65[a]	0.68
18–64 Years old	2.12	1.97[c]	2.13
65 or older	0.07	0.09[c]	0.07
Education status			
Less than high-school diploma (omitted)	0.22	0.35[c]	0.21
High-school diploma (0, 1)	0.32	0.35[c]	0.31
Some college (0, 1)	0.23	0.19[c]	0.24
Bachelor's (0,1)	0.14	0.07[c]	0.15
Advanced degree (0,1)	0.09	0.03[c]	0.09
Census region			
Northeast (0, 1)	0.14	0.13[b]	0.15
Midwest (0, 1)	0.20	0.18[b]	0.20
South (0, 1)	0.39	0.43[c]	0.38
West (omitted)	0.27	0.25[b]	0.27
Urban/rural			
Non-MSA (omitted)	0.16	0.20[b]	0.16
MSA (0, 1)	0.84	0.80[b]	0.84
Self or Proxy			
Self (omitted)	0.54	0.62[c]	0.53
Proxy (0, 1)	0.46	0.38[c]	0.47

(*Continued*)

Table 2 Continued

	Full sample 100% (n = 74 346)	*Probable depression PHQ-2 ≥ 3 10.1% (n = 7526)*	*Below threshold PHQ-2 < 3 89.9% (n = 66 820)*
Trend			
Trend	3.53	3.48[b]	3.54
Trend squared	14.85	14.55[a]	14.88

[a]Difference between probable depression and below depression threshold significant at $p < .10$.
[b]Difference between probable depression and below depression threshold significant at $p < .05$.
[c]Difference between probable depression and below depression threshold significant at $p < .01$.
Abbreviations: MEPS, medical expenditure panel survey; MSA, metropolitan statistical area; SF-12, Short Form-12.
Notes: The method of balanced repeated replications was used to correct all standard errors and statistical tests for the stratified and clustered design of the MEPS. This method also corrects for the correlation across individuals and families.
Source: Author's Calculations from 2004–09 Medical Expenditure Panel Survey, Agency for Healthcare Research and Quality, Rockville, MD.

factors. For example, depression often leads to job loss, thereby biasing downward the effects of income.

Health and Mental Health Determinants: Specification, Results, and Interpretation

Physical health

Three widely used measures of physical health are included from an earlier review. A strong correlation between depression and physical health has long been observed but the causal pathways remain unclear. Certain medical conditions, for example, heart attack, may lead to or exacerbate depression. Or patients might simply be depressed about physical ailments, especially if they lead to job loss or other life changes. On the other side of the equation, depression may lead to poor diet and exercise. In the course of treating people for physical ailments, providers might also detect depression leading to more care.

The first measure is a simple count of a set of 11 chronic conditions that are ascertained in each MEPS panel. Respondents are asked if the doctor ever told the person they had diabetes, arthritis, asthma, emphysema, stroke, high blood pressure, high cholesterol, coronary heart disease, heart attack (myocardial infarction), angina, and any other heart disease. A graph of the 0–11 condition count versus treatment was approximately linear (not shown). The regression results in Table 3 show a strong association between chronic conditions and antidepressant use in particular, with each additional condition increasing the probability of antidepressant use by 1.8% points. The association is used here because, even with reverse causality minimized, these chronic conditions are still likely correlated with depression severity, not captured in the depression index.

The second measure is the physical health summary score from the Short Form-12 (SF-12) contained in the MEPS Adult SAQ asked in the middle to later part of each calendar year. The SF-12 is a well-validated health inventory containing 12 questions on a number of dimensions of physical and mental health symptoms and functioning. This composite index is scaled from 0 to 100 and normalized to approximately 50 with a higher score indicating better health. The effect on nonspecialty use was not significant. Better physical health is associated with a reduced probability of antidepressant use as expected. Curiously, better physical health, controlling for chronic conditions and perceived health status, is associated with a small but statistically significant increase in specialty use.

The third physical health measure is derived from the standard 1-item perceived health status question asked in each of the five rounds of MEPS. Respondents are asked relative to persons their age, whether each member of the household is in excellent, very good, good, fair, or poor health. The poor and fair responses in either of the first two rounds during a calendar year into a single binary indicator (ever poor or fair vs. good/very good/excellent) have been combined. Turning to the actual results in Table 3, there is an independent effect of poor or fair perceived health on treatment, increasing the likelihood of nonspecialty visits by 1.2% points and antidepressant use by 1.3% points.

The SF-12 and poor/fair health measures bring the potential for obvious reverse causality problems because they are measured contemporaneously with treatment. In fact, they could be measured well after treatment if treatment occurred earlier in the year. Using the strategy of minimizing postdiction by measuring health at the earliest possible point during the year or using prior year values, alternative ways of constructing and using these variables have been tested. For the poor/fair measure, the round 1 responses are used only to construct an alternate poor/fair indicator. This had no appreciable effect on magnitudes of the effects. Because the SF-12 is measured later in the year, the first year of each person's observations has been discarded but used their SF-12 (and poor/fair health status) from the first year to estimate the demand models on the second year's observations. Again, nothing changed. Rather than lose half the observations, it has been opted to keep the models as they are.

Interpretation of all three physical health status measures is uncertain because they are likely associated with unmeasured aspects of health and preference. To test this, a version of the demand models presented here has been estimated, which explicitly accounts for these potential correlations. The results (not shown) suggest that the physical health measures indeed are correlated with unmeasured aspects of people's health and preferences toward care, substantially reducing the magnitudes of the observed effects of the three measures.

The MEPS contains a number of other measures related to functional limitations and disability, recent symptoms associated with chronic and other diseases, measures of work or

Table 3 Estimated marginal effects of economic, health, and demographic determinants from probit models of treatment demand

	Any nonspecialty (mean = 0.059)	*Any specialty (mean = 0.045)*	*Any antidepressant (mean = 0.110)*	*Any treatment (mean = 0.144)*
Health insurance coverage				
Any private	0.021 (0.002)[c]	0.018 (0.002)[c]	0.050 (0.004)[c]	0.059 (0.004)[c]
Public only	0.046 (0.006)[c]	0.062 (0.005)[c]	0.081 (0.008)[c]	0.102 (0.007)[c]
Uninsured (omitted)				
Family income				
Log family income	−0.0005 (0.0006)	−0.0003 (0.0005)	0.0025 (0.0009)[b]	0.0015 (0.0010)
Physical health status				
Chronic conditions	0.007 (0.001)[c]	0.003 (0.001)[c]	0.018 (0.001)[c]	0.019 (0.002)[c]
SF-12 Physical component summary	−0.0001 (0.0001)	0.0002 (0.0001)[b]	−0.0009 (0.0001)[c]	−0.0006 (0.0002)[c]
Poor/fair physical health (0, 1)	0.012 (0.003)[c]	0.002 (0.002)	0.013 (0.004)[c]	0.018 (0.005)[c]
Mental health status				
Poor/fair mental health (0, 1)	0.082 (0.005)[c]	0.097 (0.005)[c]	0.117 (0.006)[c]	0.183 (0.008)[c]
PHQ-2 score	0.013 (0.001)[c]	0.011 (0.001)[c]	0.026 (0.001)[c]	0.033 (0.001)[c]
Age				
19–29	0.003 (0.003)	0.004 (0.003)	−0.034 (0.004)[c]	−0.022 (0.005)[c]
30–44	0.013 (0.003)[c]	0.013 (0.003)[c]	−0.002 (0.004)	0.013 (0.005)[b]
45–54	0.008 (0.003)[b]	0.008 (0.003)[b]	0.009 (0.004)[b]	0.016 (0.005)[c]
55–64 (omitted)				
Sex				
Female	0.022 (0.002)[c]	0.010 (0.002)[c]	0.056 (0.003)[c]	0.061 (0.004)[c]
Male (omitted)				
Race/ethnicity				
Hispanic	−0.021 (0.002)[c]	−0.018 (0.002)[c]	−0.058 (0.003)[c]	−0.067 (0.004)[c]
Black	−0.034 (0.002)[c]	−0.022 (0.0021)[c]	−0.077 (0.003)[c]	−0.092 (0.004)[c]
Other	−0.032 (0.003)[c]	−0.024 (0.003)[c]	−0.079 (0.004)[c]	−0.094 (0.006)[c]
White (omitted)				
Marital status				
Not currently married (omitted)				
Married	−0.007 (0.002)[c]	−0.012 (0.002)[c]	0.005 (0.003)[a]	−0.006 (0.004)[a]
Household composition (number of household members)				
0–5 Years old	−0.006 (0.002)[c]	−0.006 (0.002)[c]	−0.007 (0.003)[b]	−0.012 (0.003)[c]
6–17 Years old	−0.003 (0.001)[b]	−0.005 (0.001)[c]	−0.002 (0.001)	−0.006 (0.002)[c]
18–64 Years old	−0.003 (0.0011)[b]	−0.005 (0.0011)[c]	−0.007 (0.002)[c]	−0.010 (0.002)[c]
65 or older	−0.009 (0.004)[b]	−0.003 (0.003)	−0.004 (0.005)	−0.010 (0.006)
Education status				
Less than high-school diploma (omitted)				
High-school diploma	0.004 (0.003)	0.007 (0.003)[b]	0.014 (0.004)[b]	0.016 (0.005)[c]
Some college	0.011 (0.003)[c]	0.021 (0.004)[c]	0.030 (0.005)[c]	0.041 (0.006)[c]
Bachelor's	0.009 (0.004)[b]	0.036 (0.005)[c]	0.031 (0.006)[c]	0.048 (0.007)[c]
Advanced degree	0.013 (0.004)[b]	0.059 (0.007)[c]	0.029 (0.006)[c]	0.061 (0.007)[c]
Census region				
Northeast	−0.004 (0.004)	0.009 (0.003)[b]	0.001 (0.001)	0.003 (0.006)
Midwest	−0.008 (0.003)[b]	0.002 (0.003)	0.009 (0.005)[a]	0.003 (0.006)
South	−0.010 (0.003)[b]	−0.005 (0.003)[a]	0.003 (0.005)	−0.005 (0.005)
West (omitted)				
Urban/rural				
Non-MSA (omitted)				
MSA	0.002 (0.003)	0.011 (0.002)[c]	−0.004 (0.004)	0.003 (0.004)
Self or Proxy				
Self (omitted)				
Proxy	−0.010 (0.002)[c]	−0.009 (0.002)[c]	−0.014 (0.003)[c]	−0.021 (0.003)[c]
Trend				
Trend	0.001 (0.002)	0.002 (0.002)	−0.005 (0.003)	−0.003 (0.004)
Trend squared	−0.0003 (0.0004)	−0.0003 (0.0003)	0.0006 (0.0005)	0.0002 (0.0005)

[a] $p < .10$.
[b] $p < .05$.
[c] $p < .01$.

Abbreviation: MSA, metropolitan statistical area.

Notes: Trivariate probit model of any nonspecialty visit, any specialty visit, and any antidepressant fill estimated by simulated likelihood using the Stata routine MVPROBIT. Estimated correlation between nonspecialty and specialty visits is 0.501 (0.017), between nonspecialty and antidepressant use is 0.678 (0.016), and between specialty and antidepressant use is 0.654 (0.018). Any treatment is computed at the union of any nonspecialty visit, any specialty visit, and any antidepressant fill using the estimated multivariate normal distribution. Standard errors in parentheses computed using the method of Balanced Repeated Replication (128 half replicates using a Fay's adjustment of 0.5) which accounts for the stratified and clustered design of the MEPS and correlation in observations across families and individuals.

Source: Author's Calculations from 2004–09 Medical Expenditure Panel Survey, Agency for Healthcare Research and Quality, Rockville, MD.

school days lost and bed days, and a number of other adult health measures as well as measures specific to children and adolescents. Good arguments could be made for including any one of a number of them. The main reason for sticking with just chronic conditions, SF-12, and perceived health status is parsimony. Together they do a reasonable job of representing physical health and capture many of the same dimensions of the other measures in this context.

Mental health status

Conceptually, we might think mental health is the most important determinant of demand. If you are depression free, why seek treatment? Thus, the 2-item Patient Health Questionnaire (PHQ-2), a well-validated depression screener taken from the Adult SAQ, is included. The PHQ-2 asks "Over the last 2 weeks, how often have you been bothered by any of the following problems?" "Feeling down, depressed, or hopeless," and "little interest or pleasure in doing things." Responses ranged from "not at all" (0) to "nearly every day" (3). A score of 3 or higher is suggested as a cut-point for depression screening. The linear PHQ-2 scale (0–6) is used because it measures both probable clinical depression (PHQ-2 $\geq$ 3) and severity. For example, each increment in the 0–6 scale is associated with a 2.6% point increase in antidepressant use.

The mental health analog of perceived physical health status is also included. Even controlling for symptoms of depression, we find that perceived poor or fair mental health increases nonspecialty use by 8.2% points, specialty use by 9.7% points, antidepressant use by 11.7% points, and any treatment by 18.3% points compared to those with better mental health.

Two concerns with the PHQ-2 and perceived mental health measures are noted. Most importantly, the first information we get about depressive symptoms occurs later in the first year a person enters the MEPS, and the PHQ-2 scale asks only about symptoms in the past 2 weeks. If a person sought treatment in the past because of depression, assuming treatment works, he/she may be symptom free by then. This will tend to reduce the impact of depression estimated. Like physical health, an alternative has been tested using only the second year of data for each person substituting their first year PHQ-2 measurement and perceived mental health status. Surprisingly, no appreciable differences in the effects on treatment use have been found. This may be because, although depression is a chronic illness, it is also episodic. It is also possible that reverse causality bias is offset by people with depression in the first year who do not carry symptoms into the second year and do not need treatment. However, as with physical health status, when the demand models were reestimated to explicitly account for endogeneity bias, the effects of mental health status were substantially reduced.

Second, although the PHQ-2 does a nice job for a two-item scale, it is not as sensitive to depression severity as its longer cousin the PHQ-9 or other depression instruments. A more sensitive depression measure would reduce the potential for our health insurance and demographic determinants to be confounded with unobserved severity of mental health.

Sociodemographic Determinants: Specification, Results, and Interpretation

Age

Age is represented by four binary indicators: ages 19–29, 30–44, 45–54, and the omitted category 55–64 years. For specialty and nonspecialty care, there is an upside down U-shaped relationship between age and use with the peak in the age 30–44 years range. In both, those aged 30–44 years are 1.3% points more likely to use treatment compared to those aged 55–64 years and approximately 1.0% points more likely than those aged 19–29 years. Antidepressant use showed a different pattern with respect to age with use peaking in the 45–54 year old group.

What do these U-shaped relationships mean? Age, in part, serves as a proxy for health and mental health not captured in our health measures. But other explanations are plausible. Young adults cumulatively have less exposure to the health care system, and thus, less time for providers to detect depression and recommend treatment. Tastes and preferences may change as young adults mature, or alternatively, they may suffer for years before seeking treatment. Cohort effects may also be at play here with stigma likely greater in older groups.

Sex

Sex is usually included in demand models to reflect biological differences in the prevalence, course, and severity of disease. Women, for example, are much more likely to have depression. However, controlling for symptoms of depression as much as possible in the MEPS, we find that women are still much more likely than men to use treatment, especially antidepressants. Whether this is due to unmeasured differences in depression between men and women or differences in preferences over treatment or stigma we cannot say.

Race and ethnicity

A standard representation of race and ethnicity was used in dividing the population into the following groups: non-Hispanic Whites (the omitted group), Hispanic ethnicity, Black race, and others including those of Asian and mixed race ancestry. Hispanics, Blacks, and others are substantially less likely to have nonspecialty and specialty mental health visits but proportionately even less likely to use antidepressants than Whites. For example, Blacks are almost 8% points less likely than Whites to use antidepressants controlling for other determinants. It is hard to see how unmeasured differences in depression severity might explain these magnitudes. More likely, it reflects unmeasured differences in attitudes and differential access to care. Here measures related to immigration and citizenship status have not been included because they are not available on the MEPS public use files, but they substantially reduce the magnitude of the effects for Hispanics on treatment use (not shown).

Marital status and household composition

Following standard practice (Table 1), a measure of whether the person was married at the time of their round 1 MEPS interview is included. Counts of the number of household

members between the ages of 0–5, 6–17, 18–64, and 65 years and older are also included. One reason for including household composition variables is the potential protective health benefits of marriage. Another is that increasing family size may reduce resources available, both money and time, to any one particular adult in the family for treatment. Consistent with both rationales, the measures were negatively correlated with different types of treatment use, with the exception of a small positive effect of marriage on antidepressant use. But interpretation here is difficult. Depression may also lead to divorce and family dissolution (reverse causality) reducing the magnitudes of the effects which have been observed. Family composition may also be related to unmeasured preferences for depression treatment.

Education

A series of binary indicators corresponding to degrees obtained is obtained: less than high-school diploma (omitted), high-school diploma or equivalent, some college, bachelors, and advanced degree (Masters, MD, JD, PhD). This simultaneously allows for a nonlinear relationship between education and treatment as well as potential 'degree' effects. That is, more than just another year or 2 years of college separates someone with some college from someone who earned their bachelor's degree. Certainly, this is true in the labor market but may extend to preferences over treatment through its effects on social class and norms.

The regression results show substantial differences by education, even controlling for symptoms of depression. Those with a high-school diploma or less are substantially less likely to seek treatment than their better educated counterparts. Interestingly, there is little difference in use of nonspecialists and antidepressants among those with some college, bachelor's, or graduate degree. However, there is a strong gradient for specialists, with use increasing sharply with education.

Clearly education is strongly related to depression treatment, but is it a determinant in a causal sense? Educated consumers might understand better the importance of adherence with antidepressant medication schedules. Or they may have higher quality interactions with therapists providing cognitive behavioral therapy (CBT). In both cases, better educated consumers might derive greater benefits and thus more likely to continue treatment. They may also be more likely to initiate treatment if they better understand potential benefits. However, we cannot help but suspect that unmeasured preferences and social class norms drive much of the educational differences we observe. It is hard to understand why those with graduate degrees would be so much more efficient than those with bachelor's degree in the production of CBT and other talk therapies but not with antidepressant medication. More likely, the stigma surrounding seeing psychiatrists, psychologists, and other specialists is lower among those with graduate degrees.

Reverse causality is potentially a problem with our education measures. Depression may begin in adolescence leading to lower educational achievement through decreased motivation. Such bias would tend to reduce the magnitude of educational effects measured. In the opposite direction, higher education may be correlated with greater economic resources available to pay for treatment.

Geography

Indicators for each of the four Census Regions in the United States (Northeast, Midwest, South, and West) and whether the person resided in a Census Bureau defined Metropolitan Statistical Area, a measure of whether the person lives in an urban or rural area have been included. These are likely correlated with tastes for treatment. For example, stigma for mental health treatment is thought to be stronger in rural areas and in the South. They are also attractive proxies because bias from reverse causality (health causes location) is probably small. Indeed, we find that those in urban areas are more likely to use specialists, whereas those in the South are somewhat less likely. Geography may also be correlated with availability of health care services, but it is also likely that supply follows demand (more doctors in areas where people like to use services). A growing literature also suggests substantial local variations in medical provider practices. In this context, there may be variations in preferences among psychiatrists to treat patients with talk therapy instead of medicating depressed patients.

Proxy

The MEPS is a household survey with one person responding for all household members (the Adult SAQ is one exception). Although MEPS requests that this be the person most knowledgeable about health and health care in the family, there may still be issues with proxy responses. For example, a wife may not be aware that her husband sought depression treatment. To account for the potential for underestimating treatment use obtained by proxy, an indicator is included for whether the MEPS sample member is the respondent (proxy = 0) or not (proxy = 1). Consistent with this worry, the regression results show that proxy respondents are approximately 1% point less likely to use each of the three types of treatment. Of course, proxy status could be correlated with other unmeasured aspects of individuals related to treatment.

Trend

To account for the possibility that demand increased between 2004 and 2009, two time-trend variables were included. The first is linear term for survey year (minus 2004 to normalize to 0). The second term squares the first allowing for nonconstant changes in demand. In fact, there is no discernable trend in overall demand using this or any number of alternative specifications between 2004 and 2009. This would not have been true in the late 1980s and 1990s when demand grew rapidly with the introduction of new classes of antidepressants.

Excluded Determinants

A number of potential determinants do not appear in this empirical example. Employment status, occupation, and industry have been excluded because of their direct potential for reverse causality and uncertain effects on demand. Ideally, time price would be included, but direct measures of travel and time costs as well as suitable proxies are lacked.

The MEPS Adult SAQ contains four items designed to represent individual attitudes and beliefs. However, the correlation between first and second year responses was lower than expected suggesting responses may be endogenous with current health and health care use. Good arguments could be made either way for including smoking status, but it has been excluded as lacking a clear a priori hypothesis about its effects. Although clearly relevant to depression, alcohol and drug behaviors are not available. A number of available access measures believed to be endogenous have been omitted. Finally, Local area supply side and market characteristics that can be merged onto MEPS for similar reasons have been omitted.

Need Versus Demand: Illustrating with the Empirical Example

Policymakers and advocates often speak of 'unmet need' for treatment for diseases such as diabetes, heart disease, and depression. In this context, need is some norm that is being applied to groups of individuals defined by illness and then determining the extent to which they actually receive care. In the example given, if a diagnosis of depression is used as the determinant of need then it is being said that all individuals with depression should receive treatment. Those without it have unmet need. If one prefers, the definition can be made more restrictive, as many have proposed, to include additional functional impairment criteria, but regardless we are still applying some external norm. Alternatively, the actual use of individuals in one group (say high income) as a norm for other groups can be used.

As introduced earlier, economists view demand strictly through the eyes of the individual. Even in demand–supply graphs, market demand curves are simply the sum of individual demands. The authors talk about 'need' variables being included in demand models, but individuals take this more into account than just their health in determining whether and how much care to consume. An individual with depression may or may not perceive that they need treatment at all. Some depressed individuals may not seek treatment even if their out-of-pocket price is zero. For others, whether they seek treatment may depend critically on price.

Survey data such as MEPS gives us the opportunity to study measures of 'need' as distinct from demand and use. If we look at the descriptive statistics on Table 2, we see that only 40% of people with a current PHQ-2 score of 3 or greater (suggesting probable depression and the need for further screening) receive treatment. If we use this cut-point as our norm for treatment need, it suggests that more than half of currently depressed individuals do not receive treatment and therefore have unmet need. Conversely, 12% of those not meeting our hypothetical norm for need consume treatment. The empirical model can be used to simulate the effect that changing a key determinant has on changing the relationship between our norm for need and actual demand. Health insurance coverage, because it is so amenable to policy changes, is the obvious choice. Here, the model implies that providing public coverage to all of the uninsured would reduce the gap between need and demand in the uninsured from 78% to 65% and among all adults aged 18–64 years from 60% to 57%.

Conclusion

Specifying and interpreting the empirical determinants of health care demand is as much art as science. As seen from the author's review of recent empirical studies, there is not only widespread agreement about some determinants such as age, sex, health status, and education but also wide variation in the treatment of other characteristics that might be correlated with health care use. Researchers are confronted with tough trade-offs among competing concerns in selecting and specifying determinants they think relevant in demand models. Formal models of health care demand can help guide us about the treatment of variables such as health status, income, and price. Theory also guides us in the choice of proxies and, also using statistical principles, how to best specify these proxies to represent unmeasured aspects of health and treatment seeking preferences. As seen from the empirical illustration, proxies such as education are often powerful predictors of demand. These same economic and statistical principles also aid us in interpreting our empirical determinants. But in a world where unobserved preferences and health play such a key role, we will always face some uncertainty about how to model the empirical determinants of health care demand.

Disclaimer

The views expressed in this article are those of the author, and no official endorsement by the Agency for Healthcare Research and Quality, or the Department of Health and Human Services is intended or should be inferred.

See also: Health and Health Care, Need for. Modeling Cost and Expenditure for Healthcare. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Sample Selection Bias in Health Econometric Models

Further Reading

Coffey, R. M. (1983). The effect of time price on the demand for medical-care services. *Journal of Human Resources* **18**, 407–424.

Cook, B. L., McGuire, T. G. and Zaslavsky, A. M. (2012). Measuring racial/ethnic disparities in health care: Methods and practical issues. *Health Services Research* **47**, 1232–1254.

Deb, P. (2001). A discrete random effects probit model with application to the demand for preventive care. *Health Economics* **10**, 371–383.

Freiman, M. P. and Zuvekas, S. H. (2000). Determinants of ambulatory treatment mode for mental illness. *Health Economics* **9**, 423–434.

Garfield, R.L, Zuvekas, S. H., Lave, J. R. and Donohue, J. M. (2011). The impact of national health reform on adults with severe mental disorders. *American Journal of Psychiatry* **168**, 486–494.

Kenkel, D. S. (1990). Consumer health information and the demand for medical care. *Review of Economics and Statistics* **72**, 587–595.

Manning, W. G., Newhouse, J. P. and Ware, J. E. (1982). The status of health in demand estimation; or, beyond excellent, good, fair, poor. In Fuchs, V. R. (ed.) *Economic aspects of health*, pp. 143–184. Chicago: University of Chicago Press. Available at: http://www.nber.org/books/fuch82-1 (accessed 24.06.12).

Meyerhoefer, C. D. and Zuvekas, S. H. (2010). New estimates of the demand for physical and mental health treatment. *Health Economics* **19**, 297–315.

Propper, C. (2000). The demand for private health care in the UK. *Journal of Health Economics* **19**, 855–876.

Rous, J. J. and Hotchkiss, D. R. (2003). Estimation of the determinants of household health care expenditures in Nepal with controls for endogenous illness and provider choice. *Health Economics* **12**, 431–451.

Sosa-Rubía, S. G., Galárraga, O. and Harris, J. E. (2009). Heterogeneous impact of the "Seguro Popular" program on the utilization of obstetrical services in Mexico, 2001–2006: A multinomial probit model with a discrete endogenous variable. *Journal of Health Economics* **28**, 20–34.

Wagstaff, A. (1986). The demand for health: Some new empirical evidence. *Journal of Health Economics* **5**, 195–233.

Yang, Z., Gilleskie, D. B. and Norton, E. C. (2009). Health insurance, medical care, and health outcomes. *Journal of Human Resources* **44**, 47–114.

Relevant Websites

www.nimh.nih.gov

National Institute of Mental Health, National Institutes of Health.

www.meps.ahrq.gov

Medical Expenditure Panel Survey (MEPS) On-Line Resources and Data, Agency for Healthcare Research and Quality.

Health Econometrics: Overview

A Basu, University of Washington, Seattle, WA, USA
J Mullahy, University of Wisconsin-Madison, Madison, USA

Empirical analysis of data describing relationships involving health – health econometrics – arises in a wide variety of important scholarly and policy contexts. The econometric analysis of data on topics as diverse as health insurance, substance use, provider behavior, chronic disease, evaluation, market structures, regulation, medical technologies, labor supply, and others is encountered routinely in every issue of leading field journals like the *Journal of Health Economics*, *Health Economics*, and others.

Reflecting the increased prominence of both conceptual and applied health econometrics research is an increasing array of professional activities devoted specifically to health econometrics. For over 20 years, researchers at the University of York and local sites all across the European Union have organized annual meetings on health economics and health econometrics. More recently, specialized health econometrics conferences and workshops have regularly been organized in the US, Italy, and elsewhere. Beyond these, sessions and pre-conference courses dedicated to health econometrics have been among the most popular and well-attended activities at meetings of major health economics organizations like the International Association of Health Economics, the American Society of Health Economists, and others.

The methods of health econometrics are deployed to address a wide variety of questions. At their essence, many are concerned with the estimation of treatment effects, broadly construed. These can arise in narrow small-N contexts like the evaluation of clinical interventions as well as in broad population or large-N contexts like the implementation of tax, regulatory, or other public policy interventions. Recent emphases on 'comparative effectiveness' and the empirical methods used to understand the relative value of interventions have underscored the importance of linking relevant decision-making contexts to reliable and robust analytical methods that can be deployed to inform such decisions.

How to deliver informative estimates of treatment effects in the light of observational data often utilized in the service of such questions is one of the central problems of applied health econometrics. Such observational data are now drawn from an increasingly wide set of sources: Population and community surveys, administrative data describing program participation, electronic medical records, and others. Regardless of the particular data, there is widespread recognition that many of the treatments at issue are endogenous with respect to the outcomes of interest (i.e., are correlated with unobserved determinants of such outcomes, known as 'confounding' in the epidemiological literature). To circumvent the problems that arise with endogenous treatments, quasi-experimental methods are often utilized. Instrumental variable methods, longitudinal or panel data analyses, and others are deployed with assumptions sufficient to generate consistent estimates of parameters of interest (whether the assumptions are reasonable and/or holds in the context of the particular study are separate but important questions). One such assumption is the correct specification of a model for the data at hand. Economic theory, or any other theory for that matter, often has a hard time predicting directions of covariate effects. It does not provide much guidance as to the appropriate functional form for the data at hand. Therefore, a good deal of health econometrics literature has focussed on ascertaining appropriate models using various goodness of fit measures. A good discussion of these issues can be found in the chapter by Manning on modeling healthcare expenditures that are known for their idiosyncracies. Appropriate specification of a model is then followed by identification of the parameter of interest, often a treatment effect parameter. Geographic variation in constraint sets has been one prominent identification strategy (Rosenzweig and Schultz, 1983), and indeed was – to our knowledge – the approach that introduced instrumental variable analysis to clinical and related audiences (McClellan *et al.*, 1994, in the context of differential distance instruments). More recently, approaches like propensity score or control function methods have become popular in health services research even though the extent to which such methods fail to circumvent problems arising from confounding is often underappreciated.

In this context it is often useful to bear in mind that the 'gold standard' of the randomized clinical trial against which observational data analysis is frequently held is itself an emperor that often wears little clothing. Within-trial behaviors like attrition, non-adherence, etc. (Efron and Feldman, 1991; Lamiraud and Geoffard, 2007) will typically jeopardize both the internal and external validity of results and inferences based on such data. Floras are typically compliant with treatment protocols, but human fauna will often fail to be. Whereas randomized trial provides a solid conceptual foundation for thinking about an ideal data-generating experiment (Permutt and Hebel, 1989, for a specific example executed in an instrumental variable context), its actual implementation often falls short of the ideal. When contemplating the analysis of health (or any other) data, it can generally be more helpful to appreciate that such data are themselves often generated by purposive decisions of data suppliers and demanders (Philipson, 1997).

In many instances, the particular nature of the data to be analyzed by health econometricians sets health econometrics apart from other domains of applied econometrics. Many of the measurement and sampling approaches used to describe health-related phenomena as well as the consumer, producer, and market decisions and processes from which such data arise are more or less unique to health economics. Econometric methods used to analyze such outcomes data – censored, bounded, discrete, ordered, etc. – have often been developed by analysts working primarily in health economics (Newhouse, 1987). Even so, health econometricians have sometimes failed to be sufficiently sensitive to the fundamental measurement features of the data they analyze, e.g.,

estimating moments of ordinal scale outcomes like self-reported health status obtained using Likert scale or analogous strategies (Stevens, 1946).

Regardless of the particular questions at hand, the ability to move from conceptualizing such analysis to implementing it has required both individual-level (or micro-) data describing the choices and outcomes of health producing consumers and suppliers observed over space, over time, or both, as well as a rapid evolution of analytical and data management that has permitted such data to be analyzed using state of the art methodologies (e.g., Stata, Limdep, R, and others; Renfro, 2004 for a general discussion). Given the sensitive nature of many topics with which health economists deal at the household, institution, market, and population levels, ideal data may sometimes not be available for analysis owing to a variety of privacy protection protocols that have legal standing in most countries. Nonetheless, the progress that has been made in advancing empirical understanding of such phenomena is remarkable.

Interested readers may find as a useful starting point Andrew Jones's (2000) seminal and comprehensive overview of health econometrics topics. The articles in this section complement in some respects Jones's overview and, in the light of the ongoing rapid pace of conceptual and methodological developments in the field, bring some of the topics he addressed over 10 years ago into newer light.

While the articles in this section cover a broad swath of topics in health econometrics, it could also be pointed out for context several topics that are not accorded article-length treatment in this Encyclopedia although, in some instances, they are treated in part in various articles. Among such topics of interest to health economists include specific treatment of outcome measurement, econometric analysis of experiments, prediction and forecasting, and multivariate outcomes. Also to be noted with considerable sadness is that a article on the econometric analysis of clinical trial data was planned by Prof. Tom Ten Have and was in early stages of preparation when he died of multiple myeloma at a rather young age in 2011.

References

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* **86**, 9–17.

Jones, A. M. (2000). Health econometrics. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, 1st ed., vol. 1, ch. 6, pp. 265–344. Amsterdam: Elsevier.

Lamiraud, K. and Geoffard, P. Y. (2007). Therapeutic non-adherence: A rational behavior revealing patient preferences? *Health Economics* **16**, 1185–1204.

McClellan, M., McNeil, B. J. and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Assocation* **272**, 859–866.

Newhouse, J. P. (1987). Health economics and econometrics. *American Economic Review Papers and Proceedings* **77**, 269–274.

Permutt, T. and Hebel, J. R. (1989). SImultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* **45**, 619–622.

Philipson, T. (1997). Data markets and the production of surveys. *Review of Economic Studies* **64**, 47–72.

Renfro, C. G. (2004). Econometric software: The first fifty years in perspective. *Journal of Economic and Social Measurement* **29**, 9–107.

Rosenzweig, M. R. and Schultz, T. P. (1983). Estimating a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight. *Journal of Political Economy* **91**, 723–746.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* **103**, 677–680.

Health Insurance and Health

A Dor and E Umapathi, George Washington University, Washington, DC, USA

Introduction

Most developed countries provide universal or near-universal health insurance coverage. The US has lagged behind with 49.9 million individuals, or 16.3% of the population, reportedly uninsured as recently as 2010. Policy debates in the US, where proponents of universal coverage have argued that extending coverage to the uninsured would result in better access to health care, improved health outcomes, and ultimately lower costs, have culminated in the enactment of the Patient Protection and Affordable Care Act and the Health Care and Education Reconciliation Act (collectively referred to as the ACA) in 2010. In tandem with the policy debate, health economists have explored the impact of health insurance on health outcomes using empirical methods. Nevertheless, the evidence so far remains inconclusive. Both those favoring universal coverage and those arguing for limited steps were able to find some support for their respective positions, albeit on a selective basis. Although supporters of the legislation claim that the phased implementation of the ACA will dramatically reduce the ranks of the uninsured, millions of Americans are expected to remain without any coverage. Gaining a better understanding of what the health economics literature offers to this policy discussion thus remains highly relevant.

Economic theory suggests that health insurance can serve a dual purpose of protecting people against the financial burden of illness and of increasing access to care to meet unmet health care needs. Conventional expected utility theory, which is widely applied to evaluate the demand of health insurance, considers any medical expenditure to be a loss of income. In view of this theory, the purchase of health insurance or medical care reduces a consumer's wealth. For the research question at hand that investigates whether health insurance improves health, medical expenditures may instead be considered an input in the health production function (Grossman model). Accordingly, the stock of health can be improved by investments in medical care which the purchase of insurance enables. Most of the relevant literature assumes such a model, at least implicitly. At the same time, the benefits of health insurance coverage may be partially offset by the effects of ex-ante moral hazard (Moral Hazard is the change in behavior that occurs as a result of becoming insured. Ex-ante moral refers to the change in the probability of illness or injury. Ex-post moral hazard refers to the size of the loss (medical expenditures) after the illness/injury occurred.). Accordingly, people with health insurance coverage may have less financial incentives to engage in healthy behaviors that prevent injury and illness.

A body of existing research supports a negative association between the lack of health insurance and access to care, and in turn, a positive association between access to care and health outcomes. Descriptive studies report that in the US the uninsured have less access to health care, higher risks of unmet health needs, and poorer health outcomes. For example, research has shown that uninsured adults use 60% less ambulatory health services and 30% less inpatient health services than insured adults. In addition, the uninsured are more likely to delay seeking care, to report not being able to see a physician due to costs, and to require costly emergency care. Even among patients that did see a clinician, only 18% of uninsured patients received all recommended follow-up treatment in comparison to 30% of insured patients. In comparison to insured adults, the uninsured are also more likely to have a lower self-reported health status and are more likely to be diagnosed at an advanced stage of cancer, suffer from cardiovascular diseases, exhibit worse glycemic control, and experience higher in-hospital mortality rates.

Although the positive association between health insurance coverage and health outcomes appears to be convincing, the issue has been a matter of considerable controversy among empirical economists. Mere associations may mask the fact that healthier individuals tend to be better equipped to obtain health insurance, leading one to overstate the actual contribution of insurance to health, a problem related to reverse causality and simultaneity in econometric estimation. To address this general concern different methodological approaches have been undertaken, perhaps contributing to occasionally conflicting results. In addition, this literature encounters measurement issues which merit further scrutiny.

In this article, the aim is to shed light on the nuances in the literature on the causal effect of health insurance on health outcomes. The objective here is to clarify the limitations of this literature and to provide a deeper understanding of the causal pathways between insurance and health, ultimately to better inform the policy debate. As will be seen, certain patterns have emerged: Lack of insurance may be more of an adverse factor for mortality and generic health outcomes although its effects on condition-specific measures are more complex. In addition, interruptions in insurance coverage can be just as harmful as the complete lack of insurance. This review focuses primarily on the US, and its private segment of the market, where the acquisition of insurance has largely been a matter of individual choice, at least before the implementation of ACA. However, inferences are also drawn from the literature on government sponsored insurance, namely, Medicare and Medicaid, where insurance coverage is simply assigned to individuals based on age and income, and the state in which the individual resides.

The remainder of this article is divided into five sections. First, there is a brief description of the characteristics of the US uninsured population and the anticipated changes for health insurance coverage under the ACA. Second, the methodological challenges related to estimating the impact of insurance on health are discussed. Third, the methods used in the source literature are illustrated. Fourth, general findings in the source literature are presented. Finally, challenges for future research are discussed and emerging implications of the ACA comes in as conclusion.

The Uninsured and the US Health Care System

In the US, historically, the elderly and the disabled have received public insurance coverage through Medicare, whereas the poor received public insurance through Medicaid. Otherwise, health insurance coverage has effectively been tied to employment with 90% of all privately insured individuals receiving their coverage through their employers. Roughly 50% of employees participate in employer-sponsored health insurance coverage and consequently a large portion of employees, especially temporary and part-time employees, lack insurance. Not surprisingly, the contraction in the labor market that accompanied the Great Recession in the early twenty-first century also contributed to the rising number of the uninsured. Between 2008 and 2010, nearly one-quarter of working-age adults reported that they or their spouse had lost their job and more than 50% of these people became uninsured. Historically, a significant portion of the uninsured population consisted of relatively vulnerable groups such as the near poor and near elderly. Individuals belonging to these groups are more likely to experience unemployment compared with higher income or young people, and are also more likely to suffer from adverse health events. Yet, income and age restrictions precluded these individuals from enrolling in public insurance programs such as Medicaid or Medicare, leaving them at a higher risk of remaining uninsured.

The ACA is expected to greatly reduce the number of uninsured. Under the ACA, most US citizens and legal residents will be required to have health insurance, a number of states will expand Medicaid to include the nonelderly population at 133% of the federal poverty line, and states are engaging in setting up health insurance 'exchanges' offering plan choices to previously underserved individuals. In addition, federal subsidies will be made available to small firms and individuals for the purchase of insurance. Nevertheless, the ACA will only reduce the number of uninsured by half. The Congressional Budget Office estimated that by 2019, 23 million Americans will remain uninsured even if the ACA is fully and successfully implemented. The small penalties levied against those opting out of the system, the so-called 'mandates,' may not be sufficient to outweigh the incentives not to join. Moreover, certain population groups are exempted.

Methodological Challenges

In this section, methodological challenges facing research estimating the causal effect of insurance on health are discussed, starting with identifying the uninsured, defining health outcomes, and addressing endogeneity bias.

Identifying the Insured and Uninsured

Identifying the uninsured is a difficult task. In most household surveys, individuals tend to misclassify themselves as being insured or uninsured due to the way health coverage is defined. Misclassification occurs mainly around changes in employment status, due to confusion about receiving coverage through another family member, or simply because of poor recall when survey questions require longer retrospective periods. Another type of misclassification occurs when Medicaid enrollees or beneficiaries of other public programs do not recognize these programs to be a form of health insurance, leading some to erroneously identify themselves as uninsured. Indeed, comparisons of survey data with administrative data have demonstrated that surveys consistently underestimate insurance coverage. Underreporting of insurance coverage has been proven to be a particular problem for Medicaid with specific evidence of underreporting available for all Medicaid beneficiaries in California and Maryland, and nationwide for child beneficiaries. As a result, the estimated prevalence of uninsurance varies somewhat between surveys. For instance, a comparison of the Health Retirement Survey (HRS) and the Medical Expenditure Panel Survey (MEPS) in 2006 yielded an uninsurance rate of 10% and 12%, respectively.

Another form of measurement error occurs when continuity of coverage is of interest. The majority of studies reviewed used insurance coverage at the time of interview as the key explanatory variable. However, in the US, particularly in the private sector, people frequently gain and lose health insurance coverage, a phenomenon also known as churning. Between 1998 and 2002, churning affected 22% of the population. Much like the complete absence of insurance, churning may adversely affect health due to discontinuity of care and delays in treatment. Among children and adults, loss of insurance is associated with a lower likelihood of having a primary care provider, getting check-ups, or receiving the recommended follow-up care. Intermittent health insurance coverage may thus affect health outcomes in a similar fashion as the lack of health insurance coverage. A number of nationally representative surveys including the HRS, the MEPS, and the National Health and Nutrition Examination Survey (NHANES) ask respondents to report their health insurance status retrospectively for a 3–18-month period. Several studies made use of this feature to define insurance in terms of frequency of changes or to draw comparisons between the continuously insured, the intermittently insured, and those lacking insurance coverage over a fixed period.

Although this article focuses only on the provision of insurance, note that health insurance is heterogeneous and variations in generosity of health insurance benefits can occur not only between general insurance categories such as Medicaid, Medicare, and private, but also within such groupings. The lack of information about insurance generosity in household surveys creates a major practical limitation for related research. The literature reviewed here is generally silent on this issue.

Measuring Health Outcomes

The source literature assessed the impact of health insurance on three broad categories of health outcomes: mortality, condition-specific morbidity, and generic health measures. The majority of studies relied on mortality-based measures, such as all-cause mortality, disease-specific mortality, or survival rates. Condition-specific morbidity measures pertain to the clinical status of a given illness or medical condition.

Examples from the literature include low birth weight, obesity, and cancer disease stage.

The term 'generic health measures' is used to describe aspects of functioning in health related daily activities which are not necessarily disease specific. Generic health measures can be either unidimensional or multidimensional. Unidimensional measures include self-reported health indicators, such as one's overall ranking or the number of chronic conditions. Multidimensional measures combine several subjective indicators of physical and mental health into a single additive scale.

Expert research in the psychosocial literature has validated the use of self-reported health rankings. In addition, the health services research literature offers well developed and validated methodologies to construct multidimensional health indices, such as the Short Form-36 (SF-36). Data elements that comprise these indices are now routinely included in nationally representative surveys such as the National Health Interview Survey and the HRS. For example, Dor *et al.* (2006) and McWilliams *et al.* (2007) combined self-reported general health, the number of physical limitations, and pain into a modified SF-36 health index. **Table 1** provides

Table 1 Overview of source studies by payer category

Author (year)	*Method*	*Insurance*	*Health outcome*	*Results*
Bhattacharya *et al.* (2003)	IV model uses state Medicaid eligibility and employer-sponsored insurance	Uninsured, private insurance, and public insurance	Mortality	Insurance reduced risks of dying, private insurance more than public insurance
Bhattacharya *et al.* (2011)	IV model uses state percentage of workers working in medium and large size firms and Medicaid eligibility	Uninsured versus insured	Obesity	Provision of public and private insurance to the uninsured increases body weight, with slightly larger effects for public insurance
Courtemanche and Zapata (2012)	DD compares Massachusetts to other states before and after health reform in Massachusetts	Massachusetts reform	Self-reported, physical and mental health, functional limitations, joint disorders, BMI and physical activity	The Massachusetts health reform improved all health outcomes
Dor *et al.* (2006)	IV model uses state marginal tax rates, average unemployment rate, and average rate of unionization	Uninsured versus privately insured	SF-36 type health index	Insurance improved health status by 10–11%. Separate regressions for people with asymptomatic conditions, chronic conditions, or nonchronic conditions found no substantial differences in health.
Hadley and Waidmann (2006)	IV model uses spousal union membership, immigrant status, and involuntary job loss in the past 5 years	Continuity of insurance coverage	Health index and mortality	Continuous coverage from age 55 onward reduces mortality and increases health
Kaestner (1999)	IV model uses state dummies, interaction between state dummies and high income, and mother's employment status	Uninsured, Medicaid, and privately insured	Low birth weight	Insurance coverage did not improve birth weight
Pauly (2005)	IV model uses firm size and/or marital status	Uninsured versus privately insured	Self-reported health and number of chronic conditions	Insurance status does not affect health outcomes
Thornton and Rice (2008)	IV model uses state percentage of firms with more than 20 employees and the percentage of union workers with state FE	Extending private insurance to the uninsured	Mortality	Insurance reduced mortality.
Weathers and Stegman (2012)	IV model uses randomized assignment to health insurance	Uninsured versus privately insured	Self-reported, physical, and mental health, depression, disability and mortality	Private insurance improved self-reported, mental and physical health 1 year following health insurance enrollment, but did not reduce mortality within 2–3 years of enrollment.

a brief description of the literature used in this article. Although this article focuses primarily on private insurance a summary of the evidence on the causal effect of Medicaid and Medicare is available in Table 1.

Although each of the above health outcome categories offers certain advantages, they are also affected by certain measurement issues and interpretation problems. An obvious advantage of using mortality as a summary measure of adverse outcomes is that death is completely unambiguous, and it is easily verifiable in most data systems. Thus, mortality is susceptible to minimal measurement error. However, although mortality reflects the lowest boundary of health, it does not capture the path of declining health over the individual's life cycle. In contrast, condition-specific measures may capture the stage and severity of illness, but targeting a narrowly defined condition may lead researchers to overlook other important dimensions of health. Moreover, most surveys rely on self-reporting of morbidity indicators, and thus require respondents to possess specific and time sensitive knowledge of their own disease.

Generic health measures provide a broader view of health that transcends any single condition. Because general health measures are based on a person's functioning, they can be used for more general population groupings than the above. Another advantage is that most household surveys provide validation of respondents' replies to questionnaires. However, by trading off specificity generic measures may mask insurance 'treatment' effects that might apply to certain conditions but not others. Further adding to measurement error, interpretations of good functioning may vary by respondents' age, gender, and other groupings. However, the health services literature suggests that combining several unrelated aspects of health helps mitigate reporting error in multidimensional indices. Finally, generic health measures may reflect health status changes, with a time-lag, rather than responding instantaneously.

In summary, both insurance coverage and health status indicators, excluding mortality, are subject to measurement error. In regression analysis measurement error in the dependent variable (health status) increases standard errors but does not produce a biased estimator. However, measurement error in the explanatory variable (insurance) biases the coefficient estimates, although the direction of the bias is predictable (toward zero) as long as measurement error does not appear in any other independent variables in the model.

Endogeneity of Insurance in Health

Selection, reverse causality, and omitted variable bias pose other methodological challenges. Each one of these issues presents a special case of endogeneity, whereby ordinary least square estimates of the effect of insurance on health may be biased due to a correlation between a regressor and the regression residual. A myriad of institutional and behavioral processes underlie endogeneity, making it difficult to ascertain whether the bias occurs in an upward or downward direction.

The private insurance market is affected by selection problems, which arise when there is information asymmetry between insurers and insured. Adverse selection occurs when insurance companies attract sick people who are more likely to need and use health care, and when healthy people, who do not anticipate incurring high medical expenses, choose cheaper but less generous plans or opt out of insurance altogether. Adverse selection would bias the estimated relationship between insurance and health downwards. Auspicious selection occurs when insurers try to attract the good risks (e.g., healthy or young individuals), while making plans unattractive for bad risks (e.g., those with preexisting conditions). Auspicious selection would bias the estimated relationship upwards. Both adverse and auspicious selection, and related estimation biases, will be even worse when insurance risks are experienced rated (as in the individual insurance markets) rather than community rated (group insurance).

Reverse causality occurs when health affects health insurance status. The direction of this bias is also unknown. People in poor health are more likely to buy health insurance (or purchase more generous coverage) than healthy people, as they anticipate a greater need for care. Conversely, in the mostly employer-based private segment of the US market, poor health tends to be associated with job loss and hence loss of insurance, particularly in periods of high unemployment.

Finally, a type of omitted variable bias occurs when the individual's insurance choice is determined by some traits that also affect health but are unobservable to the researcher. For example, risk-averse individuals are more likely to hedge the risk of income loss by purchasing insurance, although simultaneously displaying risk-avoiding health behaviors. Similarly, certain people are better equipped to asses both insurance and health care information and act preventively, an unobservable trait sometimes referred to as health ability. Reliable measures of risk aversion, health ability, and underlying health behavior are rarely available in observational datasets. Consequently, in classic regression analysis included explanatory variables may be correlated with the error term.

Estimation Methods

Three different approaches to measuring the causal effect of insurance on health as they appear in the literature is discussed: Studies using instrumental variable (IV) techniques, studies using quasi-experimental designs, and randomized experiments. (Quasi-experiments encompass both the natural experiments and IV studies that are discussed. For the purpose of this article, natural experiments and IV approaches are discussed separately because natural experiments rely on exogenous source of variation in the treatment assignment, whereas the IV approach uses a continuous probability distribution of the treatment assignment.) In the context of private markets, most studies used IV approaches to address the endogeneity issue previously discussed, allowing for probability distributions of the insurance choices made by individuals. In the context of government programs where insurance coverage is simply assigned to individuals based on an arbitrary (exogenous) criterion, quasi-experimental techniques are more relevant. Although the primary interest is in the private segment of the market, some attention is devoted to quasi-experimental evaluations of Medicare and Medicaid in order to draw inferences for future research directions given

the recent enactment of private mandates in the US. Finally, the very limited but important literature on controlled experiments that allow for random assignment of individuals into insured and uninsured states is discussed.

Instrumental Variable Estimation

The issue of endogeneity can be addressed by simultaneous estimation of insurance choice and a health outcome using IVs. Briefly, IVs would be included in the insurance equation but excluded from the health equation based on the following criteria: First, the instrument must be uncorrelated with the error in the health equation. As this is not easily verified, researchers' choice of IVs must rely on economic theory and solid reasoning when choosing appropriate instruments. Second, the instrument must be strongly correlated with insurance choice (the latter can easily be tested).

A variety of instruments have been used, but their validity has been repeatedly called into question. Examples include state-level variables, firm-level variables, certain individual-level variables, or some combination of all of the above (source studies and their instruments are described in **Table 1**). A number of studies employed indirect tests that provide a modicum of confidence. For instance, arguing that state-level Medicaid eligibility and average firm size are independent of health (mortality) but affect the ease with which people obtain Medicaid or employer-sponsored insurance, Bhattacharya *et al.* (2003) examine their strength and relevance as instruments when estimating the effects of public and private insurance on health among human immunodeficiency virus (HIV) patients, using data from the HIV Costs and Services Utilization Study. The authors report a strong correlation between their instruments and insurance coverage based on statistical tests (e.g., the Wald statistic), and a high degree of relevance, based on a reasonable falsification test. (The instruments would be irrelevant if they were to predict health outcomes for an unrelated population. Using a sample of Medicare beneficiaries as an alternative to the original sample of HIV patients, Bhattacharya *et al*. show this is not the case, suggesting that their instruments are valid.)

In a related example, Dor *et al.* (2006) used state marginal tax rates, average unemployment rate, and average unionization rates to instrument insurance. The study population included adults age 45–64 from the 1992 to 1996 HRS surveys. Substantial literature suggests that state-level tax burden is uncorrelated with health but to be positively correlated with insurance participation. Similarly, union membership is positively correlated with being offered insurance coverage, whereas unemployment is negatively correlated with private health insurance coverage. However, the use of marginal tax rates in the first stage results were only weakly correlated with health insurance. Some critics raised questions about the validity of unemployment as an instrument, arguing that macroeconomic downturns may affect health not only through insurance loss but also because they affect health behaviors such as drinking and exercise. It should be noted, however, that previous versions of the study used county-level firm sizes as an instrument, yielding essentially the same results for the effect of insurance on health (Dor *et al.*, 2003).

Various combinations of person-level variables have also been used to instrument insurance. Among these are employer size, marital status, spousal union membership, immigrant status, involuntary job loss, and self-employment status (**Table 1**). Again, the validity of any of these variables can be questioned, given that is unlikely that they do not affect health in some indirect way. For instance, for some people job loss may lead to depression or loss of physical activity, leading to deterioration in overall health; foreign-born workers from poor countries may have worse health status than native-born US workers, suggesting that immigration status is and negatively correlated with health. Spousal union membership may be the most appealing variable in terms of avoiding systematic correlation between the instrument and the subject's health. However, any study to date that has attempted to test the validity of this instrument by itself is unheard of.

Quasi-Experiments

Quasi-experimental methods including difference-in-difference (DD) models and Regression Discontinuity Design (RDD) models have been used to get around the difficulties of modeling endogeneity and selecting appropriate IVs. These models, which borrow from the more general program evaluation literature, rely on finding cases where insurance can be treated as an exogenous intervention. In DD models, a treatment group and a comparison group are identified and the impact of the treatment is inferred from the difference between the changes experienced by the two groups over time; DD models have been widely used to evaluate Medicaid expansions and outcomes in US states, whereas RDD models are more readily applied to evaluations of Medicare (**Table 1**). In an innovative study, Polsky *et al.* (2009) employ DD to Medicare by comparing health status for the previously uninsured and continuously insured before and after enrollment at age 65.

RDD models exploit exogenous policy rules, yielding a comparison of individuals above and below a fixed cutoff point. A critical assumption for RDD models is that by tracking individuals closely around the cut off trends unrelated to the policy are essentially filtered out. RDD models are commonly used to evaluate Medicare because of its generally arbitrary eligibility criterion which assigns individuals to the program at age 65. For example, using the 1991–2002 Behavioral Risk Factor Surveillance System, Decker (2005) estimated the effect of Medicare eligibility on breast cancer stage and survival. To ensure that other age-related changes such as retirement were not erroneously captured in her eligibility indicator, Decker also controlled for employment status; although this additional variable was statistically significant in her model, it did not alter the estimated Medicare effect. In a variant of RDD, McWilliams *et al.* (2007) used a linear spline regression to compare health outcomes for people before and after acquiring Medicare.

Randomized Controlled Experiments

Given the difficulty posed by endogeneity and concerns over nonsymmetry between treated and controls in quasi-experimental studies, ideally, the impact of insurance on

health would be inferred from randomized controlled trials (RCT) whereby people are randomly assigned to separate categories of those receiving health care coverage and those without any insurance. However, RCTs are virtually impossible to implement due to both practical and ethical reasons. Nevertheless, two recent policy experiments offer close approximations; the first was carried out by the US Social Security Administration (SSA), and the second was implemented by the state of Oregon.

Focusing on Social Security Disability Insurance (SSDI) beneficiaries, the SSA experiment was designed to test whether making medical benefits available to these beneficiaries immediately, rather than requiring a mandatory 2 year waiting period improves health outcomes. Accordingly, between October 2007 and November 2008, a subset of newly enrolled SSDI beneficiaries was asked to participate in the Accelerated Benefits (AB) demonstration. Those that agreed to participate were randomly assigned to groups receiving a relatively generous health insurance plan versus remaining uninsured for 2 more years. The AB demonstration thus provided a unique opportunity to test whether having insurance improves short-term health outcomes (Weathers and Stegman, 2012).

In 2008, Oregon had enough funding to expand enrollment to 10 000 low-income adults. Later dubbed the Oregon health insurance experiment (Oregon HIE), beneficiaries were randomly chosen by lottery from the pool of eligible candidates, thereby creating two groups of covered and noncovered individuals. The origins of the Oregon HIE can be traced to the RAND Corporation Health Insurance Experiment of the late 1970s. However, the RAND Corporation study focused on cost sharing levels with insurance rather than outright withdrawal of insurance.

Analysis of the first year's results offers valuable insights, but also highlights limitations of RCTs and their approximations (Finkelstein *et al.* 2011). At the end of the year, insurance coverage appeared to improve participants' self-reported physical and mental health in comparison to the uninsured control group. However, when the timing of these improvements was examined more carefully, the researchers found that they occurred before the actual initiation of health care. This may suggest a type of placebo effect whereby the mere availability of health insurance provides the individuals with a sense of well-being and a heightened perception of health status.

Results: Health Insurance Effects by Type of Health Measure and Study Population

Having noted methods, studies can be further classified by type of health outcome measure and type of population studied. Results are summarized accordingly:

Health Outcomes

Overall, the large majority of studies agree that health insurance coverage reduces the risk of mortality. For example, using state-level panel data from 1990 to 2000 and firm size and union membership to instrument insurance, Thornton and Rice (2008) concluded that extending private insurance to the uninsured would reduce adult mortality and save more than 75 000 lives annually. Similarly, using union membership, immigrant status, and involuntary job loss as instruments for insurance, Hadley and Waidmann (2006) concluded that extending insurance coverage to all Americans between the ages of 55 and 64 would reduce mortality in this age group. Furthermore, Bhattacharya *et al.* (2003), using the 1996–1998 HIV Cost and Services Utilization Study, concluded that HIV patients with private health insurance coverage had a 79% lower relative risk of dying than HIV patients without insurance. And, HIV patients with public health insurance had a 66% lower relative risk of mortality than HIV patients without insurance. Weathers and Stegman (2012) is the only study to find no effect of insurance coverage on mortality; their brief follow-up of 3 years did not allow for identification of longer term effects in their experimental data.

Similarly, a majority of studies found positive effects of health insurance on generic health measures. Weathers II and Stegman found that insurance improved self-reported mental and physical health of SSDI beneficiaries one year after receiving health insurance. Using the 1992–1996 HRS, both Dor *et al.* (2006) and Hadley and Waidmann (2006) found that insurance improved health as proxied by SF-36 type health indices. Similarly, Courtemanche and Zapata (2012) concluded that the Massachusetts health care reform legislation improved a number of health outcomes, including self-reported general health, physical limitations, and a health index. An exception can be found; Pauly (2005) found no significant effect on self-reported general health or the number of adult chronic conditions using the 1996 MEPS.

In contrast, insurance effects vary across condition-specific measures. Private insurance did not reduce the share of infants with low birth weights (Kaestner, 1999) and coverage did not benefit people with chronic conditions more than people without (Dor *et al.*, 2006) while initiation of Medicare coverage improved outcomes for women with breast cancer. In one interesting case, private insurance coverage actually increased obesity prevalence (Bhattacharya *et al.* 2011). One way to reconcile seemingly contradictory results would be to assume that ex-ante moral hazard (in this case more eating, less physical activity and the like) affects some conditions more than others and that the adoption of risky behaviors offsets the health benefits of insurance unequally. The association between health insurance and health behaviors was not explicitly treated in the literature surveyed in this article. Although informative, these findings may not necessarily generalize to other settings given that the efficacy of medical treatment, which insurance enables, is not the same for all medical conditions and diagnoses.

Reconciling Competing Health Measures

Using the setting of transitions into Medicare, an important discussion on the relationship between mortality and generic health measures took place through two interrelated studies (McWilliams *et al.*, 2007; Polsky *et al.*, 2009). In much of the previous literature summarized in **Table 1**, condition-specific outcomes, and generic health measures were treated as mutually exclusive outcomes. A point of agreement in both of these studies is the need to account for censoring due to mortality even when other health outcomes are of interest,

particularly when longitudinal data are employed. Indeed, the two studies using essentially the same database report that attrition due to mortality was responsible for a 15% reduction in sample size during a 13-year period for a population of people around Medicare eligibility. However, although Polsky *et al*. and McWilliams *et al*. agreed on the problem, they disagreed on the methods needed to address it, leading them to engage in a lively debate in subsequent articles.

Although both studies used similar quasi-experimental designs (DD and RDD, respectively, see Section Estimation Methods) and share some findings, their results differ for some outcome measures; the differences have been attributed to the way the authors deal with mortality-related censoring. Both studies compared previously uninsured to insured before and after entering Medicare, and both studies drew the same years and health outcomes from longitudinal HRS data (**Table 1**). Both studies found no significant effect on self-reported health status, mobility, and pain, but differed in their findings for agility and symptoms of depression. In addition, the effect of health insurance was significant for an index of health outcomes only in the McWilliams study. To attenuate sample attrition, McWilliams *et al*. used an inverse probability weighting technique to assign higher weights to individuals who had died on the basis of antecedent health trends, insurance coverage before age 65, and demographic and socioeconomic characteristics. However, this approach may not be accurate given that death is not a random event. To address the nonrandomness issue Polsky *et al*. used a novel approach simulating the predicted probability of health state transitions, with death as one of the included health states. Interestingly, when Polsky *et al*. incorporated inverse probability weighting into their original DD design they found similar results as McWilliams *et al*. This suggests that disparate results were caused by the different ways of accounting for mortality, rather than choice of general technique. The discussion underscores the need for researchers to continue to design innovative and more complete measures of health outcomes.

Vulnerable and Special Populations

The health effects of insurance vary for populations stratified by medical conditions or vulnerability, with vulnerable people benefiting more from health insurance than others. For example, although the RAND Corporation experiment did not find an effect of insurance generosity on the health status of the average adult, insurance generosity did have a positive effect on health for individuals with high blood pressure (Keeler *et al*., 1985). Similarly, private insurance positively affected the health of HIV patients (Bhattacharya *et al*., 2003), Medicaid health benefits were larger when provided at early childhood than at later childhood (Currie *et al*., 2008), and adult patients nearing the Medicare enrollment age with cardiovascular disease or diabetes benefited more from insurance coverage compared with their counterparts with any health condition (McWilliams *et al*., 2007). More generally, Weathers and Stegman (2012) attributed the larger effect of health insurance in the AB demonstration as compared with the Oregon lottery to the relatively poorer health and disability status of persons in the former cases. Further research is needed to identify which patient populations would benefit most from insurance coverage.

Continuity of Coverage: Effect of Churning

A few studies sought to go beyond the simple insured/uninsured dichotomy and evaluated the effects of discontinuities in insurance coverage over time (churning). These begin with a comparative, but mostly descriptive study, (Baker *et al*. 2001), followed by a more rigorous study by Hadley and Waidmann (2006); both studies found that adults who were continuously insured had better health outcomes, as measured by summary health scores, compared to those with intermittent private insurance. Hadley and Waidmann (2006) followed preretirement age adults up to eight years before reaching the Medicare eligibility age of 65, and analyzed the impact of health insurance on health status at that point. Insurance coverage was defined as percentage of time a person has insurance over the observation period before age 65. Although this created certain lumpiness in their insurance measure (the Health and Retirement Study, from which they draw their data, is a biannual survey thus requiring the assumption that people remain in the same insurance category in between survey years) it allowed the authors to estimate effects of continuous insurance versus intermittent insurance. They used similar IVs as those described in Section Estimation Methods to purge their insurance variables from endogeneity bias. McWilliams *et al*. (2007) also report that continuous insurance coverage appears preferable to intermittent coverage for a host of health outcomes. Despite progress made in modeling the dynamic impacts of insurance on health, there appears to be a need for additional research on the intertemporal effects of insurance.

Conclusion

This article highlights the myriad of methods used to estimate the impact of health insurance on health and their limitations. Despite the wide variation in research designs and methods applied, and in particular, the difficulty of identifying valid instruments, a number of common themes can be found. First, it appears that insurance coverage impacts mortality and generic health outcomes more significantly than most condition-specific outcomes, at least in the studies reviewed. Second, certain vulnerable populations such as infants, the disabled, and HIV/acquired immune deficiency syndrome patients appear to benefit from insurance more than the general population. Third, despite the availability of a yet small and largely descriptive body of research on the intertemporal dynamics of insurance, there is compelling evidence to suggest that continuity of health insurance coverage is particularly effective in maintaining health, and that having sporadic coverage offers little protection over little protection over having no coverage at all.

Relevance for Health Reform

With the advent of health care reform, the US appears to be moving closer to universal coverage, albeit not fully. The full

effect of reforms, in terms of reducing the ranks of the uninsured, remains to be seen. A major hurdle in the implementation of the reforms was crossed when the Supreme Court's ruling of June 2012 largely upheld the constitutionality of two major provisions of the ACA: First, the individual mandate and second, the Medicaid expansion (expanding Medicaid eligibility to almost all people under age 65 with incomes at or below 138% of the Federal Poverty Line). The individual mandate requires most people to maintain a minimum level of health insurance coverage starting in 2014; however, the ACA contains several exemptions to the mandate, which allow several millions of Americans to remain uninsured by choice. Moreover, the court's ruling made Medicaid expansion in the ACA optional for the states, and despite the availability of generous federal matching funds, some states have opted not to expand their Medicaid programs. The next hurdle in the path of health care reform and the ACA in terms of moving closer to universal coverage is the design and implementation of state insurance marketplaces (exchanges) that are meant to pool and subsidize employees of small of firms and the self-insured. These marketplaces are intended to be fully functional by early 2014. However, delays are anticipated in many states and participation rates remain to be seen.

Thus, in all likelihood, the debate regarding the value of extending coverage to the uninsured will continue to rage even after the implementation of the ACA. From a methodological perspective, studies on the relationship between insurance availability and health outcomes in the private segment of the US market were hampered by statistical identification issue, making it difficult to ascertain the precise contribution of coverage to health. The anticipated broad expansions of insurance coverage in the US should provide future researchers, opportunities to conduct quasi-experimental studies of private expansions, much like has been done previously in the context of Medicaid and Medicare. It is noted that the debate over this issue is not limited to the effects on health. Other important arguments for providing insurance include efficient use of resources, cost containment, equal access to care, and social protection. These are treated elsewhere in this volume.

See also: Access and Health Insurance. Health Insurance in the United States, History of. Moral Hazard

References

Baker, D. W., Sudano, J. J., Albert, J. M., Borawski, E. A. and Dor, A. (2001). Lack of health insurance and decline in overall health in late middle age. *New England Journal of Medicine* **345**(15), 1106–1112.

Bhattacharya, J., Bundorf, K. M., Pace, N. and Sood, N. (2011). Does health insurance make you fat? In Grossman, M. and Mocan, N. (eds.) *Economic aspects of obesity*, 1st ed., pp 35–64. Chicago, IL: National Bureau of Economic Research.

Bhattacharya, J., Goldman, D. and Sood, N. (2003). The link between public and private health insurance and HIV-related mortality. *Journal of Health Economics* **22**(6), 1105–1122.

Courtemanche, C. J. and Zapata, D. (2012). Does universal coverage improve health? The Massachusetts experience, pp 1–52. *NBER Working Paper Series 17893*. Cambridge, MA: National Bureau of Economic Research.

Currie, J., Decker, S. and Lin, W. (2008). Has public health insurance for older children reduced disparities in access to care and health outcomes? *Journal of Health Economics* **27**(6), 1567–1581.

Decker, S. L. (2005). Medicare and the health of women with breast cancer. The Journal of Human Resources **40**(4), 948–968.

Dor, A., Sudano, J. J. and Baker, D. W. (2003). The effect of private insurance on measures of health: Evidence from the Health and Retirement Study, pp 1–42. *NBER Working Paper Series 9774*. Cambridge, MA: National Bureau of Economic Research.

Dor, A., Sudano, J. and Baker, D. W. (2006). The effect of private insurance on the health of older, working age adults: Evidence from the Health and Retirement Study. *Health Services Research* **41**(3), 975–987.

Finkelstein, A., Taubman, S., Wright, B., et al. (2011). The Oregon health insurance experiment: Evidence from the first year. *NBER Working Paper 17190* Cambridge, MA: National Bureau of Economic Research.

Hadley, J. and Waidmann, T. (2006). Health insurance and health at age 65: Implications for medical care spending on new Medicare beneficiaries. *Health Services Research* **41**(2), 429–451.

Kaestner, R. (1999). Health insurance, the quantity and quality of prenatal care, and infant health. *Inquiry* **36**(2), 162–175.

Keeler, E. B., Brook, R. H., Goldberg, G. A., Kamberg, C. J. and Newhouse, J. P. (1985). How free care reduced hypertension in the health insurance experiment. *JAMA: The Journal of the American Medical Association* **254**(14), 1926–1931.

McWilliams, J. M., Meara, E., Zaslavsky, A. M. and Ayanian, J. Z. (2007). Health of previously uninsured adults after acquiring Medicare coverage. *JAMA* **298**(24), 2886–2894.

Pauly, M. V. (2005). Effects of insurance coverage on use of care and health outcomes for nonpoor young women. *The American Economic Review* **95**(2), 219–223.

Polsky, D., Doshi, J. A., Escarce, J., et al. (2009). The health effects of Medicare for the near-elderly uninsured. *Health Services Research* **44**(3), 926–945.

Thornton, J. A. and Rice, J. L. (2008). Does extending health insurance coverage to the uninsured improve population health? *Applied Health Economics and Health Policy* **6**(4), 217–230.

Weathers, R. R. and Stegman, M. (2012). The effect of expanding access to health insurance on the health and mortality of Social Security Disability Insurance beneficiaries. *Journal of Health Economics* **31**(6), 863–875.

Further Reading

Decker, S. L. and Rapaport, C. (2002). Medicare and inequalities in health outcomes: The case of breast cancer. *Contemporary Economic Policy* **20**(1), 1–11.

Sudano, J. J. and Baker, D. W. (2006). Explaining US racial/ethnic disparities in health declines and mortality in late middle age: The roles of socioeconomic status, health behaviors, and health insurance. *Social Science and Medicine* **62**(4), 909–922.

Health Insurance in Developed Countries, History of

JE Murray, Rhodes College, Memphis, TN, USA

Introduction

Institutional arrangements for health insurance long predate efficacious courses of therapy or even accurate diagnostic techniques. Let health insurance be a type of insurance in which the benefit payment is triggered by an adverse health event. Nowadays the payment is generally intended to pay the costs of health care: physicians and nurses, equipment, drugs, and hospital care. In the more distant past, the primary cost of ill-health was lost income due to the inability to work. Thus, the initial health insurance schemes set out to replace a sick or injured worker's pay, occasionally paid for medical care, and often included death or burial benefits for survivors. Throughout Europe, workers formed sickness funds to execute these sorts of risk management measures. They bound themselves into mutual aid societies that required of them regular payments into a jointly kept fund, from which they were eligible to draw benefits when incapacitated and unable to work. Details of such funds differed according to trade, nationality or region, and over time, but their numbers and the number of members they covered grew until the advent of state-sponsored health insurance in the later nineteenth century.

The Medieval and Early Modern Periods

The first health insurance schemes were established in the late middle ages. Because miners endured the greatest risk of accident and death at work, it was reasonable for them to found the earliest sickness funds. The earliest record of a miners' fund dates back probably to the year of 1300 during the reign of King Wenceslas II of Bohemia. A few early *Knappschaftskassen* (miner society funds) maintained hospitals for members and townsfolk in mining communities, but most aimed to care for widows and orphans of members killed in accidents. Medical care and short-term sick pay, for four to six weeks, generally came from mine owners. There were exceptions, however, such as the mining law for the region surrounding Trier (1546), which called for a fund managed according to insurance principles of compulsory premium and entitled benefit payments. These benefits financed care and sick pay for up to four weeks.

Miners' funds continued to grow into the early modern period, aided by complementary legal developments. Twelve German states made membership in *Knappschaften* compulsory, for example, as in the Prussian *Knappschaftsgesetz* (miner society law) of 1767. A later Prussian law required poor relief claimants in mining regions to exhaust all available benefits from *Knappschaftskassen* before receiving any government benefits. Thus began the connection between private (if government aided) health insurance and state entitlements.

Through locally based guilds, other trades also developed sickness funds in the early modern era. Dutch guilds, for example, established separate relief funds for members in the seventeenth century to protect their general operations from unexpected demands during epidemics. These funds took on the structure that characterized private sickness funds for much of the next three centuries. Although these guilds accepted donations, sickness funds required regular contributions from masters, who in turn were compelled to join the guild in order to practice their trade in a particular location. Guild members, that is, masters, who were sick and temporarily unable to work, claimed a small replacement payment to carry them through their illness. Elderly and otherwise needy guildsmen, widows, and orphans were not entitled to assistance, but might receive some aid when there was enough money in the coffers. Journeymen and apprentices were not eligible for such aid as they were to be supported by the master as long as they lived under his roof. Some guilds developed separate funds for apprentices. As early as 1608 in Antwerp, separate apprentice funds appeared for milliners' and clothmakers' apprentices and journeymen – and Louvain, Brussels, Ghent, and Bruges soon followed. In a few places, Austria in particular, apprentice and journeymen funds survived into the twentieth century. Compulsory membership in apprentice funds ensured a steady flow of new, young, and relatively healthy members into a guild of fund.

In addition to miners and skilled tradesmen in guilds, voluntary occupation-based relief funds appeared as early as the sixteenth century in Amsterdam, Delft, and Leiden, covering the great majority of journeymen and apprentices. The seventeenth and eighteenth centuries saw rapid growth in these funds as well as increasing labor mobility among their members. To prevent financial destabilization and with the aid of local authorities, many local, occupationally based funds instituted compulsory membership; these appear to have been a minority, perhaps between 10% and 20% of apprentice funds. In a nutshell, a substantial share of workers enjoyed membership in sickness funds by the end of the eighteenth century. In Amsterdam, the proportion was approximately one-third, whereas in other northwestern European cities, it ranged between 25% and 30%.

Friendly societies, as sickness funds were known in Great Britain, covered a variety of workers. They also appeared under the name of 'box clubs' to indicate the means of collecting premiums, with a box set to one side of a pub or office. For one century from the later part of seventeenth century, their effectiveness led the elite of the country to call for more such societies to enable care of the poor apart from the provisions of the Poor Laws. Daniel Defoe proposed that the sailors' mutual aid society in Chatham could be taken as a model. By two contemporary estimates, between 7000 and 10 000 friendly societies covered approximately 700 000 members by the year of 1800, or almost one-third of the adult male population. Although all of these friendly societies paid benefits to members, some paid only burial or widows' benefits and not sickness benefits.

Nineteenth Century until 1880

The most celebrated event in the history of government-sponsored health insurance occurred late in the nineteenth century: the German adoption of compulsory insurance. This development did not occur in a vacuum. Throughout the nineteenth century, the evolution of the legal environment and the secular expansion of mutual aid society membership set the stage for direct government intervention. The Napoleonic Wars spread the French Revolutionary animus against guild activity through continental Europe. For example, during the French occupation in Ghent, journeymen's aid societies operated in secret, and they were later joined by elite textile workers' secret funds. When they were allowed to operate openly in 1827, one-third of craft workers were linked to sickness funds. The number of Dutch mutual aid societies grew at this time as well, by 50% from 1800 to 1820, and the number had doubled again by 1850. For profit, commercial insurance, with benefits that replaced pay during sickness or that paid for medical costs, or both, began at this time to cover families having no members in sickness funds. Even before 1842, when physicians themselves started and operated insurance funds that enrolled their patients, nearly one-fourth of Amsterdam's population was insured against medical expenses.

In Great Britain, the friendly societies proved to be popular among all parts of the working class, both skilled and unskilled. At times, the government wished to encourage this trend, but in general active and prospective members resisted this intrusion, managing and expanding their ranks voluntarily. Early in the nineteenth century efforts to encourage county or 'patronized' friendly societies under gentry management came to naught; in 1825, the House of Commons Select Committee on Friendly Societies observed that "people themselves [prefer] clubs managed by themselves." The societies acted creatively to ensure enrollment of future workers. After the 1820s and 1830s, when some Sunday schools organized sickness funds for students as well as teachers, the Oddfellows and Foresters formed juvenile sickness funds from which, they hoped, full members of their societies would emerge.

In the New Poor Law of 1834, Parliament gave an indirect boost to voluntary enrollments of friendly societies. One intention of Poor Law reformers was to encourage the near poor to attain some degree of financial independence through membership in friendly societies. That is, ideally workers would save for hard times rather than hope for relief from Poor Law related institutions such as parish apprenticeship for unwanted children. The reform seems to have had the intended effect, as the number of societies and the number of members rose in the decade after enactment; thus workers do seem to have been saving more. However, the causative role of the New Poor Law in this trend is open to debate. Grim New Poor Law institutions such as Dickensian workhouses bore no less stigma than parish-level outdoor relief under the old Poor Law; both provided substantial incentives to the working class to avoid public assistance.

Friendly societies grew in geographic extent, membership, and operational sophistication through the midthird of the nineteenth century. Regional and national groups of societies known as affiliated orders emerged from individual societies and box clubs. There were 163 such federations by 1877, the largest two of which, the Manchester Unity of Oddfellows and the Ancient Order of Foresters, enrolled some 800 000 men in total. Centralization accompanied the process of growth and affiliation, and central offices began to require the submission of data on sickness claims. From here, it was a relatively easy task to engage in actuarial research to produce tables of claim rates and thus expected probabilities of claim rates and benefit payments in the future. The latter part of the century saw societies moving from customary levels of membership dues to actuarially determined rates, in particular rates being differentiated by age. The ability of societies – or alleged lack thereof – to assess sufficient rates to cover their liabilities became a political issue not put to rest until the 1911 National Health Insurance Act, by which the state financed these liabilities. Notwithstanding the advent of an objective actuarial science, the culture of the societies being closely tied to the local pub and its working-class bonhomie, bound members in a truly mutual fashion to examine their own claims and those of their fellows carefully. Societies sent committees of members out to examine claimants, forbade those claimants from entering pubs, and thereby reminded each man of his obligation to the society as a whole. This solidarity was one characteristic of friendly societies that could not carry on past 1911.

Despite occasional efforts to recruit women and children, the primary aim of friendly societies in nineteenth-century Britain was to cover adult men. Exactly which adult men, in class terms, has been a subject of debate. Earlier historians, such as Eric Hobsbawm, had claimed that premium levels were high enough to discourage unskilled workers from joining, so that the membership by and large consisted of skilled workers – the so-called labor aristocracy. Although more skilled and better paid workers may have composed the majority of friendly society members in the first half of the century, recent microstudies of local club membership rosters have found a broader membership base from mid-century onwards. James Riley compared distributions of occupations of Oddfellows to those of Englishmen as a whole late in the century and found a close correspondence. The representativeness of friendly society membership to British society as a whole was perhaps not evident to historians who relied on official and printed sources, and awaited those local historians who were willing to dig into the manuscript record. In any case probably, the large share of the British population who enjoyed friendly society sickness benefits did not differ in any substantial way from the uninsured population.

Later Nineteenth Century until the Great War

In cultural terms, the German situation was worlds away from that of Great Britain. The English societies had formed out of a tradition of voluntary association whereas the Prussian provident funds, or *Hilfskassen*, stemmed from compulsory organization of artisan trades through the guilds. That compulsion signals the importance of the state in the development of German funds. Care for ailing and injured journeymen played a particularly important role in the German case. Journeymen had neither the access to resources that masters had nor were they under the responsibility of a particular

master, unlike apprentices. As the German states disabled guild influence over members, they required guilds to provide closer assistance to journeymen in need. For example, the Prussian industrial code of 1845 included enabling laws that permitted local authorities to require all journeymen in their jurisdiction to belong to journeymen's sickness funds. The growth of other workers' funds was limited to opportunities left open by the lack of state action. In Germany, as in Britain, workers' sickness funds interacted with Poor Law institutions. One reason for the 1845 enabling law was the Prussian Poor Law legislation of 1842, which shifted the focus of benefit payments from the person's original place of settlement to his current place of residence. As local communities could no longer rely on guilds to care for distressed journeymen, they were granted powers by the 1845 law to shift that obligation back to the guilds. Between 1849 and 1853, some 226 Prussian municipalities made joining a sickness fund mandatory for workers. Although compulsory membership in sickness funds appears in the historiography as a reaction to the tumult of 1848, it is noteworthy that legal requirements to join these funds have predated the events of that year. A later law, the Emergency Ordinance of 1849, allowed local governments to compel factory workers to join provident funds, to which factory owners were required to contribute, thus placing artisans and factory hands on roughly similar legal footing. Again, the main concern was protecting local poor relief institutions.

Changes in the legal environment around the middle of nineteenth century affected all manner of sickness funds. A rising tide of internal migration, especially well documented in the lower Rhine region surrounding Düsseldorf, concerned local authorities who feared the newcomers would end up on their own poor rolls. The central Prussian government, however, committed itself to freedom of movement. Rather than restrict labor mobility, in 1854 it allowed local communal funds to bill the commune of birth or previous residence of a poor relief recipient for up to a year. The number of funds grew as a consequence. At the same time, the Prussian government dramatically changed its laws regarding property rights to underground resources, with consequences for the miners' funds, the *Knappschaften*. Mine owners, rather than the state, would control the disposition of assets and hold the ability to hire and fire miners and to determine their pay rates. Miners' funds could now set member contributions either as a flat percentage rate, or as a flat rate within a set of fixed categories corresponding to earnings. The rise of Liberal political influence in the 1860s led to the founding of labor union provident funds, which continued under Social Democratic influence. These funds were eager to be treated as others were; that is, to keep their members from being compulsorily enrolled in guild, factory, or communal funds, and thus paying twice over for their insurance. After the 1866 Prussian annexation of territories into the North German Confederation, the *Reichstag* did in fact issue an Industrial Code in which compulsory membership requirements could be met by joining a 'free' or voluntary membership fund, such as those operated by labor unions.

Over the course of the 1860s, the status of voluntary and compulsory funds, those 'free' of government regulation and others overseen by local government or business officials, those operated by trade unions that permitted member mobility and communal funds that did not, became muddled. Court opinions contradicted one another, and the confusion led communal authorities to cease requiring residents to join provident funds. The central government found the wide range of premiums, benefits, and claim requirements unsatisfactory, but politically untouchable at that time. One result of this relatively *laissez faire* approach to insurance regulation was the outburst of growth in both the number of funds and the number of workers they covered. Between 1860 and 1870, the number of funds for skilled craft workers rose by 29% whereas the number of covered craft workers rose by half. Over the same period, the number of funds for factory workers doubled, as did the number of such workers who were covered.

In 1876, the central government of what was then the German Empire finally achieved its goal of standardization through the Law on Provident Funds. At least nominally, the legal requirements of similar benefits, and thus similar premiums to pay for them, among the voluntary funds meant that they could not be used as an expedient method of avoiding the higher-priced, higher-benefit level compulsory funds. To do this, the central government created a new category of registered funds, membership in which might be either compulsory or voluntary, and benefits from which were strictly regulated within certain minimum and maximum time periods and levels. In addition, these funds were required to end their provision of benefits for anything other than sickness and injury, such as death benefits for widows and orphans, or partial pay benefits for striking members. Besides forbidding fund members to participate in strikes, this law also forbade investments of fund reserves in the sponsoring firm, thus detaching the funds from both workers and employers in one stroke.

For various reasons, the state's interest in health insurance regulation did not end there. In a strategic view, Bismarck wished to soften the blow of the first Anti-Socialist Law of 1878 among the working classes, and to co-opt them into believing that the state, rather than removing their political voice, was providing for them materially. This explains Bismarck's initial efforts to fund compulsory insurance through employer contributions and taxes: "If the worker must pay, the effect on him is lost," he said, because then the worker could see that he himself and not the state had produced the resources that paid for the benefits. From a more tactical perspective, the need for widespread (but not universal) compulsory health insurance arose from gaps in the current state of accident insurance that stemmed from the Accident Liability Law of 1871. Efforts to update the state of accident insurance stalled in the 1882–83 session of the *Reichstag*, and so the relatively uncontroversial provisions for health insurance were removed and placed in a separate bill. With accident insurance to be made compulsory across the Empire by a bill that assigned responsibility for the first several weeks of disability to the sickness insurance funds, it would not do to have pockets of worker autonomy concerning sickness insurance. Hence, the 1883 Sickness Insurance Law entered the books before the 1884 Accident Insurance Law.

The new Sickness Insurance Law built on the existing network of small sickness funds. It made membership in a

sickness fund compulsory for a large class of workers who earned less than 2000 marks per year. In addition, employers contributed to sickness funds at a rate of one mark for every two paid as dues by the employee, but there was to be no state funding. By inspecting employer records, cross-checking fund membership lists, and threatening employers of uninsured workers with fines, the state effectively enforced coverage requirements. For workers who toiled in other sectors, such as agricultural laborers and domestic servants, and for those workers who earned more than 2000 marks annually, membership was voluntary. Registered aid and state-registered aid funds covered those outside the compulsory system who chose to join voluntarily. The network of health insurance funds covered a large share of the working class though not immediately. Enrollment in 1885 numbered some 4.5 million, or almost a tenth of the population; by 1906, the covered share of the nonagricultural labor force (not population) had risen to approximately 70%.

Despite the broad extent of coverage, the systems still confronted problems of moral hazard and physician agency. The statutory minimum wage replacement rate was one-half, but many funds paid 60% or 70% of a worker's usual earnings to disabled members. One consequence was a steadily increasing number of missed workdays due to sickness absence. In 1885, the first year with comprehensive statistical data, the average covered worker missed six days of work due to ill-health. By 1908, that number had risen to nine days per year, an increase of 50 percent. Over this time, workers did not change the rate at which they submitted claims, so the increase in sick time was due to longer spells of sickness. For example, in establishment funds operated by particular firms, the average duration of illness rose from 12.5 days in 1885 to more than 18 days in 1908, an increase of nearly a week per sickness event. The upward trend was not affected by the 1903 law that increased the mandatory maximum duration of insured sick time from 13 to 26 weeks. Among funds in which membership was compulsory, both the frequency of sickness spells and their duration were strongly and positively associated with the level of sick pay, suggesting a moral hazard in which the availability of sick pay increased the time spent off work. Indeed, the German-American statistician Frederick Hoffman proposed that the fundamental problem behind increasing absenteeism among insured German workers was not their worsening health but rather their 'dishonesty, deception, and dissimulation' regarding missed work time.

Similar problems appeared in miners' sickness funds. Up to the turn of the twentieth century, miners had averaged between six and eight days per year of absence, but after 1900 or so, that figure jumped to as high as 12 days per year. As one observer noted, "[F]requent malingering…in the Ruhr area led to a great increase in costs." In response, *Knappschaften* ended sick pay for Sundays and introduced waiting periods that acted as a kind of deductible. Still, here too, later research found a strong, positive, and significant correlation between sick pay and absenteeism rates.

To deal with these problems, the system developed an elaborate process of obtaining second opinions. To receive sick pay benefits in the first place, a worker's claim needed to be approved by a physician associated with the fund. In German funds that offered free choice of physician, fund-employed doctors monitored independent physicians by performing second examinations. Both funds and their members enjoyed the right to demand a second opinion from a variety of 'confidential medical advisors,' either fund-employed physicians or committees were composed of physicians' and insurers' representatives. The results of the second opinions suggested that the physician-agent's diagnosis depended on the identity of the principal. Given free choice of physician paid by capitation, as in most compulsory funds, patients were the principals, and physicians who gave initial diagnoses of incapacitation were their agents. Medical advisors who monitored the primary physicians, on the other hand, were agents of the insurers. Probabilities of claim approval reflected these relationships. A report of fund groups in several northern cities during 1909 and 1910 indicated that whereas initial consultations tended to favor the worker, second examinations favored the fund. Between one-eighth and one-third of workers who had obtained statements from their own physician attesting to their incapacitation returned to work rather than be examined by a fund doctor. These workers either recovered quickly or lacked confidence in the veracity of their claims. German workers, physicians, and their supervisors all understood the implications of agency. Physicians wanted to keep even their most annoying patients, who frequently presented with dubious symptoms, in order to maintain the capitation fees that accompanied them. Contemporary observers asserted that personal physicians thus gamed the system by approving questionable claims. The fund's medical advisors then routinely rejected the claims at the second opinion stage, thereby keeping the fund financially healthy and the attending physician's pay intact, while allowing him to blame the second physician for the rejection.

During the later nineteenth century, other forms of health insurance expanded their coverage in continental Europe. With the exception of sickness funds for miners in France, membership in them was voluntary rather than compulsory. And that membership grew. French membership in adult funds, which accepted a measure of government supervision, tripled to 2.5 million from 1886 to 1905, whereas free funds, which operated without such oversight, grew by more than a third to 425 000. Similarly in Belgium, recognized funds under government oversight grew nearly ten-fold to a quarter-million members from 1885 to 1904. In Denmark sickness societies, heavily subsidized by the government, tripled their enrollments between 1895 and 1905, with another 20 percent growth by 1907.

These sickness funds managed a different set of problems from the compulsory German sickness societies previously discussed. All voluntary funds faced the threat of adverse selection, including the voluntary German funds that descended most directly from Poor Law institutions. To manage the problem of cultural differences in determining whether a worker was too sick to work, absence records from both voluntary and compulsory funds within Germany were compared to each other in the region of Leipzig. Here, membership rolls in the voluntary funds skewed older than those in compulsory funds, which suggested selection biases into membership. But then controlling for the age categories of members, voluntary funds had much higher absenteeism rates than compulsory funds among same-aged workers, which suggests that the

voluntary funds were especially attractive to those in poorer health at every age. Members of voluntary funds who were in their early 20s had extraordinarily high sickness rates, nearly as great as those of men in their 60s. A contemporary German observer has explained the reason in a classic adverse selection statement: "Practically all the male population, including the weaker and those who are physically less valuable, are sent to work in the earlier ages [i.e., and then they join compulsory funds]; in a few years, however, the weaker persons must give up the occupations in which they are engaged, but realizing their need for insurance, continue their membership as voluntary members."

In voluntary French and Belgian funds, such difficulties were compounded by the financial need for 'honorary' members. These were civic-minded men of the bourgeoisie whose membership required them to contribute premium payments but did not allow them to claim benefits. Their presence in sickness associations diluted the solidarity among rank and file members that was necessary for them to function efficiently. Both France and Belgium relied on mutual aid societies to care for sick workers through their benefit funds, with the few workers employed by large firms enrolled in establishment funds. A manual for sickness fund managers addressed a widespread concern with selection bias by recommending rejection of all applicants over the age of 40 due to "the risk of illness [being] considerably augmented after that age." French benefits were in line with those elsewhere. A large fund for store clerks in Paris charged its members two francs per month in dues and offered sick pay benefits of two francs per day for not more than 60 days plus the attention of a physician employed by the fund. Belgian funds were less generous. One coal mining company fund replaced only 22% of a miner's pay, but paid these benefits for the first six months of illness. Dependence on scarce honorary members kept Belgian dues relatively high, leading a contemporary to complain, "It is the élite of the working class alone that can stand the cost of sick insurance."

Financial problems plagued French and Belgian sickness funds as memberships aged and claim rates rose beyond the ability of fund assets to service. French establishment funds became even more dependent on subsidies from sponsoring firms. Among all French funds, the value of assets per participating member (excluding honorary members) fell by one-quarter from 1898 to 1905, whereas this measure rose by 10% among compulsory German funds. The historian Theodore Zeldin summarized the situation of the French societies thus: "Ignorance of the principles governing insurance was common, methods of administration amateur in the extreme….The most serious omission was that the whole movement was never established on an actuarial basis." Similarly, Belgian funds endured chronic financial difficulties due to their lack of actuarial soundness. A government official at the time conceded that the societies' sick funds could, in theory, "be scientifically managed," but in fact "the mutual sick-benefit societies do not fulfill the necessary requirements of a safe and rational organization." These difficulties led Catholic and Socialist legislators to agree on the need for compulsory insurance in 1912.

Consequences of sickness insurance benefits varied according to the voluntary or compulsory nature of membership. As noted above, availability of sick pay seemed to induce German workers to take additional days off, and the pattern of increasing sickness time appeared in other compulsory funds as well: in Austria and among German and French miners. Whether those days were truly evidence of malingering, or whether workers could finally afford to take necessary time off work to recover, cannot be determined from statistical analysis. Among workers who belonged to voluntary funds in France, Belgium, and Denmark, however, after about 1890 paid absenteeism began a slow and steady decline for some years. This trend is unlikely to have been caused by improving worker health. Rather, it stemmed from the financial inability of these funds to support previous levels of absenteeism benefits. French physicians, employed directly by sickness societies, ceased to approve absence benefits so readily after being ordered by fund managers to cut costs. Later, in the 1930s, Belgian funds adopted denial of benefits as an explicit policy to keep their accounts in balance. Statistically, greater expenditures per sick day on medical benefits were associated with briefer spells of absence, which may have been due to physician visits resulting in orders to return to work, at least among the voluntary funds. The French physician and statistician Jacques Bertillon wrote in 1892:

> The fact is that when these societies grant compensation they attach less importance to their regulations than to the state of their till. A rich society gives its help more liberally than a poor one; and this is absolutely the sole cause of the large English societies, which are often very old and generally rich, granting more daily indemnities than the French (for instance), who are obliged to exercise the strictest economy.

Given the limited efficacy of therapeutics in the late ninteenth and early twentieth century, the primary benefit of sickness insurance coverage was the sick pay benefit that enabled workers to take time off to recuperate. This rest enabled workers to recover from illness and injury sufficiently regularly to influence mortality rates. Various studies had found that more expansive sickness insurance coverage, whether compulsory or voluntary, was associated with reductions in mortality rates in general. In particular, infant mortality rates were also lower as coverage expanded, probably as a result of confinement benefits. Those benefits also led to relative increases in fertility rates. Finally, persuasive evidence has been adduced to show that the availability of sickness insurance in Germany had reduced the rates of emigration at the turn of the century. Thus, health insurance had measurable influences on all manner of demographic measures throughout early twentieth-century Europe.

Growth of health insurance (as it came to be called) in Great Britain trod its own path quite different from developments on the European continent. The German government was committed to elaborate intervention into, but not subsidies for, health insurance markets, and the French were equally committed to upholding a worker's choice of joining a benefit society or not. In the British case, a far larger share of working class men belonged to friendly societies than in France or even in Germany before 1883, which mitigated the perception that government action was needed to insure workers and also created a formidable political barrier to such action. The Liberal government launched its welfare reforms

only in 1906, because until that time the great concern had been to care for the elderly who had simultaneously been pushed out of the labor force by younger workers and pushed into the embarrassment of outdoor relief. How exactly to deal with the deserving aged poor remained a conundrum until the 1908 Old Age Pensions Act provided tax-financed pensions to the elderly. This landmark Act thus moved the responsibility for care of the elderly from local Poor Law Guardians to the national government.

The unusual calls for two general elections in 1910 gave the government time and space to consider the next step of compulsory health insurance. In 1907, a young William Beveridge suggested that provision of unemployment insurance could potentially mitigate a great deal of poverty, and then in 1908 after passage of the Old Age Pensions Act, David Lloyd George visited Germany for five days to study the possibilities for a similar national health insurance program in Britain. The combination of these two events led to the National Insurance Act of 1911. The health insurance aspect of the Act, as distinguished from its unemployment insurance provisions, was to be funded by weekly contributions. Unlike in the German system, these contributions were fixed as flat rates, thus imposing more of a burden on lower-paid workers. Employed men paid four pence, employed women three pence, their employers three pence, and the state two pence weekly. Coverage automatically applied to all manual laborers and to all over age 16 who earned less than £160 per year, the equivalent of 3200 marks. Insured workers could obtain free medical care from a physician who belonged to a local medical committee. Workers were eligible for a sick benefit of 10 shillings per week for men (seven shillings sixpence for women) for up to 26 weeks. After 26 weeks, an ill or injured worker might apply for a disability benefit of five shillings per week.

As the Bill proceeded through Parliament, it changed considerably. Originally, Lloyd George had intended for friendly societies to perform much of the administration of this insurance, but concluding that commercial insurers were much sounder in actuarial terms, he shifted the load of management toward them. During consideration of the National Health Insurance Bill in 1911, the British Medical Association persuaded the government to allow free choice of physician as part of a larger development that excluded more approved friendly societies from the system. Thus the great distinction today between German management of health care finance, where insurance funds determine the levels and distribution of expenditure on health care, and the British method, wherein such decisions are made by the state, is one that dates back to the early twentieth century.

After 1918

The British economy staggered out of its victory in World War I into an uneasy peace. In 1919, the earnings limit for mandatory insurance increased to £250, almost keeping pace with wartime inflation. The next year contributions rose to five pence for both men and women, and the standard benefit was increased to 15 shillings per week for men and to 12 shillings for women. Over the entire interwar period, the share of the male population entitled to benefits rose steadily from 51% to 63%; the associated share of women rose from 23% to 30%. During the high unemployment era of the 1930s, the sick pay benefit offered through the national health insurance program began to look better for workers when the comparable benefits available through unemployment insurance and workmen's compensation (accident insurance) expired. One result was that workers who became unemployed tended to make claims of ill-health against the national health insurance plan when their unemployment benefits ended. Thus, as unemployment rose during this period, so did sickness claims. From 1921 to 1927, sickness claims by men rose by almost half, as did long-term disability claims. In actuarial terms, the ratio of actual to expected costs of disability benefits for men increased by 80% in Britain between 1922 and 1935. The possible substitution of sick for unemployment benefits produced an acute strain on the insurance program's finances.

In May 1940, the Chamberlain government fell after the loss of Norway to the Germans. Only a year later, the coalition government led by Winston Churchill appointed William Beveridge to chair a new committee on the reform of social insurance. Beveridge's famous report of 1942 determined the course of the British welfare state for a generation after the war. It aimed to create a unified system of social insurance for the entire population, and not just manual workers. The safety net was to cover workers and their dependents against ill-health, unemployment, and old age, and was to be financed through general taxation funds. In the wake of successive reports from the Committee on Medical Insurance and Allied Services (1920), the Royal Commission on National Health Insurance (1926), and the British Medical Association (1930 and 1938) that emphasized the shortcomings of the existing arrangements, the Beveridge Report recommended replacing compulsory insurance for most workers with a comprehensive national health service for the entire population. British physicians fought the imposition of a salaried state medical service right up to the formal establishment of the national health service in 1948.

In France, settlement of the Great War undermined French notions of individual choice of insurance from within. After the Franco Prussian War of 1870–71 the German Empire annexed the former French Alsace-Lorraine. Inhabitants of the region were integrated with the German project of compulsory sickness insurance from its start, and by the time of the Treaty of Versailles, they were in no hurry to return to the status quo of 1870. In response to the threat of an independence movement, the French government promised Alsatian labor unions that it would maintain health, disability, and old age insurance substantially as they had been, and hinted at even using these arrangements as a potential model for the rest of France. French physicians aimed to prevent such developments, but eventually they compromised with the government and allowed the first form of compulsory insurance to be established in 1930. This insurance reimbursed patients for 80% of their medical bills. The downside of this agreement was that individual physicians felt no compulsion to abide by fee schedules negotiated on their behalf by medical groups. The share of covered population (not labor force) rose to almost 25%, but unexpected expenses and denials of benefits increased political discontent with the scheme.

The next step in French insurance policy occurred during World War II. It was conceived not in France itself but by the

Free French government in London, and then enacted in 1945. The necessary relationship between employment and insurance coverage ended, thereby enrolling greater numbers of the insured. In qualitative terms, this expansion of the *Sécurité Sociale* also proposed to limit increases in physician billing rates. By some accounts, this represented a missed opportunity to do away with fee for service medicine altogether and leap ahead to the system that began to be implemented after the 1960 reforms. Still, the postwar reforms succeeded in bringing 'the quasi-totality of the population' under coverage – a Gallicism meaning almost three-fourths, roughly same as the share of Americans with hospital insurance. But again, costs rose faster than expected, making it impossible to keep the French budgets in balance.

The Dutch interwar experience offered a fine example of the ability of a totalitarian government to break legislative deadlocks and impose politically unpopular compulsory insurance. By the end of the nineteenth century, a wide variety of sickness insurance funds was operating in the Netherlands: some formed as mutuals by groups of workers, others sponsored by employers or trade unions, still others by local governments, a few operated by commercial insurers, and a unique set of funds were operated by physicians. And here things stayed due to Parliamentary impasses. From the Great War onwards, every effort to enlarge the government's presence in health insurance markets halted due to unwanted amendments, parliamentary deadlock, dissolved governments, and other flotsam of a democratic polity. The arrival of Nazi occupation forces ended the stalemate. To bring the Netherlands in conformity with the German example, the occupiers promulgated a compulsory sickness fund decree that broke through the parliamentary clutter and established government health insurance once and for all. As for Belgium, the Allied breakout from Normandy caused the Germans to put Nazifying health insurance on hold. But soon after liberation, the Belgians too enacted compulsory insurance. Thus in the Low Countries, both occupiers and the occupied looked upon government health insurance as an idea whose time had arrived by the mid-twentieth century.

Elsewhere in the world, the rise of government intervention in health insurance markets awaited the second half of the twentieth century. In the middle of this century, the Canadian situation was in flux. Canadian physicians had become more sympathetic than their American counterparts to the prospect of state action, and the Canadian Medical Association was participating in the reform process. Creation of government insurance occurred first in the West, where Saskatchewan, British Columbia, and Alberta had adopted a tax-funded hospital insurance program. Newfoundland had already created a health insurance program that covered half the population by the time it entered the Confederation in 1949. The success of government hospital insurance in these provinces led to the Hospital Insurance and Diagnostic Services Act of 1957, by which the federal government subsidized hospital insurance in all the provinces. Pushing the principle of state insurance further, the provincial government of Saskatchewan established Medicare, as the Canadian single payer medical insurance system came to be known, in 1962. This triggered a bitter and ultimately unsuccessful strike by the province's physicians. The strike's failure caused a loss of political capital by the most important opponents of an expanded government role, and this in turn opened the door to further state intervention. The pressure for national health insurance became so great that even the physicians did not want to be seen in opposition to it, and they again moved to work with governments on the shape of insurance policy. Pushing the principle of state insurance further, the provincial government of Saskatchewan established Medicare, as the Canadian single payer medical insurance system came to be known, in 1962.

Nor has the notion of health insurance been restricted to only Europeans and their descendants. Compulsory health insurance for industrial workers began in 1950 in Taiwan, in part as a political effort to improve the protection from the risk of ill-health enjoyed by Taiwanese workers relative to those in the People's Republic. From its initial remit of coverage for workers in public factories and mines, the government expanded this health insurance to workers in private in private industry, smaller manufactories, and fisheries by 1953. Beginning in 1958, it extended compulsion to government workers and teachers, and then all industrial workers, and eventually nearly all workers, including those in agriculture. By the time of national health insurance in 1995, there were few uninsured Taiwanese remaining.

In Latin America, the more prosperous countries have succeeded in enrolling a large share of the population in health insurance of some kind. By 1986, Argentina, Brazil, Costa Rica, Mexico, Panama, Uruguay, and Venezuela offered medical care coverage to 71% of their combined populations. The covered populations tended to be city-dwellers, who were relatively easy to reach and relatively able to afford the premiums. Five of these countries covered spouses and children of the insured, and of the remaining two, Uruguay provided maternity and pediatric care whereas Panama excluded only hospital care from coverage. The origins of these programs date to much earlier in the twentieth century. For example, in the 1920s, Brazil created a variety of social insurance funds for various kinds of workers in different parts of the country. Over the next several decades, legally mandated amalgamation reduced the number of social insurance funds to seven large funds that represented major occupational groups, including rural workers.

See also: Health Insurance and Health. Mandatory Systems, Issues of. Private Insurance System Concerns

Further Reading

Companje, K. P., Hendriks, R. H. M., Veraghtert, K. F. E. and Widdershoven, B. E. M. (2009). *Two centuries of solidarity: German, Belgian and Dutch social health insurance 1770–2008*. Amsterdam: Aksant Academic Publishers.

Dutton, P. V. (2007). *Differential diagnoses: A comparative history of health care problems and solutions in the United States and France*. Ithaca: Cornell University Press.

Frohman, L. (2008). *Poor relief and welfare in Germany from the reformation to World War I*. Cambridge: Cambridge University Press.

Guinnane, T. W. and Streb, J. (2011). Moral hazard in a mutual health insurance system: German Knappschaften, 1867–1914. *Journal of Economic History* **71**, 70–104.

Harris, B. (2004). *The origins of the British welfare state: Social welfare in England and Wales, 1800–1945*. Basingstoke: Palgrave Macmillan.
Hennock, E. P. (2007). *The origin of the welfare state in England and Germany, 1850–1914. Social policies compared*. Cambridge: Cambridge University Press.
Hoffman, F. L. (1920). *More facts and fallacies of compulsory health insurance*. Newark, NJ: Prudential Press.
Hye Kyung Son, A. (2001). Taiwan's path to national health insurance, 1950–1995. *International Journal of Social Welfare* **10**, 45–53.
Khoudour-Castéras, D. (2008). Welfare state and labor mobility: The impact of Bismarck's social legislation on German emigration before World War I. *Journal of Economic History* **68**, 211–243.
Murray, J. E. (2005). Worker absenteeism under voluntary and compulsory sickness insurance: Continental Europe, 1885–1908. *Research in Economic History* **23**, 177–208.
Murray, J. E. (2007). *Origins of American health insurance: A history of industrial sickness funds*. New Haven: Yale University Press.
Riley, J. (1997). *Sick, not dead: The health of British workingmen during the mortality decline*. Baltimore: Johns Hopkins University Press.
Whiteside, N. (1987). Counting the cost: Sickness and disability among working people in an era of industrial recession, 1920–1939. *Economic History Review* **40**, 228–246.
Yamagishi, T. (2011). *War and health insurance policy in Japan and the United States: World War II to postwar reconstruction*. Baltimore: Johns Hopkins University Press.
Zschock, D. K. (1986). Medical care under social insurance in Latin America. *Latin American Research Review* **21**, 99–122.

Health Insurance in Historical Perspective, I: Foundations of Historical Analysis

EM Melhado, University of Illinois at Urbana–Champaign, Urbana, IL, USA

Introduction

Although the US, in comparison with other Western countries, was a latecomer to social insurance and the public provision of insurance for health services, it was largely in the America of the 1960s that formal economic analysis of health care first began to take root, and American ideas and practices have long since dominated health economics; hence American ideas are the focus of this and the next article. The efflorescence of American health economics emerged from (and helped alter the course of) antecedent traditions of American thought about health insurance, which began in the early-twentieth century. For a bit more than the first-half of its history, ideas about health insurance took form in and evolved from the work of two overlapping groups of analysts: a broader one, whose members took a normative perspective animated by questions of social politics; and a smaller group whose members aimed more specifically to improve public health. Figures in both were reformers and activists; they hoped to advance what they understood to be the public interest. Their normative vision little exploited formal economic analysis, which, at least in its modern, mathematized mode, was at that time only incompletely developed and thus unavailable to reformers as a basis for analyzing health policy; but the aspirations of the emergent social sciences often informed their vision.

Only during the 1960s, under the then prevailing liberal dispensation, when a significant social surplus was available to sustain expanded forms of collective provision, did formal economics of a sort more familiar to modern practitioners begin to make itself felt in application to public policy. Economists developed formal rationales for governmental involvement in the economy and articulated the principles that should govern public programs. In the case of health policy, figures such as economist Kenneth Arrow (1921–) held that much of health care qualifies as a special set of services that require collective subsidy (if not indeed public provision). This agenda gradually fractured, however, as diverse forces, both inside and outside economics, undermined a once broad faith in the value and propriety of governmental intervention in the economy, in the capacity of experts (particularly those in governmental employ) to achieve desirable goals, in the utility of regulatory regimes, and in the capacity of society to gain consensus about the goals of public policy. Convictions wavered even about the value of health services, at least at the margin. As economics developed and honed the tools to analyze public policy, analysts toned down, but did not abandon, normative orientations, and the role of the economist and expert became less that of reformer and more that of the servant of diverse interests well beyond the traditional ranks of policymakers. Gradually, the major concern became markets, at first as the best means to realize broad social goals, and later, as commitments shifted away from fostering collective provision, to serve and facilitate individual choice.

Advocates of older normative views have hardly disappeared, but their approaches, reflecting social and cultural traditions that had been eroding since especially the 1970s, have been routinely contested by advocates – both within economics and without – of markets and limited government. The public debates that preceded the Obama reforms, i.e., the Patient Protection and Affordable Care Act (PL 111–148, as amended by the Health Care and Education Reconciliation Act, P.L. 111–152; henceforth, ACA), passed at the end of March, 2010, and the persistence thereafter of pervasive disagreement in health care about the goals of public policy and the roles of government, show that both economists and Americans broadly remain profoundly divided about these questions. Many economists, by professional interest and training concerned with markets, have often presented themselves as representing a value-free perspective on questions of public policy; yet many of their critics, including other economists more rooted in traditional approaches to policymaking, find that their colleagues' claimed neutrality implicitly harbors values inimical to those rooted in older approaches, that they still sought to honor as collective commitments. Economic analysts of health insurance and related areas, despite the drift of the profession in favor of markets, still reflect the diversity of values and beliefs about the proper goals and means of public policy; neither health economics writ large nor the parts of it most directly connected with insurance have eliminated this diversity, but they have provided powerful and influential frameworks for defining, discussing, and analyzing the issues.

This article opens with discussion of several historical frameworks useful in exploring the history of American thought about health insurance and provides elements of a taxonomy of both researchers and advocates of various forms of health insurance. It then describes the earlier history of thought and advocacy in the context of the social politics that emerged in the early-twentieth century and persisted until the late-1960s and early-1970s. It takes up underlying notions of social solidarity, their tensions, and their relevance to health insurance. It next exhibits the emergence of a market-oriented perspective and its corrosive effects on ideas about social solidarity. The following article explores the two principal bodies of thought that called for market-based approaches to health care, notes their early connection with calls in the late-1960s and 1970s for expanded public insurance, elicits the main elements of these traditions, and links discussion with contemporary developments, particularly in the light of the evolution of markets on the ground. The article concludes that in America health economics has done much to enable analysts to formulate and analyze policy questions, but that policy

discussion about health insurance remains highly contested. What is clear is that the US, despite recent reforms, is not moving toward a uniform system of National Health Insurance (NHI), but continues to fragment care and coverage, organizing subsidies by income, race (through the proxy of poverty), and age. What is at stake for the future is thus not this fragmentation; but the extent to which recent reforms that aim to expand entitlement and improve benefits will survive the vagaries of administrative complexities and future political developments. Economists, meanwhile, continue to dominate analysis of these policy questions.

Historical and Conceptual Frameworks

Various frameworks have been proposed for understanding the history of health insurance in the US; this section takes up three of them. In one, Daniel M. Fox has characterized three normative models for research on health care and health policy, social conflict, collective welfare, and economizing, this last having eclipsed the former two, especially since the 1980s. In another, Paul Starr divided the field into three eras during the twentieth century, according to the ways in which advocates of health insurance addressed the costs of sickness, direct and indirect, individual and social; and in his recent book, he has revised this model in the light of subsequent developments. In a third, Deborah Stone drew attention to persistent conflicts in American insurance arrangements between 'the solidarity principle' and 'actuarial fairness,' that is, in terms that describe the opposing social and economic functions that insurance has been taken to perform.

Advocates of the two older models in Fox's scheme were those possessing knowledge of the nascent social sciences and used them in support of expanding health services and improving access to them. Under social conflict, researchers held that health services, like food, clothing, and shelter, are essential; but that those better-off (or dominant) tend to withhold them from those less well-off (or belonging to socially subordinate or marginal groups). Expanding access and improving benefits under social conflict therefore became the subject of struggle on behalf of the poor and vulnerable; research aimed, inter alia, to document lack of access, its causes, and its consequences. Under collective welfare, researchers regarded health services as special, because they determine personal wellbeing if not indeed survival, and that attitudes of social solidarity, rather than conflict, require cultivation to bring more of the benefits of medicine to more people. Research tended to exhibit, inter alia, the consequences for health of diverse levels and goals of expenditure. Both models reflected not only a conviction, owing to the scientific innovations that began in the late-nineteenth century, that health services were effective, but also a commitment to social politics to provide citizens with shelter, in policy areas thought of fundamental importance, from the market arrangements that otherwise prevailed in economy and society.

Under the recently ascendant economizing model, in contrast, researchers have thought of care as largely similar to other commodities, best organized through markets, and they have regarded research as best conducted by exploiting economics (and several other sciences, especially epidemiology and biostatistics). Research has concerned the effectiveness of services, the functionality of reimbursement mechanisms and institutional arrangements, and the means to minimize the costs of expensive programs and to structure and fine-tune markets to improve efficiency and opportunities for choice. Under the economizing model, researchers have adopted a less openly normative posture, aiming less to press for new programs than to analyze for policymakers what exists and how (in the light of policymakers' values) it might be improved. In shifting from the older models to the economizing one, researchers, as Fox had once put it, moved from reform to relativism.

The two older of Fox's models dominated the first two of the three eras that Starr marks out in the history of health insurance. The earliest, that of 'Progressive Health Insurance,' represents the body of ideas that was prominent in the American Progressive Era (roughly, 1900–20) and that focused on sickness as one cause of poverty (via the consequent interruption of wages to workers and their families), as well as on the social causes of sickness. 'Sickness insurance,' as it was initially called, to be provided on the state level, would serve workers as a cushion against lost wages and, through its financing, create incentives to exploit public-health measures and industrial reforms that would reduce the extent of sickness and thereby improve national efficiency. Starr's 'Expansionary Health Insurance,' dominant in the period from the 1930s to the 1960s, marked a redirection of researchers' concern from lost income and public health to the direct costs of care. Introduced especially by the work of the Committee on the Costs of Medical Care (CCMC), active from 1927 to 1932 under philanthropic support, they focused on the rising costs of medical care (especially hospital services), owing to scientific innovation, and on the inability of both working and middle classes to meet them (particularly in view of their highly unequal incidence); but, in view of the benefits of care, called for insurance both to cover costs and to expand the health system.

This same period witnessed the first appearance in the US of significant programs of voluntary health insurance. These programs, did not, however, stop reformers from pressing for NHI in the 1940s and beyond. Around 1930, Blue Cross plans, fostered by the American Hospital Association, provided hospitalization insurance initially to employee groups and, starting roughly a decade later, Blue Shield plans, under the control of medical societies, provided insurance for physicians' services. Governmental policies emerging in the war years encouraged the spread of voluntary insurance (inter alia by permitting collective bargaining over fringe benefits, by making health insurance a fringe benefit untaxed for employees, and by allowing employers to deduct the costs of insurance from their taxable incomes). The labor movement, although in principle committed to public provision, nevertheless preferred this privatized form of the welfare state. Reformers still nurtured hopes of creating NHI, and especially from the 1940s they repeatedly tried and failed to secure it. Only in 1965 did they achieve a partial victory with the passage of Medicare (a federal program that provided health insurance for the elderly) and Medicaid (a federal-state program that provided insurance for some of the poor), as new titles under the Social Security Act of 1935. Medicare largely

reflected a social-insurance approach, but Medicaid, enacted as a reform of antecedent welfare programs, lay in the world of welfare and public assistance.

Although the partnership of social insurance with public health that marked American interest in health insurance from the beginnings persisted into the 1940s, the concern for health insurance gradually grew more fully allied with social insurance and its advocacy became associated more with the founders, architects, and administrators of the Social Security system than with experts in public health. In envisioning health policy for the postwar years, the Public Health Service developed proposals for federal support of medical education and research as well as planned hospital construction and expansion of personal health services under public-health auspices. Some features of this program, albeit in forms that accommodated medical and other interests, did emerge in the postwar years, but as a potential site for public provision, especially for the poor, public-health institutions gained little support. At the same time, figures from public health grew less active in pressing for either direct public provision of services or insurance. Meanwhile, a mixed public–private system grew dominant, consisting of nonprofit Blue plans, their for-profit competitors, and the two large governmental programs, Medicare and Medicaid, created by legislation of 1965.

These two public programs have worked under different administrative arrangements and operated in different policy environments. Administratively, Medicare lay under the Social Security Administration until 1977, when President Carter's incoming Secretary of the then Department of Health, Education, and Welfare (HEW), Joseph A. Califano, Jr. (1931–), moved it, together with Medicaid, into his newly created Health Care Financing Administration (HCFA; becoming in 2001 the Center for Medicare and Medicaid Services). Medicaid had been lodged in the welfare bureaucracy of HEW, where it had its own bureau, something it lost at HCFA, where it was overshadowed, morally and substantively, by Medicare. These administrative changes reflected Califano's goal of gaining administrative control over health and other programs in HEW and preparing the ground for NHI. Indeed, champions of Medicaid had generally aimed to sever its links with welfare and worked to render it a suitable vehicle for NHI by reducing state-by-state variations in the program and imposing broad standards of eligibility, benefits, funding levels, and accountability. However, whenever NHI was on the table, Medicaid received little attention, seen either as a thing to be dismantled under NHI or absorbed into it. Medicaid thus became no foundation for NHI but a large, diverse, and complex program for certain uninsurable people, for several categories of the poor, for the frail elderly, and for some of the disabled. Its opponents, however, tried to undermine its character as an entitlement and pressed for devolution of its administration and management to the states. Under the ACA, Medicaid is to serve as one element not in a broad system of NHI but as one enhanced and streamlined element of the larger health system, affording coverage to most of the poor, whereas other elements, public (especially Medicare) and private, continue to cover other groups. The ACA thus crowns an incremental strategy that preserves and reforms diverse preexisting forms of health insurance and financing, thereby perpetuating the difference between the poor and those with private coverage or social insurance.

Reformers saw Medicare and Medicaid as only a way-station on the route to NHI and, at the end of the 1960s, they renewed their push, hoping to cover those still uninsured (then approximately 10% of the population) and improving what often appeared to be inadequate benefits. Expansionary thinking persisted, though the gradually dawning implications of Medicare and Medicaid, which lodged large and rapidly escalating costs in the public purse, inspired the idea that NHI would provide the levers to rationalize the health system and rein in costs. Reform of the health system, dependent on governmental supervision and regulatory measures, would render affordable the expansion of entitlement. At least as late as the mid-1970s, passage of NHI, understood in this spirit, looked imminent; its failure, however, and the recession of 1974–75, which ended the long, postwar economic expansion that had fueled diverse public programs, now gave wider scope to novel ideas about health care policy. The prevailing consensus that had sustained standard modes of organizing and financing care, via fee-for-service payment of largely free-standing hospitals and solo or very small group physician practices, began to break down. So, too, did the conviction that NHI would have to take the form of a single system, governmentally mandated, planned, and regulated. New, and often conservative, voices had begun to suggest that market-based approaches to care could offer public policies that were efficient and accountable, and liberals pressing reform began to heed this advice, while persisting in emphasizing redistributive concerns, social equity, and (for some time) planning the organization of care.

Thus in the 1970s, began Starr's third era, that of '(Cost-) Containment Health Insurance,' whereas at the same time, policymakers and the researchers they financed, employed, or consulted felt the pull and fostered the growth of Fox's economizing model. Concern with expansion of entitlement persisted but in a manner that could foster rationalization of the health system and rein in cost escalation. As Starr remarks, pressure for and resistance to NHI had become competing versions of 'comprehensive reform' of the health system. From the conservative side, comprehensive reform revolved around diminishing traditional regulatory and other barriers to the functioning of markets in health care, application of stricter antitrust enforcement (especially to rein in the anticompetitive powers of the medical profession), and support for novel organizational arrangements such as 'health maintenance organizations' (HMOs). On the liberal side, comprehensive reform still meant universal coverage but, as the actions of Senator Edward M. Kennedy (1932–2009) increasingly revealed, involved a willingness to abandon demands for a single public system (like Medicare), to incorporate private innovation in the organization and supply of health services, and to exploit the power of competition to foster efficiency. In the hope of devising system-oriented reforms, economists and other researchers began to focus on market-oriented precedents and innovations.

Two new groups of reformers emerged, the one consisting chiefly economists like Mark V. Pauly (1941–), Martin Feldstein (1939–), and Joseph P. Newhouse (1942–); and the other, comprising a diverse group of professionals, including Paul M. Ellwood, Jr. (1926–, a physician with background in rehabilitation medicine); Ellwood's associate, Walter McClure

(1937–, who came to health policy from physics); Alain C. Enthoven (1930–, an economist with background, inter alia, in defense policy); and Clark C. Havighurst (1933–, a professor of law deeply interested in antitrust). Both groups hoped to exploit the persistent interest in improving entitlement to foster a more frankly market-based system. The former group aimed to create supply-side measures that would enable consumer choice in a market setting, relying on consumer sovereignty at the point of service to discipline the supply side and using income-graduated subsidies to bring the poor into the market; the latter, while exploiting similar thinking, also believed that the problems of health care could be remedied only by transforming the supply side of the market through HMOs to apply incentives directly to physicians and competition at the point of enrollment and prospective payment by capitation to encourage efficient practice. These newer approaches to public policy, although initially exotic-seeming to policymakers and to most earlier experts, gradually grew familiar, and market-based health care, as analyzed and explored by economists such as these and those receptive to their influence, became the dominant mode of thinking about health policy. The very intellectual foundations for thinking about public policy had been transformed.

Stone's classification also exploits historical analysis but it takes up a different set of the social and economic functions of insurance from those Starr emphasized. Her central question is how one should regard medical care: as something to which citizens have a right or as merely another commodity available to consumers through markets. This bifurcation has manifested itself between the divergent appeal of equity as understood in the commercial insurance industry ('actuarial fairness' being Stone's term for risk-rating of insurance) and equity as understood among advocates of social conflict and collective welfare as providing for need medically defined (Stone's 'solidarity principle'). Actuarial fairness operates by fragmenting communities into ever narrower risk groups, by emphasizing the differences among groups and by fostering the perception that individuals are responsible primarily for themselves and far less for others. Taken to its logical conclusion, actuarial fairness could shrink the risk group to the individual level, ending the mutual aid provided by insurance. Overall, actuarial fairness distributes care in inverse relation to need (however conceived), and it undermines among citizens a sense of participation in community and a conviction that community members possess mutual obligations.

In this analysis, the solidarity principle acts in the opposite direction, by broadening risk pools, by emphasizing shared traits among members of groups and members' reciprocal responsibilities and by assuring that the healthy subsidize the sick. The solidarity principle thus preserves mutual aid through the mechanism of social insurance. The historical dimension of Stone's study lies in its recounting the emergence, development, and deployment within the life-insurance industry of underwriting as a means to reduce subsidies across risk classes; the entry of underwriting into commercial health-insurance markets; and the appearance of its diverse forms of exploitation in health-insurance mechanisms. The study also points to developments current as of when she wrote that had conspired to expose to scrutiny the propriety of actuarial fairness; however, Stone finds actuarial fairness so deeply rooted in American culture that she ends her discussion on a pessimistic note about the prospects for health reform, than becoming a reinvigorated topic. However, recent reforms that aimed to expand entitlement indeed have entailed limits on underwriting. The ACA rests on the principle that price, efficiency, and generally value for money should be the focus of competition among insurers rather than characteristics of individuals, such as their preexisting conditions and health status.

If Stone's analysis emphasizes subsidies across risk groups, so that the healthy subsidize the sick, Starr's emphasizes a different social function of insurance, subsidy across income classes, so that the rich (or the better-off) subsidize the poor (or less well-off). Because in both cases financial obstacles loom large (in the latter because the poor lack ability to pay and in the former because serious illness can entail major economic loss), discussion of 'ability-to-pay' can obscure the distinction between the two kinds of subsidies. The earlier history of health insurance in America separated them fairly clearly, later they grew blurred, but in the ACA they have again become more distinct. The early Blue Cross plans, which emerged in the 1930s to provide the working and middle classes with hospital insurance, usually as a fringe benefit of employment, rested on community rating; i.e., they charged the same premium to subscribers regardless of risk class. The healthy subsidized the ill, but the extent of redistribution was modest, given that most subscribers, being of working age and employed, were largely healthy. The appearance of competing commercial insurers, which exploited experience rating, forced the Blue Cross plans to constrain or abandon community rating, thus squeezing out the subsidy across risk classes. The rise and development of managed care, especially in the 1990s, reinstated the subsidy across risk classes, in that managed care plans promised comprehensive benefits on a capitated basis to all members of an insured employment group for the same premium. However, the 'managed-care backlash' of the late-1990s, which rested in great measure on the perception that the utilization controls exerted by managed care organizations were a back-door way to renege on the commitment to provide comprehensive benefits with low copayments, led employers and insurers to back off from utilization controls and employ a diversity of more or less flexible, networked products to cater to the wishes of both employers and employees. One result was new constraints on the subsidy across risk classes.

In the case of the American Medicare Program, the basic program, Part A, hospitalization insurance for the elderly (who had not participated in the Blue plans) took wing as a way to provide the elderly a governmentally financed version of Blue Cross. However, in this case, the boundary between the two kinds of subsidy grew blurred. In part, the elderly, having left the work force, lacked income to pay for insurance; the program therefore subsidized those who were less well-off financially. However, it was the actuarial practices of commercial insurance, the exclusion of the elderly from the community rating offered by Blue Cross and the eventual departure of Blue Cross from community rating that had in effect turned a risk class – the elderly are sicker and, with purchasing power, would use more care – into an income class. By fragmenting risk pools, private underwriting made

health insurance and thus care unaffordable to many of the elderly. Similarly, any groups facing high prices because of high risk or no prices because underwriters had labeled them 'uninsurable' could not afford (or perhaps even find a venue in which to consider the possibility of affording) to pay. A risk group becomes an income group needing a subsidy. The Medicaid program, for the poor, primarily subsidizes an income group, but to the extent that its beneficiaries have lower health status than the rest of the population and thus constitute a risk group, the program subsidizes across risk groups, i.e., healthy (and better-off) taxpayers subsidize care for the unhealthy poor; the same effect can be seen among the low-income elderly on Medicare. The diversity of programs in other advanced countries also exhibit many such complexities in the nature of the subsidies that social insurance provides.

In the US, convolutions of this kind have made for difficulty in maintaining the political stability of public programs. Neither its supporters nor its opponents thought of Medicare as an end point or irrevocable commitment in social policy; rather its opponents have continued to criticize it and attempted reforms that would reduce its costs, its economic prominence, and its character as an entitlement, whereas its defenders have seen it as an expression of social-insurance principles that they have sought to extend to the entire population. However, persistent lack of a coherent rationale for the Medicare program, whether in the failure to tailor its benefits to its target populations or to provide cogent justification for it as an element of social policy, has made it possible for diverse interpretations to come to bear on it that continue to fuel debates about its future and its reform, particularly as its costs have continued to grow. Although its proponents have seen it as a partial realization of a right to care, some analysts have argued that a different sort of stability is what had anchored health care entitlements in America: programmatic rights. Controversial programs have often found stability less in a clear rationale in social policy than in the persistence of existing programs on the ground; in their support by activist courts, congressional entrepreneurs, and activists who looked to the federal government (and not the states) for leadership in social policy; and in the expectations accumulating since the New Deal among beneficiaries (current and future) that government bears responsibility for alleviating social problems. Sometimes controversial programs like Medicare thus became invested with 'programmatic rights' that stabilized their politics. Medicare may indeed have become cloaked in such rights, particularly insofar as it had been sold by its founders as a form of insurance for which beneficiaries, while in the labor force, paid through payroll deduction. However, in the current policy environment – characterized by the high cost of governmental programs and large, governmental deficits – programmatic rights seem unlikely to sustain support for these two large public health care insurance programs. If advocates are to preserve them, clear articulation of rationale and reforms in financing and may become essential.

A similar analysis clarifies the ACA. It offers both kinds of subsidy that Starr and Stone discuss: across risk groups (and hence the importance of risk adjustment under its provisions; and across income classes (as embodied in its reforms of Medicaid and in the construction of state insurance exchanges for subsidized purchase of insurance by those not covered under employment-related or public-insurance programs). The two functions of social insurance have thus become more evident under the ACA (although persistent fragmentation of risk pools still keeps them less than fully distinct). The public, however, little understands the provisions of the act. Although the gradual implementation of its provisions would likely clarify its meaning and elicit support from its beneficiaries, its political viability, in view of the controversy surrounding it, seems dependent on the success of its advocates in articulating for it a clear rationale; in tuning its provisions to suit its target populations; and in assuring a worried public still focused on programmatic rights and confused about assaults on the legitimacy of entitlements that hitherto favored programs will not erode; and in parrying claims that budgetary imperatives must entail transformation, as opposed to reform, of costly public programs. Because many cost-control measures employed in other advanced countries have thus far proven politically unacceptable in the US, advocates of public programs have struggled to find means to rein in costs while upholding the legitimacy of continued, high levels of spending in public programs of health insurance.

Social Politics and Social Science: Securing Refuge from the Market

Analysis of health insurance began in the context of thought about social politics. From the late-nineteenth century through the end of the New Deal, American analysts of social problems participated in a largely North-Atlantic culture of social politics, in which shared conceptions of social vulnerability to the transformations wrought by industrial capitalism inspired a cluster of convictions about social policy. Thus, industrializing nations needed broadly similar policies, less to achieve specific, shared goals or a common form of polity (e.g., a welfare state or a social-insurance state) than to shelter some features of social and communal life from the reign of the market. There was also a sense that some countries had moved farther or faster in that direction than other, lagging ones (especially America) and an expectation that experiences in one country could be studied for their utility to others and perhaps imported with modifications. In America, reformers felt the appeal of European experience and hoped to import foreign ideas and modify them to suit American conditions. To analyze both European experience and American possibilities, many reformers aspired to exploit the then nascent social sciences. Some possessed either formal training in the social sciences or, in their capacities as journalists, social critics, rationalizers of business and intellectual brokers, substantive knowledge of them. A major element in the emergence of the social sciences was the tension between the participation of social scientists in reform and advocacy on the one hand and, on the other, their exercise of dispassionate scientific objectivity to gain fundamental scientific knowledge, i.e., the tension between Fox's reform and relativism. Those early health reformers who came from the ranks of social scientists and from public health clearly understood themselves as exploiting their

scientific knowledge in the service of social reform. Although their reformism eventually moderated and narrowed, the change was gradual and never complete. Only beginning in the late-1950s and especially in the 1960s, did analysts harness formal and recognizably modern economic analysis to health policy, and in that context as well, normative considerations, while circumscribed, have marked even the most ostensibly positive analyses.

Thus Starr's Progressive Health Insurance had much in common with later thinking about health insurance, but it articulated more explicitly than later proposals the rationale for distributive justice. Capitalist development, as reformers saw it, having imposed most of its costs but few of its benefits on labor, left workers facing primarily four risks, unemployment, accident, illness and old age, all of which portended the impoverishment and immiseration of workers and their families. To remedy the problems resultant from the realization of these risks, reformers recommended social insurance and, specifically in the case of health care, they pressed for 'sickness insurance' primarily to cover its indirect costs, especially loss of income. They understood that such measures would require political support and exerted themselves in various ways to achieve it. Reformers like Isaac Max Rubinow (1875–1936) aimed to enroll fellow reformers into a coalition, to which they hoped to recruit leaders of the major interests (business, labor, the medical profession). A reform tradition descending from John R. Commons (1862–1945) at the University of Wisconsin hoped to create support by showing that the workers, industry, and the public possessed shared interests in workers' well-being. Reformers aimed, in a word, to create a broad sense of social solidarity that would undergird reform coalitions. However, these reformers failed to parry opposition from diverse, well-organized interests, and, in the Progressive Era, their efforts came to naught.

In Starr's expansionary era, however, advocates again pressed for health insurance, this time emphasizing the direct costs of medical care and the social costs resulting from deficiencies in its accessibility and limitations on its availability. In doing so, however, they rarely let notions of social justice take center stage. Instead, reformers and advocates emphasized two things: the efficacy of care and the peculiar economic features of health care and the health sector. With regard to the first, reformers became deeply impressed with the advances in medical science during the late-nineteenth and early-twentieth centuries and convinced that care was of tremendous value. They therefore articulated the notion of need, urged from the outset of the expansionary era in the work of the CCMC. The committee invoked insurance not only as a mechanism to enhance access to needed services, but, out of the conviction that the health sector was inadequately developed to meet the needs of even those who could afford care, also as a method to finance the expansion of health resources (hospitals, clinics, technology, and trained personnel).

Not only efficacy suggested the importance of care but also the apparent implications emergent from early economic analysis of health care and the health sector. Analysts repeatedly identified and characterized the poor fit of health care, as opposed to most conventional commodities, with the standard tools and procedures of economic analysis, and these economic peculiarities seemed, in advocates' minds, to reflect the special moral and social significance of health and health care. Thus, analysts showed that health care differed from other commodities in several economically significant ways – in modern terms, that the demand for health care is derived from the demand for health; that health care exhibits externalities (costs or benefits involving parties outside of a transaction); that providers and patients-qua-consumers exhibit informational asymmetries (i.e., consumers are ignorant of what had recently become a recondite and technical field of scientific medicine inaccessible to those without long and arduous training); and that patients experience uncertainty regarding both the need for and effectiveness of care. In its simultaneous possession of these economically distinguishing characteristics, health care, in the eyes of reformers, was very nearly unique. In the light of these peculiarities, society had limited the extent to which market principles applied to health care, for example, through professional self-regulation, nonprofit organization of hospitals, support for programs to enlarge the health sector and to facilitate access to it; and charitable and philanthropic arrangements that served both poor and the middle class. As seen by advocates of insurance, the economic peculiarities of care, precisely because of its often little-articulated moral significance, had given rise to social arrangements that replaced standard market arrangements and thereby expressed underlying commitments to social justice.

However, corrosive forces were at work. These elicited more explicit articulation of noneconomic rationales for distributing care equitably. From the early-1970s and lasting in significant measure to the present, some voices, concerned about costs and mindful of the lack of knowledge about the effectiveness of care, expressed skepticism about the value of especially high-technology care, at least at the margin. Notions of need, that is, having begun to grow intellectually exiguous, newer analysts such as Mark Pauly began to suggest that consumers, as opposed to experts, should be allowed to exercise choice in a relatively unfettered market. In response to such growing uncertainty about the value of health care and its implications that markets need not be constrained, some proponents of redistributive policies found an additional rationale for the nonmarket arrangements prevalent in the health sector – they directly express the existence and value of social cohesiveness, of inclusive sentiments about the poor and the sick, of a will to maintain and preserve the dignity of all citizens, and of a tendency to evaluate positively lives that are not conventionally economically productive (children, the elderly, and the disabled). Figures holding these views sometimes accorded the intangible features expressed in redistributive measures in health care a priority that equaled or exceeded that of the substantive economic benefits (reduction of individual and social costs) that access to care could bring. More recent analysts, responding to the eclipse of distributional rationales for public programs under pressure of market-based health policy, have taken a similar approach, to exhibit and therefore justify perpetuating the solidarity foundations that public programs seemed to them to possess even beyond the value of the concrete medical benefits they confer.

The Eroding Aura of Medicine and the Opening to Market-Based Thinking

Cultural developments, emergent or newly prominent after World War II, exerted corrosive effects on the notion, long animating reformers, that health care and its providers possessed special qualities. Paradoxically, the organized medical profession itself was one agent of this change: while defending itself against governmental intrusion into medical care aimed at advancing entitlement to coverage, the profession portrayed the purchase of medical care as just another consumption decision, one often overshadowed by consumers' preferences for other goods and services. Lack of ability to pay seemed beside the point; supposedly unmet need, from this perspective, should be regarded not as a reflection of deficient public policy but as an anticipated outcome of a consumer society in which demand (not need) dictated the distribution of care. Health care, as another commodity, belonged not in the purview of redistributive policies, but in that of the market, where consumers could take of it what they wanted. Of course, for physicians the market was the one they had helped create and preserve, but upholding it in the face of consumerism would prove more and more difficult, for if the services that physicians purveyed were not so special, neither were the purveyors. Factors that diminished the personal ties in physician–patient relationships and substituted a remote professionalism led patients to take a more dispassionate view of their doctors. An increasingly well-insured suburban middle class viewed medical care as it did other, especially professional services, that is, as routinely available for purchase and subject to scrutiny with a consumer's eye. Social scientists, moreover, had revealed with some surprise that the high-minded professionalism of medicine seemed to cover professionally self-interested behavior. Culturally, health care formed part of the broader changes in the culture of consumption and individualism that gave precedence to the market ahead of government and politics and that gave priority to free choice over paternalism and sentiments of social solidarity and inclusiveness. Consumers increasingly expected to make market choices for services that reflected their own sense of what they wanted and needed.

This was the state of affairs that emerged in the beginning of the 1970s and set the stage for the appearance of market-based health policy: traditional reformers pressed for a governmental program of NHI in the light of their conceptions of solidarity and social justice; cost escalation, particularly under Medicare and Medicaid, suggested the need for systemic reform; medicine and its practitioners itself suffered loss of prestige; some newer voices began to doubt prevailing notions of need, thought of care as a commodity, and claimed that health care should be allowed to operate in the market; and traditional advocates of NHI responded by emphasizing that broadened entitlement to care can express and foster solidarity and social justice. Meanwhile, the social sciences, especially economics, began to suggest novel policy ideas that, their practitioners held, could accomplish system reform and redistributive goals better than further application of prevailing policy methods. The next article takes up the immediate cultural and intellectual developments that gave scope to market-based notions of health policy, it pursues the intellectual history of market-oriented health care, and it suggests how the evolution of markets have both reflected and affected novel policy positions.

See also: Efficiency and Equity in Health: Philosophical Considerations. Health and Health Care, Need for. Health Care Demand, Empirical Determinants of. Health Insurance and Health. Health Insurance in Developed Countries, History of. Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare. Health Insurance in the United States, History of. Health Insurance Systems in Developed Countries, Comparisons of. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Managed Care. Measuring Equality and Equity in Health and Health Care. Moral Hazard. Risk Adjustment as Mechanism Design. Risk Classification and Health Insurance. Risk Equalization and Risk Adjustment, the European Perspective. Risk Selection and Risk Adjustment. Social Health Insurance – Theory and Evidence

Further Reading

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5), 941–973.

Fox, D. M. (1979). From reform to relativism: A history of economists and health care. *Milbank Memorial Fund Quarterly/Health and Society* **57**(3), 297–336.

Fox, D. M. (1990). Health policy and the politics of research in the United States. *Journal of Health Politics, Policy, and Law* **15**(3), 481–499.

Institute of Medicine (IOM). Committee on the Consequences of Uninsurance (2002). *Care without coverage: Too little, too late*. Washington, DC: National Academy Press.

Institute of Medicine (IOM). Committee on the Consequences of Uninsurance (2003). *Hidden costs, value lost: Uninsurance in America*. Washington, DC: National Academy Press.

Marmor, T. R. (2000). *The politics of medicare*, 2nd ed. Hawthorne, NY: Aldine de Gruyter.

Melhado, E. M. (1988). Competition vs. regulation in American health policy. In Melhado, E. M., Feinberg, W. and Swartz, H. M. (eds.) *Money, power, and health care*, pp 15–101. Ann Arbor: Health Administration Press.

Melnick, R. S. (1996). Federalism and the new rights. *Yale Law and Policy Review* **14**(symposium issue), 325–354.

Robinson, J. C. (2004a). From managed care to consumer health insurance: The fall and rise of Aetna. *Health Affairs* **23**(2), 43–55.

Robinson, J. C. (2004b). Reinvention of health insurance in the consumer era. *Journal of the American Medical Association* **291**(15), 1880–1886.

Rodgers, D. T. (1998). *Atlantic crossings: Social politics in a progressive age*. Cambridge, MA and London: Belknap Press of Harvard University Press.

Smith, D. G. and Moore, J. D. (2008). *Medicaid politics and policy, 1965–2007*. New Brunswick, NJ and London, UK: Transaction Publishers.

Starr, P. (1982). Transformation in defeat: The changing objectives of national health insurance, 1915–1980. *American Journal of Public Health* **72**(1), 78–88.

Starr, P. (2011). *Remedy and reaction: The peculiar American struggle over health care reform*. New Haven and London: Yale University Press.

Stone, D. A. (1993). The struggle for the soul of health insurance. *Journal of Health Politics, Policy, and Law* **18**(2), 286–317.

Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare

EM Melhado, University of Illinois at Urbana–Champaign, Urbana, IL, USA

Introduction

Health Insurance in Historical Perspective, Part I explored several frameworks for understanding the evolution of American thought about health insurance; examined the belief of traditional reformers that health insurance should serve as one of a cluster of measures designed to secure citizens from the risks posed by capitalistic markets; suggested that, in an environment of escalating healthcare costs, doubts about the value of healthcare had led some reformers to stress its significance less for its substantive benefits than for its utility as an expression of social solidarity; noted the factors that undermined the special status of medicine and medical care; and indicated that medical care, in the eyes of diverse analysts, increasingly resembled other commodities traded in conventional markets. This article opens by characterizing the two broad forms that American proposals for market-based health policy initially assumed: one resting on modern economic analysis of the demand side of healthcare markets and the other, initially depending far less heavily on formal economic analysis, but reflecting the conviction that public purposes could better be realized through supply-side reforms. The article reveals the extent to which some of the founding ideas and concerns of health economics arose through analysis of the health sector when national health insurance (NHI) seemed imminent; and it briefly explores the consequences of these developments for the history of ideas about health insurance and for the development of healthcare markets on the ground. It offers conclusions about both the kinds of reform measures that American health policy has generated and role of economists in health policy.

From Advocating Care to Reforming Health Insurance

More than changing cultural perceptions of medicine helped elicit market-based thinking in health policy. From the late 1960s, and especially in the light of cost escalation that followed the introduction of Medicare and Medicaid, American healthcare became the subject of scrutiny that began to reveal shortcomings that would have to be remedied under any system of NHI. Cost escalation, the most pressing, was in a sense only a symptom of increasingly nonfunctional features of American healthcare. The health sector appeared to be an uncoordinated profusion of chiefly solo or small-group physician practices; freestanding, independent hospitals; and a diversity of public and private insurance programs. The fragmentation of the health sector, its maldistribution of resources, and its inability to tailor resources to needs on a community level or services to individuals in and among local institutions constituted a set of problems that experts as well as the public hoped to remedy. Pressure for NHI, that is, had become transformed into pressure for broad reform of the health system.

Until the end of the 1970s, NHI had seemed imminent, although in retrospect the apparently close but still abortive effort to achieve it in 1974 brought its short-term prospects to an end. For many traditional reformers, planning and regulation that they expected to take root under NHI would provide the levers to rationalize the distribution and deployment of resources and rein in costs; the resultant efficiency gains would provide the resources to expand and improve entitlements to health services. In their view, system reform amounted to more extensive and more thorough-going application of traditional policy means. However, other analysts of cost escalation and fragmentation exploited the prevailing interest in NHI as a vehicle to introduce novel ideas about the utility of markets and competition to solve the problems of healthcare. Two clusters of proposals emerged from their efforts: Reform of the demand side of the market through the imposition of increased cost-sharing under insurance (that is, increased out-of-pocket expenditures for insured individuals and families), combined with subsidies, graduated inversely with income, to insure the poor; and reform of the supply side of the market by creation of health maintenance organizations (HMOs) or other health plans that combined the delivery of healthcare and the insurance mechanisms to finance it. The roles of economics (and some other social sciences) in early studies of health insurance can be examined by tracing the emergence of these two categories of proposals.

Income-Graduated Cost-Sharing

Within economics it was the application of formal doctrines that increasingly subsumed healthcare under the rubric 'commodity.' The implications of the change emerged in at least two stages, one in which traditional advocates of NHI began to apply to healthcare (among other areas of policy) formal rationales for governmental provision and a second, in which skepticism about governmental provision combined with economic analyses to undermine the case for the specialness of care and therefore suggest the propriety of its subordination to market arrangements.

The first stage is represented by the tradition of public-expenditure analysis, which emerged in the 1960s as part of the effort to rationalize governmental financing or provision of public services. Whereas an economist like Seymour E. Harris (1897–1974), in his study of American medicine, exemplified traditional advocacy for increasing the quantity, improving the quality, and rationalizing the distribution of health services, Herbert Klarman (1917–99), in a major, early review of health economics, maintained that only those health programs that made better use of resources than alternative ones could find economic justification. However,

 doi:10.1016/B978-0-12-375678-7.00926-3

this orientation did not deflect needs-based analyses. Animated by concern for the social costs of lack of care, analysts regarded care as an investment in human capital and exploited the cost-benefit principles previously developed in analyzing governmental investments in water projects. Needs-based thinking had supposed that the demand for care depended all but exclusively on epidemiological, scientific, and technological factors; but a more dispassionate economic analysis provided evidence that care resembles other commodities in that its demand also depended on the economic variables of income and price (i.e., demand for care exhibited income and price elasticities). Nevertheless, from a needs-based perspective, such evidence could be reinterpreted: to recognize that insurance (a price subsidy) improves access to care is less to acknowledge the price elasticity of demand than to welcome the shift brought by insurance of a deprived population into the ranks of those able to acquire one of the necessities of life. Similarly, income effects among the insured wealthy need not have been taken to imply the dependence of demand on price. The wealthy buy more services because they have more education and appreciate more the value of care. From a needs-based standpoint, evidence for the commodity-like behavior of care therefore carried little weight, and it authorized reliance not on novel markets but on planning. Indeed, one of the economically characteristic features of healthcare, informational asymmetries, and the consequent dependence of patients on experts, only reinforced the conviction that nonmarket arrangements were preferred, if not indeed necessary.

Although cultural changes noted in the previous article helped divest care of its special characteristics, developments within the social sciences fostered a reorientation among formal analysts of public policy. A novel and powerful approach to analyzing both politics and policy, known as 'public choice,' particularly as undertaken by one of its founders and leading lights, James M. Buchanan (1919–2013), suggested that the virtues of public provision had been overrated. In a study of public goods, Buchanan revised the case for special social arrangements, especially public provision and production of certain goods. Acknowledging the desire of some citizens to increase the consumption of particular goods by all citizens, he could treat the individual's consumption as enhanced by an external benefit. His analysis suggested, moreover, that unlike cases such as national defense or fire and police protection, such 'externalities in consumption' need imply no monolithic supply, for example, governmental provision. Externalities in consumption could be provided in conventional markets by private producers so long as the community participates (through financing) in purchasing the goods or services. Buchanan's position departed from that represented by Paul Samuelson (1915–2009), one of the major analysts of public goods, whose approach Buchanan regarded as excessively prescriptive (i.e., paternalistic). Buchanan thus provided a path for analyzing healthcare that opened the door to subsuming it under more conventional market arrangements.

It was Buchanan's student, Mark V. Pauly (1941–), eventually to become one of the most distinguished of American health economists, who first took that path (1971), although Martin S. Feldstein (1939–), having undertaken an econometric analysis of the British National Health Service, was working along similar lines in the early 1970s; indeed, the two exerted a mutual influence. Although some others had in principle reduced calls for NHI to externalities in consumption, it was Pauly who first unequivocally translated notions about the specialness of care into support for a tax-financed program of subsidies. Pauly's question was how to optimize the subsidy. His analysis took the unequal distribution of income as given, assumed that demand for care responds to price and income – i.e., he accepted frankly that price and income elasticities suggested that care is an ordinary commodity – and anticipated that different consumers would have different levels of care. This last point also departed from traditional social justice rationales for care, which largely anticipated that NHI would provide a uniform standard of care. His analysis led him to propose 'variable subsidy insurance' (VSI). For the poorest it would prove comprehensive coverage at low or zero premium cost; for those with middle incomes, it would subsidize demand by paying part of the premium cost (perhaps to an extent that varies inversely with income) and impose deductibles and/or coinsurance that would increase with income; for the wealthy, it would supply a catastrophic policy, i.e., one that would pay only for the most expensive forms of care. The cost-sharing provisions would constrain utilization (and thus respond to growing concerns, intensified by talk of NHI, about cost escalation). Almost simultaneously, Feldstein offered a similar proposal for 'major risk insurance' (MRI).

These proposals carried important implications both for improving public policy and for exhibiting the value of economics – and some implications of its use – as a means for analyzing policy. In regard primarily to the substance of policy, several features stand out. In acknowledging the desire of some citizens to increase consumption of care by others, the proposals gave expression to social solidarity. They do so, moreover, by assuring taxpayer sovereignty: the taxpayers decide what services to subsidize, for whom, and to what extent. In recognizing that diverse consumers (because of differences in their 'tastes' for care and in their income) would exhibit diverse levels of demand for care and in according a minimal role to expert determination of need, the proposals expressed consumer sovereignty. In granting the poor, as traditional reformers had wanted, the same rights as those better-off to make choices from among the same providers in the same private markets, the proposals emphasized that aspect of social solidarity that focused on inclusiveness and mutual regard across income classes. However, in rejecting a universal standard of care, the proposals drew back from the distributive imperatives underlying older notions of solidarity. This result followed in part from the economic tools that underlay the proposals. The optimization procedures of welfare economics aimed to enhance allocative efficiency – the efficiency with which resources are distributed among consumers - in that a system of graduated subsidies under cost-sharing would achieve a reasonably tight match between the income of consumers and the socially desired enhancement to their consumption of care. In addition, any market operating under these proposals would help constrain social costs by fostering productive efficiency: In competing for the business of patients with purchasing power, healthcare providers would have to show themselves frugal in using the funds brought to

healthcare transactions by insured consumers who would have to foot a significant part of the bill. Providers would seek either to produce a given level of care more efficiently or offer services of perhaps reduced (but still positive) benefit but at lower cost. Physicians and hospitals, that is, would have to become the financial, as well as the medical fiduciaries of their patients. Finally, because the proposals left market mechanisms largely undisturbed (except to the extent the supply side would evolve on its own under such a system of subsidies and cost-sharing), they offered a means to resolve controversy among economists about separating efficiency in the allocation of health services (achieved through the exercise of consumer choice under cost-sharing) from distributional equity in access to health services (achieved through income-graduated public subsidies).

Two additional features of the work by Pauly and Feldstein merit attention. One is their discussion of 'moral hazard,' the tendency of insurance itself to foster the occurrence of the risks against which it provides protection, so named by the insurance industry to signal the 'abuse' of insurance by policy holders. In the case of healthcare, under an insurance scheme that, absent cost-sharing, affords a zero price at the point of service, the insured will purchase more care than otherwise. Pauly regarded the effect not as morally dubious but as rational. It implies that a taxation scheme that compels citizens to pay for insurance against certain risks is inefficient, because, under the scheme, some consumers would have to pay more than they would want to; some consumers, in a word, would benefit from purchasing a lower standard of care. Moreover, cost-sharing, by reducing demand and thus constraining utilization, would reduce the premium of insurance and therefore could make desirable a policy otherwise unattractive to some consumers. The effect of coinsurance, for example, depends on the elasticity of demand, which varies among consumers; an optimal policy would thus similarly vary. Hence, the utility of such schemes as VSI: Income-graduate subsidy would encourage socially desired utilization (i.e., the increased consumption by some consumers that others desired); and income-graduated cost-sharing would improve the efficiency of the resultant allocation.

Feldstein, responding to Pauly's view of moral hazard, showed in the case of the hospital industry that the stimulation of demand that insurance occasions has a special characteristic: it results in increased prices, which in turn elicit more insurance, i.e., it produces a circular effect that, although not explosive, provided strong evidence in support of cost-sharing. Moreover, the government further stimulates demand via the tax treatment of health insurance (primarily, that the health benefits that employers provide employees are exempt from employees' income tax), from which Feldstein drew two conclusions: (1) tax subsidies make the net cost of an insurance premium fall below the expected value of the benefits; and (2) they encourage employees to substitute for taxable wages more comprehensive (but shallow) insurance. Insurance then provides first-dollar coverage for modest expenses, but little coverage for catastrophic ones. It was in the light of these concerns that Feldstein devised MRI. Health insurance, previously seen as a solution to the problem of achieving access to health services, itself now became the source of two problems: intensive price inflation and inappropriate forms of coverage. Older advocates of health insurance had insisted on universal, comprehensive benefits as following from the high social valuation of healthcare; now, their approach seemed to be an artifact of faulty policies. In the newer view, allowing some consumers to purchase a lower standard of care would not only serve the cause of efficiency but it might also help overcome the political obstacles to NHI. As Pauly observed after over a decade of discussion about the virtues of NHI and of its possible forms, advocates of comprehensive NHI had kept the poor and those suffering from catastrophic illness from obtaining a standard of care that, if lower, was nevertheless, for them, more desirable.

These concerns lay in the background to the RAND health insurance experiment (HIE), one of the most ambitious social experiments ever undertaken. Conducted over the period from about 1974 to about 1982 by the RAND Corporation, a nonprofit organization that contracts with diverse organizations to carry out research and policy analysis, the experiment emerged from the War on Poverty amid discussions of how to arrange financing of care for the poor. The principal issue around which it took form was the lack of consensus about the effects of increased demand (through expanded entitlement and improvement of benefits) and about the effects of cost-sharing, on both utilization and health, in constraining demand. This is not the place to discuss either its origins and evolution or underlying economic concepts; for present purposes, only some of its conclusions merit attention. The experiment indicated a price elasticity of -0.1 to -0.3 for most kinds of health services (i.e., an increase in price of 1% would decrease quantity demanded from between 0.1% and 0.3%). Although the measured elasticities were modest, the experiment seemed to show that consumers do adjust usage to price; that excessive insurance does seem to result from moral hazard; that cost-sharing does constrain use, even for hospitalization; that these changes, for all but the sick poor, had little effect on health; and that therefore cost-sharing can serve as a sound instrument of public policy that aims to constrain costs. The implication seems to have been that much of the care provided to most consumers lay on what Alain C. Enthoven (1930–), a prominent advocate of healthcare markets, called the 'flat-of-the-curve' (i.e., where the initially upward graph of benefits of care as a function of their costs flattens, indicating that additional expenditures on care provide no health benefits).

However, in constraining use, cost-sharing did not, as its advocates hoped, limit chiefly ineffective care. Cost-sharing was therefore a blunt instrument, but its impact, at least on the nonpoor, seemed positive, for the reduction of utilization it achieved did not have an adverse effect on health. Pauly and Feldstein had justified consumer sovereignty with reference to lack of knowledge about the outcomes of care. The RAND group, which had classified forms of care into the categories of 'effective' and 'ineffective,' now argued that the failure, even of care it classified as effective, to affect health under a variety of insurance schemes that fostered reduced utilization authorized the same conclusion.

Cost-Sharing, the Poor, and the Value of Services

However, the message issuing from the experiment was not univocal. The HIE revealed that, in regard to the poor,

especially the ill poor, cost-sharing could entrain adverse effects on health; that is, the failure of the poor, under cost-sharing, to obtain some effective services led to reductions in their health status. Diverse policy responses could be devised to bring such services to the poor. One would establish targeted programs to supply specific services to the poor, although not all services are amenable to this approach. Others might exploit screening programs, but in large populations their costs exceed their benefits. Moreover, as critics of market-based care have argued, the likely confinement of this measure to public programs risks offense to standards of equity and the dignity of the poor. Yet another would supply insurance but exempt the poor from cost-sharing, as Pauly had suggested and as, under Medicaid, they largely had been, although maintaining a separate Medicaid program rather than the imposition of a general income-graduated cost-sharing would continue to stigmatize the poor, which is indeed the approach taken by the Obama reforms under the Affordable Care Act (ACA, i.e., the Patient Protection and ACA, PL 111–148; as amended by the Health Care and Education Reconciliation Act, P.L. 111–152, passed in March 2010). Another approach, also taken under the ACA, would impose on individuals and families a modest level of income-graduated cost-sharing and provide income-graduated subsidies, as the likes of Pauly, Feldstein, and the RAND group had been discussing. Yet another measure would be to structure coinsurance so as to foster coverage of effective services, an approach that draws strength from recent research on the effectiveness of care, although the still small proportion of services that have been evaluated limits the usefulness of this practice.

Other policy responses might be devised; however, more important than the prospect of modest cost-sharing in any version of NHI, the experimenters acknowledged, was the difference between some insurance and none. Nevertheless, in regard to the poor, the RAND group was reticent, leaving to policymakers to decide whether the experimental results should authorize public provision of care to the sick poor. The normative case for cost-sharing for the nonpoor, in other words, was for the researchers overwhelming; but for the poor they aimed only to narrow public debate by providing concrete experimental results, not to propose whether and if so how to expand entitlement to services. For the poor, if not for the better-off, relativism, not reform, is what characterized analysts of health policy.

The experiment has exerted an enduring influence in American health policy, particularly in its emphasis on the utility of demand-side measures – which have received far greater application in America than in other advanced countries – to constrain utilization. However, subsequent developments have changed the context for assessing its implications. A growing body of more recent research has suggested much more strongly than the HIE that uninsurance and underinsurance, especially for the poor, entrains poor health outcomes and that improving Medicaid and other kinds of coverage entails positive health benefits. By strengthening previously attenuated convictions about the effectiveness of care, these results have enhanced the case for redistribution to cover effective services, whether routine and inexpensive (such as blood pressure monitoring and in general management of chronic diseases) or less frequent but much more costly (such as organ transplants or care for heart disease or cancer), especially but not only for the sick poor.

Indeed, Nyman argues that advocates of cost-sharing have failed to understand a point that reformers have been making since early in the past century: insurance is needed to secure access to forms of care that are not affordable even by the middle class and that are medically valuable, even life-saving; insurance, that is, possesses what Nyman dubs its 'access value.' In these cases, the exercise of moral hazard, that is, the purchase of more care than would be purchased without insurance, is precisely the point, for it gives access to valuable services that would otherwise be inaccessible. Because the payoff from insurance amounts to an increment to income, Nyman argues, the purchasing decisions of a seriously ill person with insurance reflect not a shift along the demand curve, as most economists assume, but a shift of the curve outward. Discouraging consumption through cost-sharing of services that are valuable and expensive is therefore welfare reducing, because it limits the access value; at the same time, excessive consumption of less urgently needed or less valuable care may be a relatively minor effect of insurance. Pauly, a major architect of the moral-hazard argument, eventually recognized that its applicability to the seriously ill and the services they need had not been adequately studied. The HIE therefore provides little assistance for policymakers in deciding the extent to which especially expensive services should become available to Americans, both poor and better-off.

These reflections, which result from new research that occasioned reevaluation of the RAND HIE, clearly implicate both sides of the market, although the HIE itself had focused on the demand side. The figures such as Pauly and Feldstein who had suggested demand-side reforms at first actively opposed reconstruction of the supply side of the market. Reconstruction would require a major role for government, but the newer approaches to public policy took inspiration precisely from what their advocates regarded as governmental failures, especially under traditional regulatory regimes (which had been under attack since the Carter administration). By contrast, incentive-based reforms, to which Charles L. Schultze (1924–) later gave systematic articulation, sidestepped any meddlesome and likely counterproductive governmental intrusion into the economy, and it reduced the risk of antagonizing the major interests, especially the providers of healthcare. Instead of what Schultze called the 'command-and-control' characteristic of regulatory regimes and the 'perverse incentives' operating under them – terms that helped put much wind in the sails of Schultze's ideas – incentives that aligned the interests of actors with public purposes could serve public policy more efficiently in both economic and political senses. Moreover, economists believed that demand-side reforms that responded to concerns for the inflationary effects of insurance (e.g., reducing the tax subsidies of health insurance), could achieve with reasonable promptness and certainty the savings anticipated by theorists, whereas ambitious structural reforms might not work and entail severe unintended consequences. Supply-side innovations, nevertheless, had their advocates, and the evolution of markets on the ground has taken place in a context defined by their concerns and their vocabulary.

Healthcare Plans

Indeed, it was roughly simultaneously with demand-side analyses that an alternative, supply-side approach emerged. It called for combining insurance with the provision of care through competing large, bureaucratic institutions (healthcare plans, initially, chiefly the HMO). The early proponents of this approach shared some views with advocates of cost-sharing, especially that healthcare is a commodity suitable for sale to consumers in markets, and a commitment to an incentives-oriented approach to public policy as preferable to regulation; but they departed from advocates of cost-sharing by calling for government to assist in reorganizing the supply side of the market and then to withdraw and let it evolve. Moreover, unlike advocacy of cost-sharing, the call for reform of the supply side did not at first result chiefly from applications of economic theory.

Reforming the Market

Instead, their views arose from at least three convictions: (1) although under cost-sharing physicians would have to compete on economic as well as medical grounds, aiming to serve as fiduciaries of patients' money as well as their health, incentives toward economy could become truly effective only if they were made to bear more directly on physicians; (2) large bureaucratic organizations could accomplish this task in ways not possible under market conditions characterized by solo practice and freestanding hospitals; and (3) traditional healthcare policy, which relied on professional self-regulation, planning, and regulation of institutions, especially hospitals – the very features of healthcare markets that, for older theorists, distinguished them from conventional markets and reflected the unusual characteristics and fundamental importance of healthcare – if carried out effectively, would either lock-in the causes of dysfunction in healthcare or, because draconian, erode in the face of opposition from patients and providers. In a period when cost escalation elicited characterizations of complex problems facing the health sector, when calls for NHI grew coupled with calls for the reform of health system itself, and when traditional forms of governmental regulation were in decline, the market solution, based on new organizations, seemed to cut through the Gordian knot that advocates of traditional NHI then still hoped to unravel by strengthening established practices.

Hence Paul M. Ellwood, Jr. (1926–) and his colleagues, in their classic summary of the 'health maintenance strategy' Ellwood *et al.* (1971), held that the "health system is performing poorly because its structure and incentives do not encourage [systemic] self-regulation" and that "[m]arket mechanisms, such as competition and informed consumer demand, which might provide a check on the provision of unnecessary services, inflation, and inequitable distribution, do not exist in the health industry." Their conclusion (p. 298) was as simple as it was bold:

> The emergence of a free-market economy could stimulate a course of change in the health industry that would have some of the classical aspects of the industrial revolution - conversion to larger units of production, technological innovation, division of labor, substitution of capital for labor, vigorous competition, and profitability as the mandatory condition of survival. Under these conditions, HMOs would have a vested interest in regulating output, performance, and costs in the public interest, with minimal intervention by the federal government.

To sharpen the contrast between prevailing arrangements and the market-based system, Ellwood, from 1972, invoked a locution that until then had been little exploited in discussions of healthcare, 'cottage industry.' In the early 1970s, it allowed him, together with his colleagues and allies, to epitomize the inadequacy of what they perceived to be a still preindustrial health sector; and it has remained a handy resource that has enabled them and their successors to deprecate subtly traditional healthcare policy and practice, while enhancing the legitimacy of the novel, market-based ones and buoying their prospects.

The confidence evident in the preceding quotation rested far less on economic theory than on enthusiasm for a textbook notion of competition and from the knowledge that the archetype of the HMO, the prepaid group practice (PPG) – of which a few then existed, several having emerged especially from the 1930s – had successfully provided high-quality care more cheaply. Their principal tools were capitation payments (per head or per family) from plan members and either the staff or group model of provider organizations – in the former, the plan itself employs physicians, whom it pays a salary; in the latter, the plan pays the physician group, which pays its physicians a salary. In both kinds of plans the incentive structure of fee-for-service medicine had been reversed – for example, as Enthoven saw it, of Schultze's call for reform of perverse incentives – as neither plans nor physicians benefited from increased utilization. Moreover, by owning or contracting with hospitals paid on a global budget, the plans had incentives to provide hospital care efficiently. Yet the numbers and market penetration of such organizations was small and, in areas in which patients did have choice of insurers, the success of health plans may well have reflected their case mix and the tastes of their clienteles. Moreover, early analysis of their performance suggested that their economies resulted primarily from limiting hospitalization rather than from constraining the other aspects of practice that exposure to a fully competitive market might have led plans to target. In other words, as a model for a competitive health system, the HMO was suggestive but hardly compelling. To call for expanding these modest precedents to dominate the entire health system and create a novel, competitive market was thus to pose an enormous gamble (as advocates of cost-sharing under fee-for-service had believed). Proponents found it appealing because, in the face of the complex problems of the health sector, competing health plans seemed a conceptually simple approach, one as yet little encumbered with a body of experience and a long history in the policy sphere. In comparison with what market advocates saw an apparently exhausted tradition of regulation and planning, markets populated on the supply side by competing, large, capital-intensive organizations looked fresh and promising.

However, there was some pertinent history in the policy sphere. The modest degree of market penetration that bureaucratic practices had attained by the end of the 1960s

reflected in part the successes of the organized medical profession in controlling not only the narrow dimensions of medical practice and training but also the organization and financing of healthcare. PPGs had long been a target of the profession, which had generally succeeded in constraining their growth and proliferation. To assist him in taking on this legacy of professional control, Ellwood coined the term, 'HMO.' It expressed not only the hope that he, as a rehabilitation physician, entertained about the importance of prevention (especially regarding chronic disease) and its utility in an anticipated cost-control regime but also his expectation that additional organizational forms beyond the traditional PPGs could serve the purposes that advocates of plans envisioned. However, the new term recommended itself chiefly as a way to appeal to physicians without eliciting memories of the history of conflict over the organization of medical care. Like earlier reformers, who saw that social and economic developments presaged transformations of healthcare and called on the medical profession both to lead and, by so doing, protect its interests, Ellwood hoped to engage physicians and enroll them in his project of reform. However, another advocate of market-based reforms, the law professor Clark C. Havighurst (1933–), took a more adversarial stance toward the profession, holding that professional self-regulation underlay the profession's anticompetitive practices. The cottage industry was the profession's creature; it existed to serve the interests of the profession, not those of patients or polity. From the standpoint of his concern for antitrust, he believed that reorganizing the supply side would break the back of medical dominance over the market for healthcare, permit the evolution of large provider organizations that the profession had long succeeded in inhibiting, and expose physicians to market discipline. Even more important, he took on the role of policy entrepreneur who disseminated his views among those able to make decisions and act in practical circumstances. A major goal for his activity was to establish the market as a realm for the exercise of choice by consumers.

The Evolution of Healthcare Markets

From the mid-1980s, the reduction of constraints on supply-side innovation resulting from antitrust activity; the diminished threat, after the failure of the Clinton health-reform plan, of increased federal regulation; and the restraints on state regulation resulting from federal preemption, under the federal Employee Retirement Income Security Act of 1974 (P.L. 93–406), of state regulatory powers in healthcare, helped open the door to the rapid evolution of healthcare markets on the ground. A new coinage, 'managed care,' emerged in the late 1970s and became commonplace from about the mid-1980s to encompass the early emergence of diverse and novel supply-side arrangements in addition to HMOs as originally conceived. Under that term, analysts included organizations and practices that supposedly generated efficiency gains (and thus cost-controls and quality improvements) through corporate control over the practice of medicine and that supposedly fostered competition among managed care entities and between them and conventional fee-for-service practice. From the late 1990s, with the 'managed care backlash,' the apparent consensus on the virtues of managed care had dissolved, but dynamic evolution continues.

That dynamism is one of several themes that emerges from the growth of markets. In both extent and degree, the dynamism of healthcare markets has surely exceeded the expectations of most of their early advocates. An industry formerly heavily sheltered from market forces now, under the profit motive – and the resultant imperative for nonprofit entities to emulate for-profit ones – has become subject to chaotic impulses that have created, reconstructed, and destroyed novel organizations and managerial and professional practices, as well as built and upended institutions and relationships among employers, insurers, providers, and patients. Indeed, so rapid have markets evolved that scholars have been in continuous struggle to keep up with events, characterize changes, and assess their implications. Such changes arouse concern not only with the services that healthcare markets provide but also likely more so the economic advantages and the profits that issue from them. A focus on market share and profit making is surely what anyone expects of markets; but roiling market dynamics seems incompatible with the stability that patients and consumers would hope for in a system intended to provide services of an often intimate nature and existential import.

Nevertheless, the concern of market-oriented analysts and policymakers to widen the scope of consumer choice is a second theme in the evolution of markets. The managed care backlash seemed to suggest that consumers were disillusioned with paternalism, whether of employers or providers, and that they wanted to exercise choice in an environment that made the relationship of costs, benefits, and accessibility more evident than the combination of community rating and sub-rosa utilization controls that managed care had created. Private insurers backed off trying to influence physicians (the fundamental goal of managed care), aimed instead to influence patients in an environment of diverse choice, and tried to appeal to employers who sought to offer employees a menu of options rather than to select plans for them. Under such arrangements, the consumer would have greater room for making choices and greater responsibility for exercising them. 'Consumer-driven healthcare,' a particular set of financial and insurance arrangements, is perhaps the fullest expression thus far of this trend. It reflects the appearance of the middle-class shopper given to evaluating professional services, a phenomenon that market advocates had favorably anticipated. However, studies have shown that the extent to which consumers enjoy clear choices and, where they do, the extent to which they take advantage of them, have been highly limited.

A third theme has been the tendency of market advocacy and attention to market evolution to eclipse the public-interest goals of traditional reformers. After policymakers grew convinced that not only NHI but system reform was also necessary, after Senator Edward M. Kennedy (1932–2009) altered his thinking about healthcare reform to accommodate private markets, and after the failed Clinton plan marked a new check in the work of reformers to achieve NHI and opened the floodgates to dynamic market change, conceptions of the purposes of healthcare markets that depart from traditional collective thinking gained increasing prominence. Indeed, many have argued that the growth and growing

familiarity of markets and the continual rehearsal of their anticipated virtues have entailed consequences many of which were foreseen with apprehension by the earliest critics of markets: Diminished interest in entitlement, access to care, continuity of care; waning of patients' trust in providers; and loss of interest in fragmentation of the health system. What advocates of markets have deemed most important is enhancing efficiency, constraining cost escalation, avoiding paternalism, fostering choice, all without 'rationing care,' long demonized as paternalistic, unaccountable, and simply dangerous. This approach comports with recent cultural developments that have rendered 'the market' an idealization that lacks historical or social content or context. In the minds of their advocates, healthcare markets have not yet reflected or achieved an ideal state, but confident that such a state can be attained, they persist in searching for it.

Accordingly - and here is a fourth theme in the evolution of markets - policymakers' focus on efficiency and consumer choice has compelled reformers oriented to traditional public-interest goals to continually rehearse them and insist on their pertinence and viability. Even early advocates of markets like Pauly and Enthoven, for all their emphasis on care as a set of commodities and markets as the best way to distribute them, held to Schultze's notion that markets existed (or should be created where they did not) to serve articulated public purposes, in the case of healthcare, not only efficiency and cost-control but also improved entitlement; and they stuck with the conviction that the same markets that served the better-off should also accommodate the poor, albeit to buy a lesser standard of care. Moreover, Enthoven originally proposed that governmental regulation was needed to organize a market so as to meet public goals, and therefore he called early for 'procompetitive regulation'. Later he substituted 'managed competition' – not to be confused with 'managed care,' i.e., provision of care by cost-efficiency-oriented bureaucratic organizations – as a means to avoid such problems as risk-selection (e.g., selling insurance to the well and avoiding the ill) and product differentiation that hinders consumers from making comparisons and circumnavigates price competition. He and others held, in brief, that markets required regulation or management to keep their evolution in conformity with public purposes. That such concerns have managed to persist in the face of enthusiasts who reject governmental intervention in markets find testimony in the ACA, which both expands entitlement and organizes markets. The controversy that this legislation has aroused, however, shows that the struggle between market enthusiasts and advocates of traditional public-interest goals has scarcely ended.

These last two themes contrast sharply with experience in most other advanced countries. There, the traditional focus of policymakers lay on regulating or constraining the supply side of the market. Cost-constraining measures in advanced countries have included lower levels of funding; upstream limits on capital; planning; limits on the exploitation of technology; constraints on the size of the medical profession, its composition by specialty, and its geographic distribution; limits on professional fees; global budgets; bargaining among 'peak associations' (i.e., national-level interest groups); gatekeeper systems; explicit rationing and waiting lists; price controls (e.g., on pharmaceuticals); and simpler administrative and payment mechanisms, all of them practices to which the American polity has thus far been vastly less hospitable. Moreover, even the recent experiments that other advanced countries have undertaken with competitive measures – on both demand and supply sides – to foster choice and with it improve efficiency often have been accompanied by regulatory measures to keep their healthcare systems in conformity with underlying solidaristic values. In America, pressure in support of efficiency and choice pose a constant threat to traditional public-interest goals. However, regulation, which market advocates had seen as impediments to the achievement of efficiency and securing of choice, constantly returns through the back door. As diverse market arrangements provoke dissatisfaction from consumers-cum-voters, they demand and get piecemeal protective regulation from the sequelae of market operations. However, few policymakers draw the conclusion that their focus on the efficiency of markets may fail to serve the public and thus require something resembling the practice in other advanced economies of subordinating market arrangements to other social values; rather, they suppose that the ultimate in market arrangements remains to be found.

Concluding Reflections

The themes that this and Health Insurance in Historical Perspective I develop suggest that the ACA is a profoundly American product, tempering as it does the traditional goals of social policy with support for markets and consumer choice. It aims to cover most of the hitherto uninsured, and it preserves and reforms existing market arrangements and adds new ones; but it does not transform the healthcare system into a version of uniform entitlement to comprehensive benefits that traditional reformers long desired. Given the persistence under the ACA of employer-based insurance, of the diversity across employers in costs and levels of coverage, of regressive tax subsidies for private insurance, of Medicare, of Medicaid and its variations across states, of safety-net institutions devised for the poor; and the appearance of new provisions for income-graduated subsidy and cost-sharing, the US has clearly decided to persist in subsidizing care according (primarily) to income, and thereby also (by proxy) according to race, and (secondarily) according to age. Proponents of reducing health disparities (i.e., different levels of health status prevalent among different ethnic and income groups) have recently come to apply the term 'fragmentation' – formerly employed with regard to such things as the 'cottage-industry' structure of the health sector, its lack of focus on the patient, and its inability to coordinate care – to the distinctions drawn in our health system by race, class, gender, and income. These distinctions find expression in the differentials that persist across social groups in access to care, extent and depth of coverage, magnitude of reimbursement, and the kinds and numbers of accessible insurers and providers. Although other health systems in advanced countries also took form with references to such social categories, they persist in the American system to a far greater extent. The ACA offers not a uniform system of NHI, no 'Medicare for all' that some have advocated – HR 646, first introduced into the 111th Congress – no reckoning of care as a prerogative or right attached to citizenship to be equitably

assured, but a system that expresses differential degrees of social success and approval, that affords differential degrees of freedom and responsibility in seeking and gaining access to care, and that provides differential access to care according to socioeconomic status and ethnic and gender identity. Americans have not utterly eschewed a sense of collective responsibility and social solidarity; but their choice of a market-based system seems entirely consonant with their persistence in classifying and discriminating citizens from one another, their privileging the goals of choice and efficiency over social protection, and their seeking in the market an exalted path to realizing and expressing personal autonomy and responsibility.

What role has economics played in the evolution of American health insurance? In no sense has it been determinative of policy choices, in part because economists came to see themselves much more as servants of their masters, public and private, than as reformers or decision makers. Yet economists have scarcely been strictly neutral analysts, for like those for whom they work, they reflect (and in turn have reinforced) the broader cultural and social changes that have helped give rise since the end of the 1960s to a polity and to a population of policymakers more attuned to the values associated with the market – the home turf of economists – and more hostile to government, professional expertise, and paternalism (whether public or private) – than the concerns that traditional policymakers still strive to uphold. If economists have been more influential than in the past, it is the result, in great measure, of this convergence of values. However, their influence also reflected developments in the capacity to analyze public programs that economics as a discipline had begun to show in the late 1950s and 1960s. In a context marked by the problems emergent in the health sector under traditional policy, by the growing concern about cost escalation, and by the fear that expanding access to health services through NHI by extrapolating previous approaches to policy would be too expensive, economists applied to public policy their increasingly mathematized and powerful intellectual tools that had matured in the postwar era. From there flowed the influence of their fundamental individualism, of their arguments about the failures of traditional health insurance, about moral hazard, and about cost-sharing. Moreover, through efforts of this kind, they gave rise to the subdiscipline of health economics and heavily informed the emergent, interdisciplinary field of health services research.

As for analysis of the supply side, the push for competing health plans, rather than only for competition inside a traditional cottage industry, was less an argument of economists than the harnessing of modest institutional precedents by a new set of analysts to remedy the problems in healthcare that cost escalation had rendered acute. Yet as markets involving novel organizations and practices emerged and grew, their development provided grist for the economists' mill. The efficiency of integrated insurers-cum-providers, their incentive-structures, their marketing methods and market shares, their access to capital, their likelihood of serving goals increasingly defined by market-oriented sensibilities (and decreasingly defined by collective sentiments), all this and much more proved amenable to economic study and analysis. Even if the pace of events has often outrun the ability of economists and other health services researchers to keep up, the dynamism of markets and their capacity to serve the preferences of payers, of individuals, and the goals of mostly market-oriented policymakers have opened a vast field for economic analysis. There, too, economists will not and cannot make the value-based decisions that drive policy; but their powerful tools, their professional argot, and the market orientation they share with their employers and many policymakers assure that their will remain influential voices.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Efficiency and Equity in Health: Philosophical Considerations. Efficiency in Health Care, Concepts of. Health Econometrics: Overview. Health Insurance and Health. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Moral Hazard. Welfarism and Extra-Welfarism

Reference

Ellwood, Jr., P. M., Anderson, N. N., Billings, J. E., et al. (1971). Health maintenance strategy. *Medical Care* **9**(3), 291–298.

Further Reading

Ameringer, C. F. (2008). *The health care revolution: From medical monopoly to market competition*. California/Milbank Books on Health and the Public 19. Berkeley, CA: University of California Press and New York: Milbank Memorial Fund.

Buchanan, J. M. (1968). *The demand and supply of public goods*. Chicago: Rand McNally.

Enthoven, A. C. (1980). *Health plan: The only practical solution to the soaring cost of medical care*. Reading, MA: Addison-Wesley.

Feldstein, M. S. (1971). A new approach to national health insurance. *Public Interest* **23**, 93–105.

Helderman, J.-K., Bevan, G. and France, G. (2012). The rise of the regulatory state in health care: A comparative analysis of the Netherlands, England and Italy. *Health Economics, Policy, and Law* **7**(1), 103–124.

Institute of Medicine (IOM) (2009). America's uninsured crisis: Consequences for health and health care. *Board of health care services, committee on health insurance and its consequences*. Washington, DC: National Academy Press.

Jost, T. S. (2007). *Health care at risk: A Critique of the consumer-driven movement*. Durham, NC: Duke University Press.

Klarman, H. E. (1965). *The economics of health*. New York: Columbia University Press.

Melhado, E. M. (1998). Economists, public provision, and the market: Changing values in policy debate. *Journal of Health Politics, Policy, and Law* **23**(2), 215–263.

Newhouse, J. P. and the Insurance Experiment Group (1993). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press. A RAND study.

Nyman, J. A. (2003). *The theory of demand for health insurance*. Stanford, CA: Stanford University Press.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**(3, pt. 1), 531–537.

Robinson, J. C. (1999). *The corporate practice of medicine: Competition and innovation in health care*. Berkeley: University of California Press. California/Milbank Series on Health and the Public 1.

Rodgers, D. T. (ed.) (2011). The rediscovery of the market. In *Age of fracture*, ch. 2, pp 41–76 (text), 280–288 (notes). Cambridge, MA and London, UK: Belknap Press of Harvard University Press.

Schultze, C. L. (1977). *The public use of private interest*. Washington, DC: Brookings Institution. The Godkin lectures at Harvard University, 1976.

Health Insurance in the United States, History of

T Stoltzfus Jost, Washington and Lee University, Harrisonburg, VA, USA

Glossary

Adverse selection A situation that arises when high-risk individuals are more likely than low-risk individuals to purchase insurance. As a result, the average riskiness of people who buy insurance exceeds the average riskiness of the population as a whole. Low risk individuals may choose not to insure at all.

Community rating The setting of health care insurance premia according to the utilization of a broad population (e.g., one defined by employer type or geography).

Experience rating Setting premiums based on an individual or group's claims history.

Introduction

Given the central role that health insurance plays in the American healthcare systems, it is remarkable how short a time it has been with us. Many Americans alive today were born before modern health insurance became available in the United States around 1930. Although brief, the history of health insurance in the United States is sharply contested.

The history of health insurance in the United States is often presented as a narrative of missteps and missed opportunities. Indeed, two contending narratives of policy failure dominate much of the literature describing this history.

The predominant narrative, both in terms of the length of the tradition and volume of scholarship it has produced, emphasizes the failure of the United States to join other developed nations in embracing universal care coverage. Time and again, during the progressive period, in the New Deal, during the Truman Administration in the 1960s and 1970s, and during the Clinton Administration, efforts to establish a universal national health insurance program had come to naught. There were certainly victories along the way – most notably, the enactment of Medicare and Medicaid in 1965. But repeatedly, national health insurance proposals had gone down in defeat.

There is, however, also an alternative narrative of failure, favored by opponents of government intervention in healthcare finance. According to this narrative, repeated unwarranted government intervention in our healthcare system through regulation and subsidies has resulted in excessive cost, inadequate quality, and limitation on choice. Our biggest policy failure has been our refusal to unshackle the free market to work its magic on our healthcare system.

This article recounts both narratives. It will then, however, offer yet another alternative narrative – a story of 'muddling through' and of modest success. In fact, throughout the second half of the twentieth century, the vast majority of Americans were insured. The number of Americans covered by employment-based health insurance expanded very rapidly during the 1940s and 1950s, whereas the scope and extent of coverage continued to expand until the 1980s. Beginning with the 1960s, the Medicare, Medicaid, and State Children's Health Insurance Programs in the 1990s filled the most serious gaps in private coverage. Besides noninsurance 'safety net' programs, the Emergency Medical Treatment and Active Labor Act, which requires hospitals to provide emergency treatment regardless of ability to pay (although not for free), filled yet another gaps. Only with contractions of private coverage in the 1990s, greatly accelerated in the 2000s, did this patchwork of insurance coverage become truly unsustainable.

The article concludes with an analysis of the Patient Protection and Affordable Care Act of 2010, which attempts to build on the United States' unique mix of private and public health insurance to fill the growing gaps in coverage that have become apparent at the beginning of the twenty-first century. The extent to which this fix, in fact, succeeds, certainly, remains to be seen.

A History of Political Failure: Attempts to Achieve Universal Coverage

The dominant account of the history of health insurance in the United States focuses on failed attempts to create universal health coverage. The first attempt to establish universal health coverage in the United States was led by the progressive movement in the late 1910s. Germany had inaugurated a social health insurance program in 1883, followed by a number of other European countries in the 1890s and early 1900s. The success of the efforts of the progressives to expand social welfare programs at the state level led the American Association for Labor Legislation (AALL) to believe that a national sickness insurance program might also succeed. The AALL marshaled a coalition of progressive academics and enlightened business leaders, who pushed for reform based largely on the German model. By 1917, the AALL's standard health insurance bill was being considered in 15 state legislatures.

Then everything fell apart. Some labor leaders opposed the government taking over the provision of welfare benefits to workers, a role that they coveted for themselves. Business leaders consolidated their opposition to the legislation. After a brief initial period of openness to change, organized medicine retreated to a stance of obdurate and highly effective opposition, which it assumed toward public health insurance for decades thereafter. Insurance companies, which as of yet sold little health insurance but had developed a substantial market

 doi:10.1016/B978-0-12-375678-7.00903-2

for industrial life insurance policies, opposed the proposal, which would have offered burial policies as part of the sickness benefit. Finally, as America was drawn into the First World War, enthusiasm for German things quickly waned. Compulsory health insurance legislation was defeated in California and New York, and by 1918, social health insurance was no longer on the table.

The possibility of a national health insurance program flickered to life again briefly during the 1930s. The severe economic dislocation of the Great Depression quickly overwhelmed state, local, and private relief efforts. The Social Security Act enacted in 1935 created a national social insurance retirement income program for the elderly and offered federal subsidies for state cash assistance program for the poor elderly, dependent children, and the blind. Although there was considerable support for a federal program that would provide health benefits, fervent opposition led by organized medicine threatened to bring down the entire social insurance program if health insurance was a part of it. President Roosevelt ultimately abandoned social health insurance.

Repeated attempts to create a universal social health in the aftermath of war also proved unsuccessful. Although President Truman campaigned for a national health insurance program more vigorously than what Roosevelt did before him, the United States turned rightward following the war, electing a Republican congressional majority. The most important parts of Truman's program that survived Congressional debate were the Hill-Burton hospital construction program (which between 1947 and 1971 disbursed US$3.7 billion in federal funds for hospital construction, contributing up to 30% of all hospital projects during the period) and a heavy federal commitment to healthcare research. There was also quiet expansion of healthcare for the poor. The Social Security Act Amendments of 1950 for the first time committed the federal government to match, to a very limited extent, state expenditures for in-kind medical services through the matching fund provisions of the federal/state public assistance programs for the elderly, blind, and disabled, as well as families with dependent children. Federal assistance for state indigent healthcare plans program was further expanded by the Social Security Act Amendments of 1960, which created the Kerr-Mills program to provide federal matching funds for a medically needy elderly.

The anticommunism of the late 1940s and early 1950s and continued opposition from organized medicine put a hold on any further attempts to create a national health insurance program. Nevertheless, pressure for national health insurance was quietly building among organized labor and the elderly, who adapted strategically scaling back their expectations by limiting their immediate goals to cover only social security beneficiaries with social health insurance, and by 1960, to the coverage of hospital care.

With the election of President Kennedy in 1960, efforts to provide healthcare for the elderly were redoubled. The landslide election of President Johnson and of a liberal Democratic Congress following the assassination of Kennedy finally made health reform inevitable. In 1965, Congress created the Medicare program to insure hospital and medical services for the elderly as well as the Medicaid program to pay for healthcare services for public assistance recipients and the medically needy.

Social insurance advocates had hoped that the enactment of Medicare and Medicaid would be followed up by expansion of public insurance to the entire population. The 1970s, however, brought little progress as Democrats in Congress failed to reach agreement with the Nixon administration regarding the way forward, and the Carter administration focused (largely unsuccessfully) on cost control rather than on coverage expansion. Medicare coverage was expanded to the disabled, but no further.

Although the 1980s saw expansion in the Medicaid program, universal health coverage was off the agenda during the Reagan administration. The election of Bill Clinton in 1992, who campaigned for healthcare reform in the light of a growing number of uninsured Americans and rapidly increasing healthcare costs, brought new hope to reform advocates. However, Clinton administration stumbled politically. It took a year and a half to craft a reform plan in secret, giving interest groups and political opponents time to rally opposition and devising at last a plan that was too complex and could not gain traction. The late 1990s saw the creation of the State Children's Health Insurance program, followed by the expansion of Medicare to cover outpatient prescription drugs in 2003, without which another two decades were likely to be lost in the quest for universal coverage.

Analysts offer a variety of explanations for America's inability to adopt universal coverage. These include a national ideological aversion to strong government, powerful interest groups that benefit from the status quo, the absence of a strong political left, political institutions that make it far easier to block than accomplish change, and path dependency. Each of these explanations explains part of the problem, although the saliency of any particular explanation varies from one decade to another.

The 'failed attempts to adopt universal coverage' narrative would seem to be an accurate description of the history of health insurance coverage in the United States as far as it goes but does not fully acknowledge the remarkable expansion of private health insurance, which has played a more central role in the United States than it has in most other developed nations (Switzerland and, more recently, the Netherlands being the main exceptions). It is to that story to which the authors will shortly turn. The author will also consider whether the adoption of the 2010 Affordable Care Act provides a happy ending to the narrative of failure. But first, the free market advocate alternative for 'history of failure' narrative will be considered.

The 'Government Interference with Health Care Markets' Narrative: A Narrative of Economic Failure

Although the failure of universal coverage narrative focuses on the plight of uninsured and underinsured Americans, the government interference narrative contends that Americans are 'overinsured.' Americans have too much insurance because of government policies that have encouraged private insurance for routine as well as catastrophic medical costs, thus resulting in severe moral hazard (as well as too much public health insurance and government regulation).

The history of American overinsurance begins, according to this narrative, with the exemption of fringe benefits in wage-price controls during World War II, thus stimulating the former's growth. Also dating from the 1940s, are tax subsidies for employment-related insurance that have encouraged the provision of excessive health insurance coverage for most Americans. Because insurance premiums have largely been covered by employers, the true cost of health insurance has been concealed from Americans. Because the predominant forms of health insurance have imposed little costsharing, the true cost of healthcare has been concealed as well. Finally, the Medicare and Medicaid programs have driven up healthcare prices and utilization, limited choices for the elderly, and discouraged provider innovation. Repeated attempts by the government to fix health insurance market failures have only worsened the situation.

There is some truth in this narrative despite offering only a partial picture of American developments. In fact, labor was scarce during World War II and excess profit taxes were very high, up to 85%. The Stabilization Act of 1942 did allow the National War Labor Board (NWLB) to exclude a 'reasonable amount' of insurance benefits from wage controls. An IRS administrative ruling of 1943 also allowed businesses to deduct payments toward health and welfare funds as business expenses, contending that these benefits would not be taxable to employees.

Yet, there is reason to be skeptical of the oft-repeated claim that wage policy was the primary reason for the expansion of health insurance coverage during the War. First and foremost, health insurance as an employee benefit was already well established and rapidly growing before the war began, as described below. Second, most of the growth in wartime employment and health insurance coverage took place before the NWLB policies came into effect in 1943. American industry had been gearing up for the looming war since 1939, and while the number of American employees insured through commercial plans (the plans most likely to be paid for in part by the employer) increased from 960 000 in 1939 to 4.3 million in 1943, it only increased to another 71 000 between 1943 and 1945. Employment-related insurance coverage increased again rapidly after wage price stabilization controls expired in 1946, suggesting yet again that expansion was not driven primarily by wage stabilization policy. The wage stabilization policy was in any event routinely circumvented, as it allowed wage increases in conjunction with promotions, which quickly became common. Finally, throughout the war, Blue Cross coverage, the most common form of health insurance, continued to be paid for largely by employees rather than employers. By the end of the war, only 7.6% of Blue Cross enrollees were participants in groups to which employers contributed.

There is more reason to credit the employee benefit tax exclusion and deduction for the increase in health insurance coverage in the United States. The most rapid growth in health insurance coverage, however, took place in the late 1940s and early 1950s before the tax subsidies were enshrined in the 1954 Tax Code, and probably had more to do with aggressive collective bargaining by the unions than tax subsidies. The tax subsidies, however, undoubtedly contributed to the expansion of the scope and depth of health benefits well into the 1990s.

It is also very likely that the expansion of benefits has contributed to the growth in healthcare costs. Free market advocates assert that the Rand Health Insurance Experiment (HIE) conclusively demonstrated that more comprehensive health insurance coverage leads to higher healthcare spending. Although the meaning of the HIE and its relevance to contemporary health policy continue to be debated, the correlation between broader insurance coverage and increased healthcare spending seems plausible. It is also clear that the creation of Medicare and Medicaid has resulted in higher healthcare spending at least due to more people being insured.

Market advocates generally argue for the removal of tax incentives for private health insurance coverage and for the scaling back the operation of government healthcare programs through the use of vouchers to pay for private health insurance. Their most significant legislative victory has been the Medicare Modernization Act of 2003, which provided tax subsidies for health savings accounts coupled with high deductible health plans. High deductible health insurance has spread rapidly during the early 2000s and now dominates the individual market. This has resulted in increased financial difficulty for insured families and reduced access to healthcare. However, increased cost sharing has also arguably had a restraining effect on healthcare costs.

An Alternative Narrative: A Modestly Successful Patchwork of Coverage

The Origins of Modern Health Insurance

There is yet a third narrative of the history of health insurance in the United States that is somewhat more sanguine. Health insurance came into existence in the United States in the first half of the twentieth century as advances in medicine made healthcare of real value and increases in the cost of healthcare rendered it increasingly less affordable to those with serious medical problems. The prestigious Committee on the Costs of Medical Care concluded in its 1932 final report that the high cost of medical care for those most in need necessitated the provision of either private or public insurance, but by that time private insurance was already in use.

Describing the early history of health insurance is problematic because of a different meaning of the term 'health insurance' before the mid-twentieth century. The late nineteenth and early twentieth centuries saw the rapid growth of what was then called health insurance or sickness insurance. This coverage insured against wages lost due to illness. After a short waiting period, an insured individual would be able to collect a fixed amount per week until he (or, rarely, she) was able to return to work or until the benefit was exhausted. This insurance was offered by employer-funded 'establishment funds', labor organization funds, and commercial insurers; as well as by ethnic, religious, and community-based fraternal organizations. Some of these insurers and funds also provided life or burial insurance. Although a few offered insurance to cover medical costs, most did not. Not only was the value of most medical care questionable, fund members were also apparently concerned

that a doctor paid for by the fund might be too eager to certify the member healthy enough to return to work.

Other precursors of health insurance also emerged during the late nineteenth and early twentieth centuries. Some fraternal organizations hired physicians to provide care to their members – the much maligned 'lodge practice'; some even built their own hospitals. Employers in remote areas like in the case of railroad, mining, or logging companies also provided medical services through company doctors or through industrial medical plans.

Modern health insurance was born in 1929. In that year, the first 'hospital service plan' was started by Baylor Hospital in Dallas in 1929. Baylor entered into a contract under which white public school teachers paid 50 cents a month into a prepaid hospital services annual plan with the assurance that they would receive up to 21 days of hospital care, and a one-third discount for the remaining 344 days.

Hospital service plans did spread quickly during the 1930s. In 1936, the American Hospital Association established the Commission on Hospital Services, which ultimately became the Blue Cross Association. This commission encouraged and supported the spread of state and regional Blue Cross plans. By 1937, Blue Cross plans had 894 000 members; by 1943, membership reached almost 12 million.

Blue Cross plan members paid a fixed sum every month for the assurance that their needs would be covered if they had to be hospitalized. Blue Cross plans were available on a community-rated basis, that is, all members paid the same rate, regardless of health status. The plans negotiated 'service benefit' contracts with the hospitals under which the plans would cover up to a fixed number of days of hospitalization for a per diem fee established in the contract. Blue Cross plans also provided either service benefit or indemnity coverage (under which insureds would pay medical providers in cash and then file a claim with the insurer for an indemnity payment) for 'extras' such as emergency and operating room charges, or laboratory tests.

As it became increasingly clear in the late 1930s that there was a substantial market for hospital benefits, private commercial insurers too entered the group insurance market. Whereas only 300 000 Americans were covered by commercial hospitalization policies in 1938, nearly six million had coverage in 1946. Unlike Blue Cross plans, commercial insurance covered hospital care besides offering surgical coverage. By 1945, over five million Americans had commercial surgical coverage. Commercial plans even began to cover medical costs (nonsurgical physician's services) in the hospital. By the late 1950s, home and office visits also began to be covered, especially under individual policies. Commercial health insurance was sold on an indemnity basis. Indemnity payments would be for fixed sums per service, which were set forth beforehand in the insurance contract.

The success of the hospitals in offering prepaid benefits was soon noticed by physicians. In 1939, the first of the physician service benefits plan that came to be known as Blue Shield plans appeared. Blue Shield plans initially covered surgical benefits in hospital, expanding later on to cover in-hospital medical and eventually ambulatory medical benefits.

Blue Shield plans combined the Blue Cross and commercial insurance approaches for providing benefits. Although some plans offered only service benefits or only indemnity coverage, most of them offered both. Doctors agreed to accept negotiated payments from the plans as payment in full for patients whose income fell below a specified level. Members with incomes above such levels, however, received indemnity payments and had to pay the difference between the doctor's charge and the indemnity amount. Blue Shield plans were initially community-rated, but over time moved to experience rating like the commercial insurers.

The year of 1929 saw the birth of other models of health insurance as well. In that year, the first consumer's cooperative providing prepaid medical care was created in Elk City, Oklahoma, whereas the Ross-Loos Clinic, a physicians' cooperative, began offering a prepayment plan for an employment-related group in Los Angeles. During the 1930s and 1940s, other models of health care coverage appeared based on comprehensive prepayment for healthcare. Some of these, such as the Kaiser plan, were initially industry-sponsored wherease others, like the Washington Group Health Insurance Plan, grew out of consumer-sponsored plans. The Farm Security Administration encouraged consumer cooperatives, which covered 725 000 persons by the early 1940s, but largely disappeared when government support ended. Industry-sponsored plans also continued to exist, covering approximately a million people in 1930.

These precursors of modern staff-model health maintenance organizations (HMOs) were vigorously opposed by organized medicine. Organized medicine preferred cash-and-carry medicine (as it does today), but was willing to tolerate insurance that did not subject doctors to lay control. Lay control of medical practice was unacceptable, and health plans that employed doctors were fought vigorously by the American Medical Association (AMA) through much of the twentieth century, resulting in a criminal conviction of the AMA for antitrust violations in the 1940s. These efforts by the AMA kept prepaid medical practice marginal until the final quarter of the twentieth century.

Initially, Blue Cross, Blue Shield, and commercial plans were sold primarily to groups. It was much less expensive to market health insurance to groups than to individuals. Insuring employment-related groups in particular helped for addressing the problem of adverse selection. Blue Cross plans sold insurance to groups of various types, primarily, however, they contracted with employment-related groups. Employers permitted the sale of group policies to their employees, facilitated the formation of groups, and often deducted the premiums from pay checks through a payroll check-off system.

At the outset, employers themselves rarely contributed to premiums for the Blue Cross plans. As late as 1950, only 12.2% of Blue Cross plan participants received employer contributions. Employer contributions were more common with commercial plans. By 1950 employers contributed approximately half of the 'gross cost' of health insurance for employees and 30% of the cost of dependent coverage. Because employers commonly received rebates from insurers, their actual 'net cost' was in fact much lower, approximately 38.5% for employees and 20% for dependents. A major focus of collective bargaining agreements was to shift more of the cost of the premium to the employer.

Health Insurance in the Mid-Twentieth Century

In the booming American economy following World War II, health insurance coverage expanded dramatically. By 1950, nearly 76.6 million Americans constituting half the American population had hospitalization insurance – 54.2 million had surgical benefits, and 21.6 million had medical benefits. By 1965, when Medicare and Medicaid were adopted, private hospital insurance covered 138.7 million Americans, that is, approximately 71% of the American population.

As coverage expanded, it also became more comprehensive. In the early 1950s, commercial insurers began to offer major medical coverage that provided catastrophic coverage for hospital and medical care. Major medical policies usually supplemented basic hospital and surgical-medical coverage. Comprehensive coverage followed soon on its heels, bundling basic and major medical coverage into a package to provide the most complete coverage available. During the 1950s, Blue Cross and Blue Shield plans began to combine forces to offer similarly comprehensive coverage. Finally, during the 1960s and 1970s, insurance coverage began to expand to cover dental care and pharmaceuticals, with improved coverage for maternity care, mental health, and some preventive services within basic coverage.

Another important trend after the War was the increased employer responsibility for employee health benefits. During the late 1940s and early 1950s, employer contributions to collectively bargained plans increased exponentially. By 1959, employers paid the entire premium for hospital insurance for virtually all unionized employees in multiemployer plans and for 37% of employees subject to collective bargaining agreements in single-employer plans.

Employer contributions to premiums in nonunionized places of employment increased more slowly. By 1964, however, approximately 48% of employees had the total cost of their health insurance covered by their employer. Employer contributions to health insurance expanded even more quickly during the 1970s and 1980s. By 1988, employers covered 90% of the cost of individual coverage and 75% of the cost of family coverage.

Among the several reasons for the impressive postwar expansion in the number of workers covered, the benefits provided, and the level of employer contributions in the third quarter of the twentieth century, the most important one was probably pressure from the labor unions. Unions were at the peak of their strength in the mid-twentieth century. Improved fringe benefits were a high priority for the unions. The National Labor Relations Board clarified in 1949 that employee benefits were included within the 'terms of conditions of employment' subject to collective bargaining under the National Labor Relations Act, giving new impetus to union demands for health benefits.

In the beginning, some of the major unions such as the United Mine Workers had operated their own health benefit funds. The 1947 Taft-Hartley Act prohibited union-run benefit plans, but established multiemployer Taft-Hartley plans, which were operated jointly by labor and management. Most employee benefit plans, however, were established by management. Unions tended to favor Blue Cross and Blue Shield contracts, which offered more comprehensive coverage, but large employers favored commercial insurers that offered more flexibility in the design of plans as well as generous rebates, which substantially reduced the employer's net contribution to premiums. Employers with healthy workforces also favored commercial insurers because they used experience rating and thus could offer lower rates.

Health benefits were not limited to unionized firms. Even firms that were not unionized offered liberal fringe benefits to forestall unionization. Employers also saw health insurance as a means to stabilize employment (by making it more difficult for employees to leave), to keep workers healthy and productive, and to ward off a national social health insurance program.

Another factor underlying the growth of employment-related insurance was the continuing increase in healthcare costs. The proportion of the gross domestic product spent on healthcare grew from 3.6% in 1928–29 to 5.4% in 1958–59. Changes in medical technology were making medical care much more effective and thus more valuable, although medical care was becoming less affordable. The growing burden of healthcare costs led, in turn, to an increased desire to spread costs through insurance and pass it on to employers.

Tax policy also certainly played a role. The 1954 Internal Revenue Code explicitly recognized the nontaxability of employment-related benefits. As more and more Americans began to pay income tax (which was paid primarily by the wealthy before World War II), the tax benefits of health insurance became more important. Tax subsidies played a particularly important role in increasing the share of premiums covered by employers as well as the scope of coverage.

A final factor that drove the expansion of employee coverage was the enactment of the Employee Retirement Income Security Act (ERISA) in 1974, which blocked the application of state insurance regulation and premium taxes to self-insured plans. Self-insurance gave employers increased power to control healthcare costs and the opportunity to receive interest on reserves, as well as protecting them from state premium taxes, insurance mandates (which became common in the early 1980s), capital and reserve requirements, and risk pool contribution requirements. Whereas only 5% of group health claims was paid by self-insured plans in 1975, an estimated 60% of employees were in self-insured plans by 1987.

Although most American employees had hospital coverage (and increasingly surgical and medical coverage) by the 1970s, that coverage was often quite thin. Until the 1980s, commercial insurance was predominantly indemnity coverage and balance billing was very common. Moreover, dollar limits on coverage were often quite conservative. As late as 1959, when 72% of the population had hospital insurance, 18.4% of personal care expenditures was covered by insurance, whereas 56.5% had to be paid. Blue Cross plans offered first-dollar coverage, but initially limited the number of days of hospitalization they would cover, whereas Blue Shield plans often offered indemnity coverage to higher-income enrollees.

Coverage, moreover, did not reach many who were not employed. The one group that was most noticeably left behind during the coverage expansion was the elderly. Retiree health coverage expanded rapidly during the 1950s and 1960s, and many of the elderly purchased individual insurance, yet many

remained uninsured too. Efforts to provide public insurance for this group came to fruition in 1965 with the creation of the Medicare program, described earlier. The Medicaid program too offered supplemental coverage to the elderly and disabled besides basic coverage to impoverished families with dependent children.

Other new programs also began to partially fill other gaps left by private insurance. Community health centers that provide services to lower-income families on a sliding scale basis were launched in 1964. Provisions of the 1949 Hill Burton hospital funding program, requiring grantees to provide free or reduced cost care to those in need, finally began to be enforced in the 1970s. The 1986 Emergency Medical Treatment and Active Labor Act required Medicare-participating hospitals to provide emergency services even to those unable to pay (although not free). The 1986 budget bill also included a provision that allowed persons who lost their employment or their dependency status to purchase continuation coverage for a period of time at full cost (so-called, COBRA continuation coverage).

By 1980, the vast majority of Americans had health insurance coverage through their employment, and this coverage was increasingly comprehensive. 82.4% of the population had private health insurance that year, a proportion not yet repeated. Most employers paid the full premium for individual coverage and the majority of the premium for family coverage. Deductibles and coinsurance remained common, and indeed spread to Blue Cross and Blue Shield plans, but with the advent of major medical and then comprehensive coverage, out-of-pocket expenditures decreased and insured expenses increased in the final quarter of the century. By 1980, the proportion of healthcare costs covered by private health insurance exceeded that covered out-of-pocket, and with the advent of HMOs in the 1980s, cost-sharing virtually disappeared. The United States had apparently solved through private initiative, supplemented by public programs for those whom private markets could never protect, the problem of health security that other nations addressed through social insurance or public provision.

Private Health Insurance Unravels

However, America's health security system began to unravel during the early 1970s. The driving disruptive force was the increase in healthcare costs. Inflation generally was a serious problem during the 1970s, but healthcare costs grew even more rapidly than other costs. Public initiatives were adopted to restrain healthcare cost growth – including health planning, professional standards review organizations, and in some states, hospital rate review – but none achieved great success.

During the 1980s and early 1990s, health insurers responded to cost increases by turning from being passive payers to becoming care managers. Within a decade, conventional indemnity insurance and service benefit plans gave way to plans, initially called HMOs and preferred provider organizations, which offered limited provider networks, attempted to review and control utilization, and experimented with incentive structures that would discourage rather than encourage provision of services. This strategy worked for sometime. By the mid-1960s, healthcare cost growth had declined dramatically, indeed it briefly fell in line with the general growth of the economy.

But cost increases also began to have an impact on coverage. Beginning on with the 1980s, the percentage of Americans with health insurance began to decline. The first to lose coverage were retirees, who fell victims to the declining power of the unions (which had been their strongest champions), to the steady increase in healthcare costs, and to a change in accounting standards after 1990 that required firms to consider the cost of future retiree health obligations as a current liability on their books.

Moreover, small businesses had never covered their employees to the same extent as larger businesses, and as the American economy shifted from a manufacturing to a service-based economy – and concomitantly from large unionized employers to small businesses, the percentage of employees who were insured began to fall further.

Small groups have been underwritten for decades on the basis of expected claims costs of their members, and coverage can be very expensive, even difficult to find, for older groups or groups in hazardous occupations. A number of states took steps in the 1990s to make health insurance more accessible for small groups. This included statutes guaranteeing insurance issue and renewal, limiting variations in rating among groups; and restricting the preexisting condition exclusions. A few even required community rating. The 1996 federal Health Insurance Portability and Accountability Act established guaranteed issue and renewal requirements throughout the country and imposed limits on the preexisting condition exclusions. Administrative costs, however, remained significantly higher for small groups than for large groups and health status underwriting continued for small groups in most states.

Even for larger groups, managed care succeeded in stemming the growth of healthcare costs only temporarily. The more extreme forms of managed care proved intensely unpopular. Although Congress failed to adopt a national managed care bill of rights when the issue came before it in 2001, most states adopted legislation restraining managed care in the late 1990s. As the economy improved in the late 1990s and early 2000s, employers backed off from the most stringent forms of managed care, moving to broader provider networks and away from strict utilization controls.

Healthcare costs began to rise dramatically again by 2000, however. As the economy worsened again in the mid-2000s, the cost of employment-related health insurance began to reach levels that employers found intolerable. Employers reacted primarily by increasing employee cost-sharing, although some employers dropped coverage or increased the employee share of health insurance premiums. Many employers shifted to high-deductible policies, sometimes offering contributions for health savings accounts (held by the employee) or health reimbursement accounts (held by the employee), which received tax subsidies under the 2003 Medicare Modernization Act. As health insurance became more costly and less valuable to employees, more employees passed up employment-related coverage.

Public programs grew steadily for some time, offsetting the decline in private coverage. Employment-related coverage had never covered dependents on the same terms as workers, and many lower income workers could not afford the premiums

required for family coverage if their employers even offered it. Many children, therefore, remained uninsured. Medicaid coverage for children had steadily expanded through the 1980s and 1990s, and in 1996, the State Children's Health Insurance Program was created to cover children in families with incomes up to 200% of the poverty level and above. Medicaid was also expanded after 1981 to cover pregnant women, recognizing the cost-effectiveness of timely prenatal care.

The massive layoffs and economic retrenchment that accompanied the economic decline in 2008 and 2009, however, accelerated the decline in private coverage, overwhelming the expansion of public coverage. Only 55.8% of Americans had employment-based coverage by 2009, down from 63.9% in 1989. The decline of insured retirees had been even steeper. Only 28% of large firms that offered health benefits covered retirees in 2010, down from 66% in 1988.

A small percentage of Americans have always been insured through the individual market. Administrative costs are even higher in the individual non group market than in the small group market, and premiums vary sharply from individual to individual based on health status, age, and other underwriting factors. Individual insurance, however, is often the only alternative available for a growing number of self-employed Americans, including, early retirees, the unemployed, part-time and temporary workers, and individuals who do not have insurance available through their place of employment. A number of states attempted in the 1990s to reform the nongroup market, but in most states reforms were not as ambitious as small group market reforms. The Health Insurance Portability and Accountability Act required only guaranteed renewal and imposed limits on the exclusions of preexisting conditions for individuals who transfer from group insurance or some equivalent public insurance. Many states also established high risk pools for otherwise uninsurable individuals, but risk pool premiums were high and participation was generally low. Individual plans are now predominantly high-deductible plans, with the most common deductible levels in 2009 being US$2500 for individual policies and US$5000 for family policies. The individual market is characterized by high premiums and high turnover, but it is the only coverage currently available to many Americans.

In summary, the history of American health insurance is a story of successes and failures. It is true that healthcare costs have been growing at rates in excess of general inflation almost without interruption for the past half century and that the number of uninsured Americans has now reached critical levels – 50.7 million or 16.7% of the population in 2009. But the vast majority of Americans had access to healthcare for a half century through private health insurance and those who had the most difficult time accessing insurance were covered through public insurance. Can we, however, do better?

The Patient Protection and Affordable Care Act of 2010

The Patient Protection and Affordable Care Act of 2010 (ACA) represents an additional article in each of these three narratives. Some, although not all, of its supporters laud it as finally achieving the long-dreamed of goal of healthcare coverage for all. In fact, if all goes according to plan, the legislation should dramatically expand health insurance coverage and reduce the number of the uninsured. The legislation expands Medicaid to cover all American citizens and long-term legal residents with incomes of up to 138% of the federal poverty level and offers tax credits to help cover the cost of health insurance premiums for Americans and legal residents with incomes of up to 400% of the poverty level. It imposes a penalty on Americans who do not purchase health insurance and penalizes employers who do not offer health insurance or provide inadequate coverage to their employees. The Congressional Budget Office estimates that by 2020, the legislation will reduce the number of the uninsured by 32 million, but it will still leave 23 million Americans (including undocumented aliens) without health insurance. The dream of universal coverage is not yet fulfilled.

Free market advocates loudly criticize the ACA as a 'government takeover' of the healthcare system. They complain that the legislation extends government subsidies for and regulation of the healthcare system even further. They fret that the expansion of health savings and reimbursement accounts that they achieved in the early 2000s will be overturned. They assert that the legislation will result in unconstrained growth in healthcare costs.

The ACA does dramatically expand federal funding and regulation of private health insurance. It does not, however, significantly expand federal regulation of the healthcare delivery system. Fundamentally, moreover, the ACA adopts a market-based approach to healthcare reform. It establishes 'health insurance exchanges' at the state level to organize competition among health plans. It establishes new programs to increase competition by encouraging the extension of multistate private plans to every state and the formation of interstate insurance sales compacts and nonprofit insurance cooperatives. Finally, the legislation has no effect on health savings or reimbursement accounts other than to limit their use for over-the-counter drugs. Indeed, the normal employment-based policy currently has an actuarial value of over 80%, whereas the standard subsidized 'silver' policy under the ACA will have an actuarial value of 70% ('actuarial value' refers to the percentage of total medical costs of a standard population paid for by an insurer, so the lower the actuarial value, the higher the percentage of medical costs borne by the insured. Most current health savings accounts-affiliated high deductible plans will be permissible as 60% actuarial value 'bronze' plans. There is likely to be most, not less, cost-sharing under the ACA.

However, the ACA is best understood finally in terms of the 'patchwork of coverage' narrative. The legislation is in the long American tradition of expanding private health insurance coverage and filling gaps with public coverage. Once the ACA is fully implemented, most Americans will continue to be covered by employment-related health insurance, Medicare, and Medicaid. The ACA significantly expands Medicaid, acknowledging that Americans below 138% of the poverty level simply cannot afford health insurance although the Supreme Court decision limits the number of poor Americans who will benefit from this expansion. Tax credits and cost reduction subsidies are offered to allow Americans earning up to 400% of poverty to purchase health insurance and to limit their

exposure for cost-sharing, thus making insurance affordable to 19 million more Americans.

The biggest change made in the American health insurance is that the legislation outlaws health status underwriting and bans preexisting condition exclusions. Insurers must no longer compete based on risk selection but rather do so based on price and value. The original Blue Cross plans community-rated premiums, and community rating has long been the norm (and required by law since 1996) within employee groups. Outlawing health status reinforces this tradition while rejecting an equally long tradition of health status underwriting. The legislation also prohibits or limits other health insurance practices and policy provisions – some of which, like the imposition of annual limits, date back to the beginning of health insurance, whereas others, like limitations on access to certain specialists, are more recent.

The ACA fits within the narrative of the quest for universal coverage, and can also be understood as imposing additional government regulation and subsidies on healthcare markets (albeit to the prospect of making them function better), but it is best understood as one more article in the ongoing story of helping our patchwork private/public health insurance system to hobble along.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance and Health. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare. Health Insurance Systems in Developed Countries, Comparisons of. Health-Insurer Market Power: Theory and Evidence. Managed Care. Medicare. Moral Hazard. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Private Insurance System Concerns. Social Health Insurance – Theory and Evidence

Further Reading

Avnet, H. H. (1944). *Voluntary medical insurance in the United States*. New York: Medical Administration Services.

Committee on the Costs of Medical Care (1932). *Medical care for the American people*. Chicago: University of Chicago Press.

Committee on Labor and Public Welfare, United States Senate (United States Senate, 1951). *Health Insurance Plans in the United States*. Washington: Government Printing Office.

Cunningham, III, R. and Cunningham, Jr., R. M. (1997). *The Blues: A history of the Blue Cross and Blue Shield system*. Dekalb: Northern Illinois University Press.

Dobbin, F. R. (1992). The origins of private social insurance: Public policy and fringe benefits in America, 1920–1950. *American Journal of Sociology* **97**, 1416–1450.

Field, M. J. and Shapiro, H. T. (eds.) (1993). *Employment and health benefits*. Washington: National Academy Press.

Goodman, J. C. and Musgrave, G. L. (1992). *Patient power: Solving America's health care crisis*. Washington: Cato Institute.

Hacker, J. S. (2002). *The divided welfare state*. New York: Cambridge University Press.

Health Insurance Association of America (1959–2002). *Source book of health insurance data*. Washington: HIAA.

Ilse, L. W. (1953). *Group insurance and employee retirement plans*. New York: Prentice-Hall.

Jost, T. S. (2007). *Health care at risk: A critique of the consumer-driven movement*. Durham: Duke University Press.

Marmor, T. R. (2000). *The politics of medicare*, 2nd ed. Hawthorne, NY: Aldine de Gruyer.

Murray, J. E. (2007). *Origins of American health insurance*. New Haven: Yale University Press.

Quadagno, J. (2005). *One nation, uninsured*. New York: Oxford University Press.

Starr, P. (1982). *The social transformation of American medicine*. New York: Basic Books.

Thomasson, M. (2002). From sickness to health: The twentieth century development of U.S. health insurance. *Explorations in Economic History* **39**, 233–253.

Weiner, J. P. and De Lissovoy, G. (1993). Razing a tower of Babel: A taxonomy for managed care and health insurance plans. *Journal of Health Politics, Policy and Law* **18**, 75–103.

Health Insurance Systems in Developed Countries, Comparisons of

RP Ellis, T Chen, and CE Luscombe, Boston University, Boston, MA, USA

Glossary

Ceiling The limit on the dollar payments or visits covered by a health plan.
Claims The payments for consumer losses covered by health plans.
Coinsurance The proportion of healthcare cost paid by the consumer, for example, 20%.
Complementary insurance Insurance that covers part of the consumers' cost share of their primary plan.
Copayment A fixed-money amount paid per day or unit of service, for example, US$10 per office visit.
Cost-sharing, demand-side The healthcare costs paid by the consumer, which can be copayments, coinsurance, deductibles, or amounts paid above a coverage ceiling.
Cost-sharing, premium The share of premium paid by consumers rather than a sponsor.
Cost-sharing, supply-side The healthcare costs borne by providers.
Deductible An amount up to which the consumer pays the full price for healthcare; hence, the consumer might pay the first US$500 deductible without any copayment.
Duplicate insurance Insurance that provides coverage for benefits already included in the primary insurance program, which may have further benefits, including jumping ahead in a waiting line.
Health savings account A system of self-insurance in which funds are deposited by a consumer or sponsor and available for reimbursing healthcare expenses in the current or future year.
Managed care An insurance program in which utilization constraints are used to control costs.
Pay for performance The payments determined based on some observed measures of providers.
Premium Fixed payment per unit of time (e.g., per year) for a defined set of healthcare services.
Primary insurance The system of insurance used for the dominant group in every country, who are employed workers and their dependents.
Replacement insurance Insurance purchased as an alternative to the primary insurance. It is not clearly defined for the US.
Risk adjustment The use of information to explain variation in healthcare spending, resource utilization, or health outcomes over a fixed period of time.
Secondary insurance Insurance that adds to, or replaces, the coverage provided by primary insurance.
Selective contracting Providers can choose whether to contract with some or all health plans, and health plans can choose whether to contract with only some providers.
Self-insurance Consumers bearing the full risk of health expenditures through savings. Consumers are also their own sponsors.
Social insurance A system of insurance in which benefits are defined by statute, revenue generation is primarily income based, and participation is mandatory.
Specialized insurance The insurance programs designed to serve specialized populations, which could be elderly, children, disabled, or having certain specified chronic conditions or high health costs.
Stoploss A limit on the amount of payment by an agent, such as a consumer or health plan.
Supplementary insurance Insurance that provides coverage for services not covered by the consumer's primary insurance plan.

Introduction

There is an enormous literature evaluating and comparing health insurance systems around the world, which this article attempts to synthesize while emphasizing systems in developed countries. The authors' approach is to provide an overview of the dimensions along which health insurance systems differ and provide immediate comparisons of various countries in tabular form. To organize their analysis, they focus their discussion on coverage for the largest segment of the population in all developed countries: workers under the age of 65 years earning a salary or wage, which they call the primary insurance system. They later touch on the features of special programs to cover the elderly, the poor or uninsured, and those with expensive, chronic conditions. They do this not because these groups are less important, but rather because special programs are often used to generate revenue and provide services to these groups, and including these programs in their discussion adds considerable complexity. For the same reason, they also focus on primary insurance coverage of conventional medical care providers – office-based physicians, hospital-based specialists, general hospitals, and pharmacies – knowing that there are many specialized insurance programs for long-term care, specialty hospitals, informal providers, and certain uncovered specialties.

A key feature of their analysis is that they focus on providing a broad framework for evaluating different systems rather than immediately comparing specific countries. They initially distinguish between the alternative contractual relationships used in different insurance settings and the choices available to each agent or decision maker. They then provide an overview of the alternative dimensions in which healthcare systems are commonly compared, which include the breadth of coverage, revenue generation, revenue redistribution across

 doi:10.1016/B978-0-12-375678-7.00905-6

health plans, cost control strategies, and specialized and secondary insurance.

Throughout the article, the authors use the health insurance systems of Canada, Germany, Japan, Singapore, and the USA. As shown in Table 1, insurance systems in these five countries span much of the diversity exhibited by health insurance systems around the globe. These countries include both: the most expensive system (US) and the least expensive (Singapore); single payer as well as multiple insurer; and government-sponsored and employer-sponsored insurance. Unlike many comparisons, the authors try to emphasize the general nature of the institutions used to provide care rather than the specifics of the institutional arrangements. More unified discussion of each country is reserved until after they characterize the dimensions in which healthcare systems can be compared.

The topics in this article relate to almost every other article in this Encyclopedia, but are particularly relevant for the topics of health insurance, risk adjustment, equity, demand-side incentives, and provider payment.

Agents and Choices

Agents

As summarized in Table 2, it is useful to distinguish six classes of agents in all health insurance markets. Consumers are agents who receive healthcare services, but in some systems they may have other choices to make. Providers actually provide information, goods, and services to consumers and receive payments; the article focuses on providers covered by insurance contracts. Health plans are agents who contract with and pay providers, also known in some countries as sickness funds. The sponsor in a health system serves as an intermediary between consumers and health plans, allowing for consumer contributions for insurance to differ from the ex ante expected cost of healthcare across consumers. In most countries, the sponsor is a government agency, although in the USA and Japan the sponsor for most employed workers is their employer. The key role of the sponsor in most countries is to ensure that the insurance contribution by a consumer with high expected costs (such as someone old, chronically ill, or with a large family) is not many times larger than the contribution of a consumer with low expected costs. Despite the enormous complexity of diverse intermediaries in many health insurance systems, consumers, providers, health plans, and sponsors can be viewed as the fundamental agents in every healthcare market.

Two other types of agents deserve mention. Insurers are agents that bear risk in their expenditures. In a given system, they can be identified by asking who absorbs the extra cost of care from a flu epidemic or accident. The insurer is not always a health plan as many health plans do not actually bear risk, but instead simply contract with and pay providers and pass along the expense to someone else. Insurance (or risk sharing) in a healthcare system can be shared by any of the four main agents in the healthcare system. Finally, regulators set the rules for how the healthcare and insurance market is organized, and this role can be played by sponsors (e.g., government), health plans, or providers (such as the American Medical Association in the US). Sometimes the functions of two or more agents are combined in the same agent. For example, some health plans own hospitals, and hence are simultaneously a health plan and a provider.

Systems of Paying for Healthcare

Fundamentally, there are four different ways of organizing payments and contracts in healthcare systems. Schematic diagrams of these are shown in Figure 1. System I is a private good market, in which consumers buy healthcare services directly from providers. This system is still used in all countries for nonprescription drugs and many developed countries for certain specialized goods (e.g., routine dental and eye care, and elective cosmetic surgery,) but is rare for the majority of

Table 1 Overview of health insurance systems in five countries

	Canada	*Germany*	*Japan*	*Singapore*	*USA*
Simple characterization	Single payer	Universal multipayer	Employer-sponsored insurance	Subsidized self-insurance	Employer-sponsored insurance
Primary sponsor	Government	Government	Employers	Self	Employers
Numbers of health plans	1	200	>3000	0	>1200 companies
Mandatory coverage	Yes	Yes	Yes	Yes	No

Table 2 Six classes of agents in every health insurance system

Consumers: People actually receiving healthcare, and in some countries choosing health plans or sponsors
Providers: Agents actually supplying healthcare services, such as doctors, hospitals, and pharmacies
Health plans: Agents responsible for paying and contracting with healthcare providers
Sponsors: Intermediaries between consumers and health plans who are able to redistribute the ex ante expected financial cost of health care across consumers and among health plans
Insurers: Agents who bear risk (insure), who can be any combination of the consumers, providers, health plans, or sponsors
Regulators: Agents who set the rules for agents in the health-care system

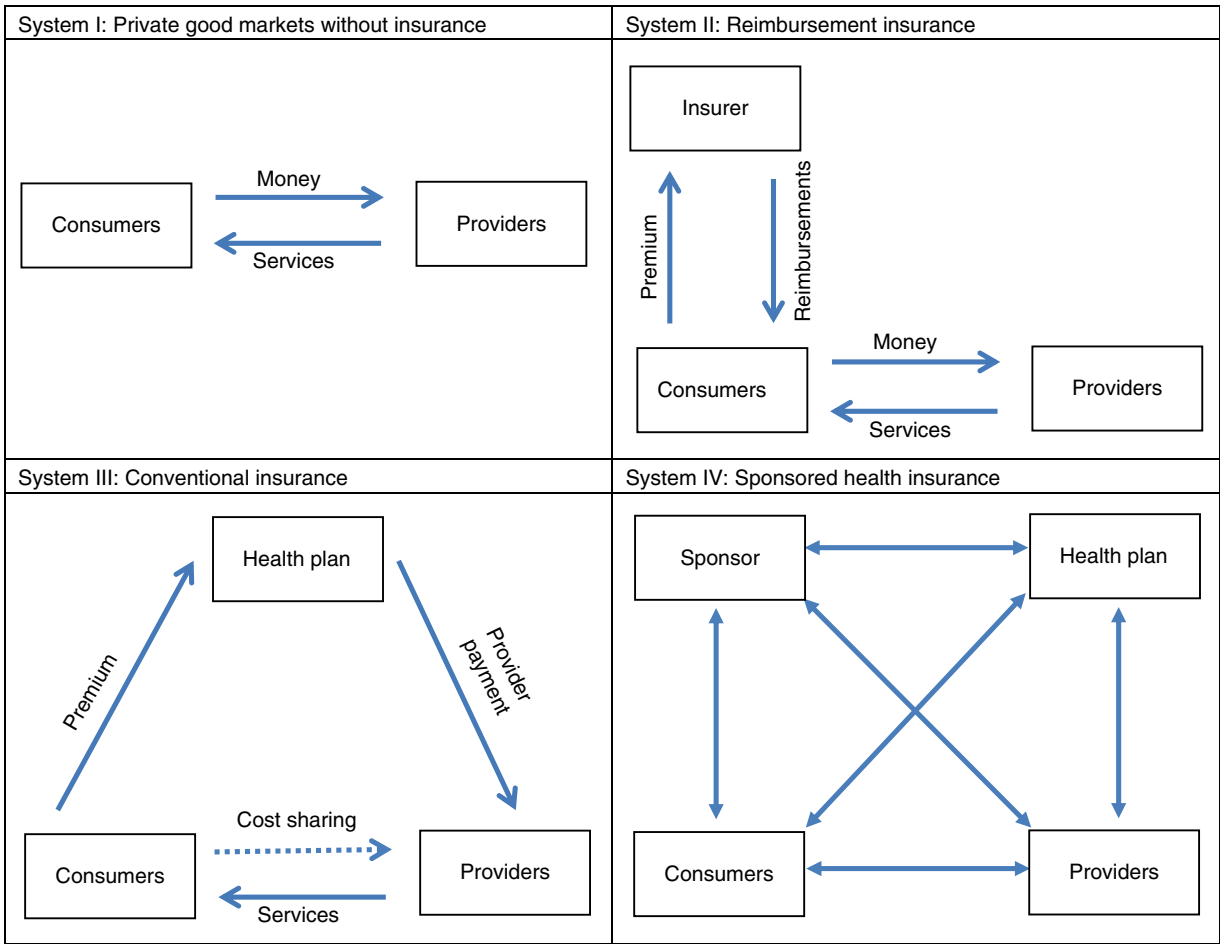


Figure 1 Four structures of healthcare payments.

healthcare services. Most consumers in Singapore and uninsured consumers in the US rely on a private good market, and pay for their healthcare when needed, without insurance.

System II is a reimbursement insurance market, in which consumers pay premiums directly to an insurer in exchange for the right to submit receipts (or 'claims') for reimbursement by the insurer for spending on healthcare. Under a reimbursement insurance system the insurers need not have any contractual relationship with healthcare providers, although the insurers will need rules for what services are covered and how generously. As will be seen, System II is the most common for secondary insurance in developed countries, and it is also widely used for automobile and home insurance.

System III is a conventional insurance market in which the consumer pays a premium to a health plan, which in turn contracts with and pays providers. Although popular in theoretical models of insurance System III is not used for the primary insurance system in any developed country, but is sometimes used for secondary insurance programs. Note the key difference in incentives between these two systems: System II incents the consumer, but not the health plan, to search for low price, high-quality providers, whereas System III does the reverse, reducing consumer incentives but enabling the health plan to negotiate over price and quality.

System IV is a sponsored insurance market in which the revenue is collected from consumers (directly or indirectly) by a sponsor who then contracts with health plans, who in turn contract with and pay providers. All developed countries that were studied involve a sponsor, although in some developing countries the sponsor may be a health plan.

Choices

Each of the line segments shown in Figure 1 reflects a contractual relationship, in which money or services are transacted. These relationships are generally carefully regulated. Countries differ in the extent to which they restrict or allow choice in each of these contractual relationships. Although many comparisons of international insurance systems do not emphasize these choices, they vary across countries significantly. Table 3 summarizes them for the five countries that are the focus of this article.

Every developed country insurance system allows consumers to choose among multiple providers, but only a few allow providers to turn down consumers, or charge fees above the plans' allowed fees (Singapore, the US). In some countries (notably the US, and legal but rare in Germany), health plans may choose which providers they want to contract with, and

Table 3 Health system choices allowed in five countries

	Canada	*Germany*	*Japan*	*Singapore*	*USA*
Consumer choice of providers	√	√	√√	√√	√√
Consumer choice of health plans		√	√	√	√
Consumer choice of sponsors			√		√√
Provider choice of consumers				√	√
Provider choice of health plan				√	√
Provider choice of sponsor					
Health plan choice of consumers				√	√
Health plan choice of providers				√	√√
Health plan choice of sponsors				√	√√
Sponsor choice of consumers			√		√
Sponsor choice of providers					ε
Sponsor choice of health plans			√√	√	√√
Simple count of system choices allowed	1	2	5	8	10

Note: √, allowed; √√, dominant; ε, allowed but minor.

providers may in turn choose the health plans they contract with (selective contracting). An especially important dimension of choice is whether the primary system has more than one health plan (Germany, Japan, the US), and how choices among health plans are regulated. In countries like the US and Japan, employers implicitly choose who to sponsor when they hire workers, and hence employers play a key role in redistributing the costs of healthcare between young and old, healthy and sick, or small and large families. In the US, consumers and their sponsors (employers) are allowed to choose not to purchase any insurance at all; some Japanese consumers ignore the mandate and do not purchase insurance, making it similar to the US. The 2010 US Affordable Care Act (ACA) will start imposing tax penalties on consumers and employers in 2014 if they do not purchase insurance, but the system will remain voluntary.

Breadth of Coverage and National Expenses

Breadth of Coverage

With the exception of the US, all developed countries have universal coverage for their own citizens through their primary insurance programs. As shown in the first row of Table 4, insurance coverage approaches 100% of the population in Canada, Germany, Japan, and Singapore, whereas only 83% of the US population has coverage. The 2010 ACA in the US will increase the percentage covered, but there is considerable uncertainty about how much coverage will increase.

Because these measures are often a focal point of international comparisons of healthcare systems, Table 4 also contrasts the dollars per capita and percentage of gross domestic product (GDP) spent on healthcare. US spending of US$8233 per capita (18% of GDP) is by far the highest, whereas Singapore's spending of US$2273 per capita (4% of GDP) is by far the lowest. In recent years, not only has the US been the most expensive, but it has also been experiencing more rapid cost growth relative to a share of its GDP (Figure 2).

Countries differ considerably in the proportion of their healthcare spending done by the public versus the private sector. This dimension is commonly a focus of international comparisons, but the proportion is not a direct choice of the country, rather it is the result of all of the other choices and regulations made in the country. Of greater interest is the percentage of spending by the primary health insurance plan or plans. This ranges from 70% in Canada to 34% in the US. Also of interest is the relative importance of the primary insurance program versus various specialty insurance programs. The US has specialized insurance programs for the elderly, the poor, children, and persons with disabilities, which collectively accounts for 56% of total healthcare spending.

Revenue

Revenue Generation

Developed countries vary significantly in how they generate revenue used to fund health plans (Table 5). In most countries, proportional or progressive taxes earmarked for healthcare are used as the primary source of revenue (e.g., Canada, Germany, Singapore, and Japan), although in some cases general tax revenues predominate. In the US and Japan, because employers are the primary sponsors, revenue comes from premiums paid by each worker. In the US, the premium is typically shared between the employer and the employee with the employer being free to choose the portion of the premium paid by the employee. State and federal tax systems partially subsidize health insurance in the US, by allowing these health insurance contributions to be exempt from income taxes, a widely discussed subsidy of health insurance and potential distortion. In Japan and Germany, premium contributions are set by law at a fixed rate, which is evenly split between employees and employers.

Revenue Redistribution

In countries with a single health plan option, there is no need for redistributing revenue between multiple health plans. However, such systems typically have to allocate budgets among different geographic areas, a similar task to reallocating

Table 4 Measures of health insurance breadth of coverage in five countries

Breadth dimensions	*Canada*	*Germany*	*Japan*	*Singapore*	*USA*
Population covered by primary insurance (%)	100	100	100	100	83
Dollars of health spending per capita	5948	4218	2878	2273	8233
GDP spending on health care (%)	11.6	11.5	9.3	4	17.9
Public healthcare expenditures (%)	70	77	80	36	56
Spending on the primary health insurance (%)	70	58	70	67	34
Specialized insurance for selected populations	No	Yes	Yes	Yes	Yes
Prevalence of secondary insurance	Common	Common	Common	Common	Common
Data from year	2012	2010	2011	2009	2010
Population in 2012 (in millions)	35	82	127	5	316

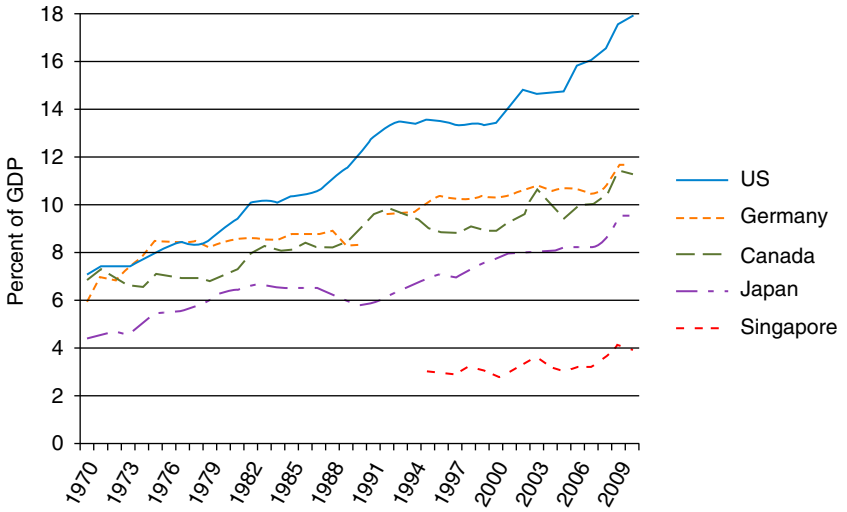


Figure 2 Healthcare spending as a percentage of GDP in five countries.

Table 5 Revenue generation and revenue redistribution in five countries

	Canada	*Germany*	*Japan*	*Singapore*	*USA*
Sources of health-care spending revenue					
Proportional payroll taxes		√√	√√		√
Progressive income taxes	√	√	√		
General tax revenue	√√	√	√	√	√
Implicit subsidies from employers		√	√√	√	√√
Fixed dollar premiums	ε	√	√	√	√√
Charitable donations		ε	ε	√	ε
Consumer out-of-pocket payments		ε	√	√√	√
Revenue redistribution: The use of risk adjustment					
Primary insurance program	√	√√	√		
Specialty insurance programs	ε		√	√	ε
Public programs	ε	ε	√	√	√

Note: √, allowed; √√, dominant; ε, allowed but minor.

money between competing health plans. In Canada, explicit risk adjustment formulas are used to allocate funds among geographic areas within each province. In systems with multiple competing health plans (i.e., Germany, Japan, the US) risk adjustment is sometimes used to redistribute money away from plans enrolling predominantly healthy enrollees and toward plans that enroll disproportionately sick or high-cost enrollees. (This topic is explored in a separate entry on risk adjustment in this Encyclopedia.) Explicit risk adjustment for this purpose is done only in Germany, where age, gender, and diagnoses are used to reallocate money among competing plans. In the German system, redistribution is done not only to adjust for health status, but also to undo unequal revenues due to the average income of health plan enrollees. This is due

to the fact that plans enrolling predominantly high-income enrollees will have greater revenues than plans with low-income enrollees, as a proportional payroll tax is used as the dominant revenue source.

Despite having multiple competing health plans, Japan and the US do not use risk adjustment to redistribute revenue, although in the US the ACA will expand the use of risk adjustment to the individual and small group markets. Risk adjustment is already used extensively in the various US public programs offered to the elderly and disabled populations and plans serving low-income and high-medical cost consumers.

HealthCare Cost Control

Although every country faces the challenge of controlling healthcare costs, countries vary significantly in their methods for doing so. Fundamentally, there are only four broad strategies for controlling healthcare costs: demand-side cost sharing, or using prices imposed on consumers to encourage them to reduce utilization; supply-side cost sharing, or using prices paid to suppliers to reduce utilization and/or reduce plan payments per unit; nonprice rationing, or setting limits on the quantity of key resources available to provide healthcare, whether done by the government sponsor or by individual health plans; and information provision that influences care provision and demand.

Table 6 summarizes the various cost control features used in the five countries that the article focuses on. It is interesting to note that Japan and the US rely extensively on demand-side cost sharing to control costs, whereas Canada and Germany rely heavily on supply side cost sharing. Singapore utilizes both. A growing number of countries have moved to bundled payment for hospital care, which originated in the US where hospital payments are based on Diagnosis Related Groups

Table 6 Cost containment in five countries

	Canada	*Germany*	*Japan*	*Singapore*	*USA*
Demand-side cost sharing					
Is it used to control costs?			√√	√√	√√
Copayment for office visits			√√	√	√√
Deductibles			√√		√√
Coinsurance				√√	√√
Coverage ceilings			√	√√	√
Stoploss					√
Tiered provider pricing					√
Supply-side cost sharing					
Is it used to control costs?	√√	√√	√	√	√
Prevalence of MD fee-for-service	√√	√√	√√	√√	√√
Use of bundled hospital payment	√	√√	√√		√√
Bundled payment for primary care	ε				ε
Salaried hospital physicians	√	√	√√	√	ε
Capitated provider groups					√
Monopsony pricing	√√	√√	√		
Government sets fee levels	√√	√	√		
Global budgets	√	√√	√		
Pay for performance bonuses			√		√
Nonprice Rationing					
Government regulation of:					
Hospital beds	√√	√	√	√√	ε
Imaging equipment	√√	√	√	√	
Numbers of doctors	√√	√	√√	√	ε
Health plan use of:					
Selective contracting		ε			√√
Utilization controls	√			√	√√
Managed care				√	√√
Gatekeepers	√√			√	√
Information					
Hospital quality measures		ε	√	√	√
Physician quality measure		ε	√		√
Health plan quality measures		ε			√
Patient satisfaction surveys			√	√	ε

Note: √, allowed; √√, dominant; ε = allowed but minor.

(DRGs). This system is now used in Germany, Japan, and many other countries. Experimentation with other forms of bundled payment, such as for primary care and multispecialty clinics, is ongoing but not yet widespread in Canada and the US.

Nonprice rationing techniques are used quite differently in the different countries. In Canada, gatekeepers and provincial-level restrictions on capacity are common. In the US, the government uses these tools very little, though many private health plans use selective contracting and some managed care plans use gatekeepers, though they are rarely mandatory. Gatekeepers are rare in Germany, Japan, and Singapore. Consumer information about hospitals, doctors, and health plans is of growing availability in the US and Japan, but rare or nonexistent elsewhere.

Specialized, Secondary, and Self-Insurance

So far the focus has been on characterizing the primary insurance mechanism used by employed adults in each country. Some countries have separate specialized insurance programs, for which only certain individuals are eligible, such as the elderly, people with a serious disability, children, low-income individuals, individuals with high medical costs, the unemployed, the self-employed, and individuals employed in small firms. In some cases, these programs cover a sizable fraction of the population and an even higher fraction of total healthcare spending. As shown in Table 7, specialized insurance programs are very common in the US and Japan. At the other extreme, Canada, with its universal, largely tax-funded system, does not need any specialized programs for subsets of its population.

In addition to specialized insurance for which only certain individuals are eligible, many countries have secondary insurance programs that reduce the cost to consumers for spending not covered by the primary insurance policy. This can be of four forms: supplementary insurance covers services not covered under the primary insurance; complementary insurance provides additional reimbursement for services not covered by the health plan; duplicate insurance provides coverage for services that are already included in the primary insurance program; and replacement insurance serves as a substitute for primary health insurance coverage. Although conceptually distinct, in some countries, a single insurance policy may have elements of all three. In Australia, for example, a single private policy may cover out-of-pocket costs for some services (complementary), cover new services (supplementary), and also allow the enrollee to opt out of using the public insurance system for a specific hospitalization or service (duplicate). Germany allows specified high-income households to purchase replacement policies instead of the primary policy.

The type of secondary insurance available in a country depends on the regulatory environment and the structure of the primary insurance mechanism. For example, replacement insurance is banned in Canada, but encouraged in the US for elderly or disabled Medicare enrollees. In countries where primary health insurance does not utilize consumer cost sharing, consumers will have no incentive to purchase complementary insurance. Almost every health insurance system will create a demand for supplementary insurance, i.e., coverage for services not covered by the primary policy. Chiropractic care, dental care, optometry, physical therapy, and pharmaceuticals are common examples of services excluded from primary insurance but often covered by supplementary insurance. Coverage for nonhospital-based prescription drug spending is in some cases covered in the primary policy (Germany, Japan, and some Canadian provinces) but not in others (many US plans, Singapore), though in Singapore there is a short list of prescription drugs that can be obtained free of charge from approved providers.

A relatively unusual alternative for insurance is self-insurance, in which consumers are required or encouraged

Table 7 Specialized insurance, secondary insurance, and self-insurance in five countries

	Canada	*Germany*	*Japan*	*Singapore*	*USA*
Specialized insurance for:					
Elderly			√√	√	√√
Disabled			√√	√	√√
Children			√		√
Low income			√√	√	√√
High medical cost				√√	√√
Unemployed		√	√		√
Self-employed		√	√		√
Secondary insurance					
Complementary insurance				√√	√
Supplementary insurance	√	√√	√	√√	√
Duplicate insurance					√√
Replacement insurance		√	√		√
Self-insurance programs					
HSAs				√√	√

Note: √, allowed; √√, dominant.

to save for their own current and future medical expenses. Self-insurance is typically encouraged through a tax-exempt health savings account (HSA). This mechanism is particularly important in Singapore, where health spending from HSAs comprises the majority of total healthcare spending. HSAs also received increased tax preference in 2003 in the US, and in 2012 were used by approximately 4% of all Americans. The institutional structure of HSAs varies between the US and Singapore, but both have a common point, in that consumers are encouraged through the tax system to put money in when young. For most consumers the account will grow over time. In some systems (Singapore), unspent money in the account can be used for other household members, or spent on education, housing, or other retirement consumption.

The attraction of self-insurance is that consumers purchase healthcare services with money that is valuable to them, and hence they have more incentive to shop around. The experience of Singapore, discussed further below, provides evidence that the savings can be substantial. However, the challenges of self-insurance are numerous. First, it presupposes that consumers can become enough well informed to shop around intelligently. This is unlikely in most countries where there is inadequate price and quality information for consumer shopping. Countries such as Canada and Germany, which do not use demand-side cost sharing, demonstrate that supply-side incentives can be equally or more effective than demand-side cost sharing. Also of concern is that self-insurance works well only for the 80% or so of the population with below average healthcare costs. Individuals with the highest healthcare costs, particularly those with chronic conditions, will tend to spend all of the money in their HSA, and be severely constrained in their ability to afford healthcare. In effect self-insurance fails these consumers when they need it most. Finally, self-insurance raises equity concerns. Studies show that wealthier households accumulate far more resources than low-income households and the tax-advantaged savings are of much lower value to low-income households. Together, both imply that most of the benefits of HSAs go to relatively healthy, higher-income households.

Country-Specific Comparisons

Canada

Canada has a universal single-payer, sponsored health insurance system called Medicare, which is administered independently by the 13 provinces and territories. Every citizen and permanent resident is automatically covered. The only choice available to consumers in the primary insurance system is a choice of providers. The only provider choice is whether to be in the dominant public system, or be an independent private provider, which is rare of most specialties. As of 2012, Canada spends approximately 11% of GDP on healthcare expenditures. Medicare provides medically necessary hospital and physician services that are free at the point of service for residents, as well as some prescription drug and long-term care subsidies. In addition to Medicare coverage, most employers offer private supplemental insurance as a benefit to attract quality employees, and a few Canadians purchase replacement insurance. Each province/territory is responsible for raising revenue, planning, regulating, and ensuring the delivery of healthcare services, although the federal government regulates certain aspects of prescription drugs and subsidizes the provinces coverage of services to vulnerable populations.

Because all services covered by primary insurance are free at the point of service, medical expenditures in this system are financed primarily through general tax revenue, or in some provinces with small income-based premiums, which together cover 70% of healthcare expenditure. Private supplementary and replacement insurance make up for the remaining 30% of medical expenditure. Employment-based supplementary insurance is the status quo among large employers and tends to cover services such as optometry, dental, and extended prescription drug coverage.

In most provinces, there are no selective contracts, hence the consumers are not limited to any particular network of providers; however, gatekeepers are often used so that consumers must obtain referrals from their family physicians to see specialists. Office-based providers are paid fees for the services. Each province/territory sets its own fee schedule. Bundled DRG payments are used to allocate funds to hospitals in a few provinces (e.g., Ontario), but this system of payments is largely invisible to patients. Whereas providers are able to charge alternative fees, the provincial insurance programs will not pay for any of the services not charged at the regulated rates. This means a provider who does not accept the government's rates must bill the patient, or the patient's secondary insurance, for the full amount of the fee. The patient will not be reimbursed by the government's insurance program for any out-of-pocket expenses. It is important to note under most provincial and territorial laws, private insurers are restricted from offering coverage for the services provided by the government's program.

Although provider shortages and long wait times to receive services push costs down, Canada is also struggling to control rising healthcare costs. The elderly population is increasing in size and it is difficult to maintain the level of benefits Canadian citizens have become accustomed to; cutting covered services is causing frictions in the country.

Germany

The German government sponsors mandatory universal insurance coverage for everyone, including temporary workers residing in Germany. Germany's primary insurance system is a social health insurance system that covers approximately 90% of the population in approximately 200 competing health plans (called Sickness Funds), with the remainder of the population (primarily high-income consumers) purchasing private replacement health insurance system. Although employers play a role in tracking plan enrollment, collecting revenue from employees and passing it along to a quasi-government agency, they are not sponsors: Insurance is not employment based in that all plans are available without regard to where a consumer works. Germany spent approximately 12% of GDP on healthcare in 2009.

Germany's health spending, excluding private insurers, is mostly funded by an income tax. This tax is a fixed portion of

income, usually 10–15%, depending on age, that is the same no matter which health plan an individual is enrolled in, and is shared equally by the employee and employer. Health plans are required to accept all applicants and pay all valid claims. Health plans are free to set premiums but due to strong competition there is almost no variation in price. Germans stop having to pay any payroll tax for healthcare at the age of 65 years even while continuing to receive healthcare benefits. Patients are also expected to pay a quarterly copayment to their primary care doctor. Collection of payroll taxes and premiums is managed by employers, although employers play no role in defining choice options and merely pass along taxes and premiums to an independent government agency. Government subsidies are provided for the unemployed or those with low income. Risk adjustment is used to reallocate funds among the competing health plans, based on age, gender, and diagnoses.

In response to the acceleration of healthcare costs, Germany has implemented various cost-cutting measures. These include accelerating the transition to electronic medical records, introducing quarterly consumer payments to primary care doctors (although visits remain free). Nonprice rationing methods are also used; for example, in order to see a specialist, patients must first be diagnosed and receive a referral from a physician who acts as a gatekeeper. Selective contracting by health plans is allowed, but rare.

The German system uses a unique point-based global budgeting system to control annual healthcare expenditures, whereby the targeted expenditures are achieved by ensuring that total payments to all providers of a given specialty are equal to the total budget for that specialty in a year. The Federal Ministry of Health sets the fee schedule that determines the relative points for every procedure in the country. Each year the total spending on a specialty in a geographic area is divided by the number of procedure 'points' from specialists in that area to calculate the price per point, and each physician in that specialty is paid according to the number of accumulated points, up to quarterly and annual salary caps.

The primary insurance coverage offered through the funds is among the most extensive in Europe, and includes doctors, dentists, chiropractors, physical therapy, prescriptions, end-of-life care, health clubs, and even spa treatment if prescribed. There are also separate mandatory accident and long-term care insurance programs. A majority of consumers also purchase supplemental coverage from private insurers, and the supplemental coverage typically provides patients with dental insurance and access to private hospitals.

Japan

Japan has a mandatory insurance system that comprises an employment-based insurance for salaried employees, and a national health insurance for the uninsured, self-insured and low income, as well as a separate insurance program for the elderly. The employment-based insurance system is the primary insurance program in which employers play a significant role as sponsors and health plans have considerable flexibility in designing their benefit features. Employment-based insurance is of two kinds, distinguished between small and large firms. Health insurers offer employer-based health insurance that provides coverage for employees of companies with more than 5 but fewer than 300 workers and covers almost 30% of the population. Large employers (an additional 30% of the population) sponsor employee coverage through a set of society-managed plans organized by industry and occupation. Employer-based health insurance coverage must include the spouse and dependents. A public national health insurance program covers those not eligible for employer-based insurance, including farmers, self-employed individuals, the unemployed, retirees, and expectant mothers, who together comprise approximately 34% of the population. Health insurance for the elderly covers and provides additional benefits to the elderly and disabled individuals. Finally, any household below the poverty line determined by the government is eligible for welfare support. Altogether Japan spends approximately 9.3% of GDP on healthcare (2011).

Health insurance expenditures in Japan are financed by payroll taxes paid jointly by employers and employees as well as by income-based premiums paid by the self-employed. Fees paid to the healthcare workers and institutions are standardized nationwide by the government according to price lists. The largest share of healthcare financing in Japan is raised by means of compulsory premiums levied on individual subscribers and employers. Premiums vary by income and ability to pay.

Employers have little freedom to alter premium levels, which range from 5.8% to 9.5% of the wage base. Premium contributions are evenly split between employees and employers. Cost-sharing includes a 20% coinsurance for hospital costs and 30% coinsurance for outpatient care. Employer-based insurance is further subdivided into society-managed plans, government-managed plans, and mutual aid associations. Patients may choose their own general practitioners and specialists and have the freedom to visit the doctor whenever they feel they need care. There is no gatekeeper system.

All hospitals and physician's offices are not-for-profit, although 80% of hospitals and 94% of physician's offices are privately operated. Japan has a relatively low rate of hospital admissions, but once hospitalized, patients tend to spend comparatively long periods of time in the hospital, notwithstanding low hospital staffing ratios. In Japan, the average hospital stay is 36 nights compared to just 6 nights in the US. This high average is likely to reflect the inclusion of long-term care stays along with normal hospital stays in the average.

Health insurance benefits designed to provide basic medical care to everyone are similar. They include ambulatory and hospital care, extended care, most dental care, and prescription drugs. Not covered are such items as abortion, cosmetic surgery, most traditional medicine (including acupuncture), certain hospital amenities, some high-tech procedures, and childbirth. Expenses that fall outside the normal boundaries of medical care are either not covered, dealt with on a case-by-case basis, or covered by a separate welfare system.

United States

The US system is at its heart an employment-based health insurance system in which employers play a key role as

sponsors of their employees. By one count, there are over 1200 private insurance companies offering health insurance in the US, which are regulated primarily by the 50 states and not at the federal level. These companies offer tens of thousands of distinct health insurance plans, each with their own premiums, lists of covered services, and cost-sharing features. In addition to this private system, there are also many overlapping public specialized insurance programs designed to cover consumers who are elderly, disabled, or suffering from end-stage renal disease (Medicare program), the poor or medically needy (Medicaid), children, veterans, and the self-employed. Because the US relies on both private and public insurance it is sometimes called a mixed insurance system. As of 2012, approximately 17% of the US population was without primary insurance, although many of these consumers are in fact eligible for Medicaid coverage but do not realize it. Altogether, the US spends nearly 18% of GDP on healthcare, the highest of any developed country.

Although the government acts as the sponsor to all of the public specialized insurance programs, employers are the key sponsor for most Americans. Choice is available to almost every agent in the US system: consumers choose providers, health plans, and sponsors; and employers, health plans and providers can generally turn down consumers who they prefer not to insure/employ, enroll, or provide services to. Employers generally contract with health plans while trying to control costs, but find little competition to hold down prices or control utilization. Many health plans negotiate fee reductions with provider groups, who tend to have substantial market power, but fees for medical care services in the US are with few exceptions the highest in the world. Although the US Medicare program sets provider fees for all regions without negotiation, all health plans must negotiate prices to be paid to providers, and the resulting fees reflect bilateral bargaining with market power.

The 2010 ACA dramatically changed many features of the US healthcare system and should greatly reduce the number of Americans who are uninsured. Starting in 2014 consumers who are without insurance will have to pay a tax penalty, and employers above a certain size will have to offer insurance to their full-time employees or pay a penalty. This US system also entails setting up insurance exchanges to cover the self-employed and small employers, who have the hardest time obtaining insurance in the US. The ACA does relatively little to address cost-containment issues, but does work toward expanding the number covered by insurance. It is unclear whether the national reform will work as well as it has in Massachusetts, where it has reduced the percentage that is uninsured to less than 2% of the population.

Cost containment is a huge issue in the US with such high spending in relation to its income. Demand-side cost sharing is used widely, with copayments, coinsurance, deductibles, coverage ceilings, and tiered payments all being used to deter demand. Many health plans use supply-side cost sharing, such as DRG bundled payments, and some are beginning to bundle primary care payment. Tiered provider payment, a form of 'Value based Insurance,' is also beginning to be used. Recent innovations include capitated provider networks, known as Accountable Care Organizations and reorganizing primary care providers to work and be paid as a Patient Centered Medical Home. Pay for performance systems and electronic medical records are other innovations being tested. It is too early to know which of these systems will be most successful in controlling costs.

Much can be written about the US public insurance programs – Medicare, Medicaid, the Children's Health Insurance Program, and The Department of Veterans Affairs – which also have their own payment systems and cost containment issues. The key point is that there is a huge amount innovation, from which other countries can learn. A positive feature of the US system is the exploration of diverse payment, nonprice, and informational programs to try to control costs. Individual-level healthcare data is more available from the US than from any of the other four countries studied here. Also, consumer information about doctors, hospitals and health plans are all available and can potentially play a role in consumer choice.

With the exception of Singapore, the US healthcare system is arguably the most unfair healthcare system, with consumers who are poor or ill with chronic illnesses paying a high share of their income for medical care. Healthcare spending is a common source of individual bankruptcy.

Singapore

Singapore has a unique-to-the-world healthcare system where the dominant form of insurance is mandatory self-insurance supported by sponsored saving, although complementary and special insurance programs are also central to their system. Remarkably, despite having a per capita GDP of approximately US$60 000 in 2011, Singapore spent a mere 3–4% of GDP on healthcare (2012). The centerpiece of its system is a mandatory income-based individual savings program, known as Medisave, that requires consumers to contribute 6–9% (based on age and up to a maximum of US$41 000 per year) of their income to an HSA. This HSA can be spent on any healthcare services a consumer wishes, including plan premiums. Funds not spent in a consumer's HSA can be carried forward to pay for future healthcare, used to pay for healthcare received by other relatives or friends, or if over the age of 65 years, cashed out to use as additional income, though there are some restrictions. A complementary insurance plan, known as Medishield, is available to cover a percentage of expenses arising from prolonged hospitalization or extended outpatient treatments for specified chronic illnesses, though it excludes consumers with congenital illnesses, severe preexisting conditions and those over the age of 85 years. As of 2011, this specialized program, which is optional, covered approximately 65% of the population. The government also supports a second catastrophic spending insurance program, known as Medifund, which exists to help consumers whose Medisave and Medishield are inadequate. The amount consumers can claim from this catastrophic insurance fund depends on their financial and social status. Singapore's system also includes a privately available, optional insurance program covering long-term care services (called Eldershield), with fixed age of entry-based payments. Consumers are automatically signed up for Eldershield once they reach the age of 40 years but they may opt out if they wish. Subsidies are available for most services, but even after the subsidies consumers must pay something out of pocket for practically all services. Some, but not all,

subsidies depend on the consumer's income, and consumers often have a choice over different levels of subsidy.

Funding for all three of the secondary insurance programs (Medishield, Medifund, and Eldershield) comes from general tax revenue. There are also five private insurance companies offering comparable plans, some of which are complementary to Medishield. Singapore has both public and private providers with the public sector providers serving the majority of inpatient, outpatient, and emergency care visits and the private sector serving the majority of primary and preventative care visits. Singapore's system receives positive publicity for its low percentage of GDP spending on healthcare but has been criticized as not replicable elsewhere. The relatively small population and high GDP per capita allows Singaporeans to avoid some of the costs associated with regulating health insurance in larger, more populous countries. Perhaps Singapore's most substantial criticism is insufficient coverage for postretirement healthcare expenses. Between potentially diminished savings and being cut off from Medishield at the age of 84 years, there is little support for financing catastrophic illnesses. Other criticisms of the country center on fairness concerns. The system favors high-income over low-income households, as they will have much greater funds contributed to their HSA. Also, consumers with high-cost chronic conditions, such as diabetes and mental illness, will repeatedly deplete their HSA and need to fall back on the various secondary insurance programs. Stigma is also an important cost containment mechanism. Finally, although consumers are incented to shop around among providers, as of 2012 there are no readily available report cards or other information sources available to guide consumers to lower cost or high-quality doctors and hospitals.

Concluding Thoughts

From the above descriptions, it is clear that there are an enormous number of ways that healthcare insurance programs vary around the world. Most country systems can be viewed as combinations or variations on the five systems described here. Although it would be wonderful if there were a way of identifying the characteristics of the most effective systems and the most equitable ones, unfortunately doing so in this article would require going beyond the boundaries of what is feasible. There are several excellent surveys of country healthcare systems, notably from the Organization of Economic Cooperation and Development and a series by the Commonwealth Fund that are excellent and are worthy of further reading.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance in Developed Countries, History of. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in the United States, History of. Health Microinsurance Programs in Developing Countries. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Long-Term Care Insurance. Priority Setting in Public Health. Private Insurance System Concerns. Rationing of Demand. Risk Adjustment as Mechanism Design. Risk Equalization and Risk Adjustment, the European Perspective. Risk Selection and Risk Adjustment. Social Health Insurance – Theory and Evidence. Supplementary Private Insurance in National Systems and the USA. Value-Based Insurance Design

Further Reading

Breyer, F., Bundorf, M. K. and Pauly, M. V. (2012). Health care spending risk, health insurance, and payments to health plans. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *Handbook of health economics*, vol. II. Amsterdam: Elsevier North-Holland.

Busse, R., Schreyögg, J. and Gericke, C. (2007). Analyzing changes in health financing arrangements in high-income countries: A comprehensive framework approach, health, nutrition and population (HNP). Discussion paper of The World Bank's Human Development Network. Washington, DC: The World Bank.

Cutler, D. M. and Zeckhauser, R. J. (2000). The anatomy of health insurance. In Cuyler, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. I, pp 563–637. Amsterdam: Elsevier North-Holland.

Davis, K., Schoen, C. and Stremikis, K. (2010). *Mirror, mirror on the wall: How the performance of the U.S. health care system compares internationally*. New York, NY: The Commonwealth Fund.

Ellis, R. P. and Fernandez, J. G. (in press). Risk selection, risk adjustment and choice: Concepts and lessons from the Americas. Boston, MA: Boston University.

European Observatory on Health Systems and Policies (2013) Health systems in transition (HIT) series. Available at: http://www.euro.who.int/en/who-we-are/partners/observatory/health-systems-in-transition-hit-series (accessed 15.04.13).

Henke, K.-D. and Schreyögg, J. (2004). *Towards sustainable health care systems – strategies in health insurance schemes in France, Germany, Japan and The Netherlands*. Geneva: International Social Security Association.

McGuire, T. G. (2012). Demand for health insurance. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *Handbook of health economics*, vol. II. Amsterdam: Elsevier North-Holland.

Meulen, R. T. and Jotterand, F. (2008). Individual responsibility and solidarity in European health care. *Journal of Medicine and Philosophy* **33**, 191–197.

Physicians for a National Health Program (2013) International Health Systems. Available at: http://www.pnhp.org/facts/international_health_systems.phppage=all (accessed 15.04.13).

Rice, N. and Smith, P. C. (2001). Ethics and geographical equity in health care. *Journal of Medical Ethics* **27**, 256–261.

Saltman, R. B., Busse, R. and Figueras, J. (2004). *Social health insurance systems in Western Europe*. Berkshire: Open University Press.

Thomson, S. and Mossialos, E. (2010). *Primary care and prescription drugs: Coverage, cost-sharing, and financial protection in six European countries*. New York, NY: The Commonwealth Fund.

Thomson, S., Osborn, R., Squires, D. and Reed, S. J. (2011). *International profiles of health care systems*. New York, NY: The Commonwealth Fund.

Van de Ven, W. P. M. M., Beck, K., Buchner, F., et al. (2003). Risk adjustment and risk selection on the sickness fund insurance market in five European countries. *Health Policy* **65**, 75–98.

Van de Ven, W. P. M. M. and Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. I, pp 755–845. Amsterdam: Elsevier North-Holland.

Relevant Websites

http://www.syndicateofhospitals.org.lb/magazine/jun2011/english/Health%20System.pdf
Syndicate of Hospitals.

http://www.ciss.org.mx/pdf/en/studies/CISS-WP-05122.pdf
The Inter-American Conference on Social Security.

http://www.kaiseredu.org/Issue-Modules/International-Health-Systems/Japan.aspx
The Kaiser Family Foundation.

Health Labor Markets in Developing Countries

M Vujicic, Health Policy Resources Center, Chicago, IL, USA

Glossary

Dual job holding The situation where health workers hold more than one job. Typically the primary job is a salaried position within the public sector and the second job is after-hours in a private clinic.

Performance-based pay A method of remuneration that aligns the incentives and rewards provided to health workers with the health outcomes-related objectives of a district or facility employer.

Shortage of health workers When employers are willing to hire more health workers, but there are no health workers available who are willing to accept employment at current wages.

Introduction

Health workers are at the center of health systems, and the health workforce plays a key role in increasing access to health services for populations in developing countries. There are numerous challenges in this critical area of health policy in developing countries. At the global level, a 2006 World Health Organization analysis found that an additional 4.3 million health workers were needed to provide basic health-care services to populations in developing countries. A more recent analysis found that for 31 countries in subSaharan Africa, there will be a needs-based shortage of 800 000 health workers by 2015 and addressing this shortage would require more than 2.5 times the projected financial resources that will be set aside for health worker salaries in these countries. Various global, national, and regional analyses have demonstrated the link between having an adequate number of health workers relative to population and achieving key health service delivery and population health targets. The evidence suggests clearly that having an inadequate number of health workers is limiting the effectiveness of health service delivery in many developing countries.

However, the availability of health workers is not the only health workforce policy challenge in developing countries. In fact, growing empirical evidence would suggest that it is not even the main issue, at least in the short term, in many settings. Geographic maldistribution of health workers is one of the most persistent and widespread issues in developing countries. A recent analysis by the World Health Organization found that one-half of the world's population lives in rural areas, but these areas are served by only 38% of the total nursing workforce and by less than a quarter of the total physicians workforce. Lack of health workers in rural areas is a major constraint to improving health service delivery. Low health worker productivity and quality limit the effectiveness of the existing health workforce. An analysis in five countries found an average health worker absenteeism rate of 35%. A recent review found that in India and Tanzania, doctors completed less than one quarter of the medically required tasks for patients presenting with tuberculosis (TB), diarrhea, or malaria. If these issues of geographic maldistribution, low productivity, and poor quality of care delivered by health workers is resolved, this could often have an immediate impact on health service delivery and population health outcomes in developing countries.

Why is it that countries with relatively similar epidemiological and disease profiles have vastly different numbers of doctors and nurses? Why are there unemployed nurses in countries that have far fewer nurses than needed to deliver basic care? Why do rural areas that often have the highest need for health services have the lowest staffing levels? Why are doctors absent in public facilities yet see patients in their private office? Why do health workers deliver care that is of lower quality than what they are trained to deliver?

As shown in this article, a labor economics perspective is extremely useful in understanding the underlying causes of these and other health workforce challenges developing countries are facing. Specifically, this article reviews the key factors that influence the demand for and supply of health workers and reviews the special features of the health labor market in developing countries. It also discusses how the labor economics perspective is extremely useful for policy makers when designing policy responses to the numerous challenges developing countries face.

A Labor Economics Perspective

A major focus of health workforce policy in developing countries historically has been to identify the 'need' for health workers of various skill sets, in various types of facilities and locations. Need can be defined as the number of health workers required to provide some mix of health services to the population. Need is a completely normative concept and takes into consideration only the epidemiological profile of a population, the preferences of policy makers over disease priorities, and technology considerations such as optimal skill mix, models of care delivery, and the expected productivity of health workers. Determining the need for health workers involves a great deal of priority setting among policy makers, but no economic factors such as prices or budgets enter into the needs discussion. There have been numerous studies that focus on identifying needs-based staffing levels. The World Health Organization estimates that worldwide greater than 4 million additional health workers are needed to deliver basic health services to the population. In Ethiopia an analysis

estimated that 36% more physicians are required in order to expand antiretroviral treatment to target level. To scale up 42 priority health services, Tanzania is estimated to require greater than 100 000 full-time health workers by 2015, compared to a projected availability of less than 30 000.

Demand for health workers is defined according to the standard definition from labor economics: the total amount of labor, or in the simplest sense the total number of health workers, employers are willing to hire at current wages, holding constant other important variables such as health worker productivity, household income levels, political considerations, and government budget levels. The key distinction between need and demand is that many factors other than the health status of the population influence the demand for health workers. In other words, financial, economic, and political factors can be thought of as driving a wedge between the demand and the need for health workers. More importantly, the demand for health workers – and not the need – drives hiring behavior and, as a result, policy makers need to understand employer behavior in order to influence hiring decisions.

In the health sector, particularly in developing countries, there is a diverse set of employers of health workers. The main ones include global nonprofit employers (e.g., multilateral organizations that directly employ health workers), a country's public sector (e.g., national, state, or local government directly or a government-owned hospital), a country's for-profit sector (e.g., for-profit clinics), the nonprofit sector (e.g., mission clinics), and individuals (e.g., sick people who seek care from health workers and pay for their services out of pocket). The factors that influence the demand for health workers within each employer category are different. For the global nonprofit employers, a key factor is the level of resources various bilateral and multilateral agencies provide for health initiatives. For example, the large increases in donor resources for human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) in recent years led to a sharp increase in the demand for health workers who provide HIV care.

The public sector is a significant employer of health workers in both developed and developing countries. Within the public sector, hiring decisions are often influenced as much by political, macroeconomic, and social factors as by the needs of the population. In settings where health workers are employed as civil servants, the demand for health workers is influenced heavily by the total wage bill allocated to the health sector, which, in turn, is often a highly politicized process dependent on macroeconomic and fiscal policy priorities. In developing countries there tend to be constraints, for very sound reasons, on how fast the overall wage bill can expand. For various reasons, these overall wage bill constraints often, but not always, restrict the demand for health workers in the public sector. For example, in Kenya in the mid-2000s the overall wage bill policy of the government limited the Ministry of Health's ability to hire health workers and expand service delivery, leading to high health worker unemployment rates. Even in decentralized settings, fiscal transfers to subnational governments (or even to facilities) are only very loosely based on the health-care needs of populations. As a result, the demand for health workers in the public sector often fluctuates with government (or facility) fiscal conditions, and this has been well documented empirically in both developed and developing countries. As noted in the Introduction section, the fact that addressing the needs-based shortage of health workers in subSaharan Africa would require more than 2.5 times the projected financial resources set aside for health worker salaries helps explain why there are so few health workers (relative to need).

The nonprofit sector operates similar to the public sector, except that specific agencies will focus on particular diseases, populations, or geographic areas. In developing countries this is important because the nonprofit sector is often a major employer of health workers, especially in very poor countries. Moreover, specific to developing countries, if significant levels of donor assistance for health are channeled through nonprofit organizations with little coordination with the government, this further increases the demand for health workers within the nonprofit sector.

In the for-profit sector, the demand for health workers is driven by profit maximization. Among individuals, the demand for health workers is influenced by the demand for health-care services, which is driven by a person's health status, ability to pay, and other factors. In many developing countries, the individual-level market for health services is large, and as a result, individuals and households are a significant source of direct demand for health workers. In a sample of 15 countries in subSaharan Africa, for example, out-of-pocket payments accounted for a low of 6% of the total health spending (Namibia) to a high of 62% (Chad).

The supply of labor in the health sector can be defined as the total amount of labor, or in the simplest sense the number of health workers, willing to work at current wages, holding constant other important variables like working conditions. A more refined definition could incorporate various aspects of effort, including productivity (e.g., hours worked, number of patients treated) and quality (e.g., care provided according to treatment guidelines). It is important to highlight several key decisions that influence the supply of health. These include the migration decision (whether to stay in the country), the labor force participation decision (whether to work or not), and the health-care labor force participation decision (conditional on working, i.e., whether to work in the health sector or in some other field).

Migration of health workers is an important issue in many countries, but especially in developing countries. As much as 70% of the medical workforce in subSaharan African countries eventually migrate. High rates of migration are also found in other regions. Many view migration as the single biggest challenge to strengthening health systems in developing countries. The health-care labor force participation decision is often overlooked by policy makers, yet has important implications. Several studies have shown that even small changes in the health-care labor force participation rate have important effects on the supply of health workers.

Migration and labor force participation decisions determine the supply of health workers within a country. Beyond that, the internal migration decision (which geographic area to work in), the sector decision (whether to work in the public or private sector), and the 'effort' decision (productivity and quality) influence the supply of labor in various settings of interest (e.g., a rural public clinic). When delineated this way,

it is clear that there are intervention points to influence the supply of health workers that go well beyond simply adjusting enrollment levels within education institutions. Too often in developing countries, policy makers overlook several of these critical labor supply decisions.

Just as with the demand for labor, a host of factors unrelated to the health-care needs of the population influence the labor supply decisions of health workers. Migration decisions are influenced by relative wages, working conditions, individual and family characteristics, and preferences. Labor force participation of health workers depends on factors such as wages and working conditions and family income, and the health-care labor force participation decision is influenced by the wages and working conditions of jobs in the health sector compared to relevant jobs outside the health sector. All of these labor supply decisions are also heavily influenced by demographic characteristics such as age, gender, family size, and parental education levels.

It is clear that both the demand for and supply of health workers in developing countries are influenced by a complex set of factors that are unrelated – or at best loosely related – to the health-care needs of the population. This has an extremely important implication: the health labor market – even at market-clearing employment and wage levels – will not necessarily generate health worker employment outcomes that meet the needs of the population. From the policy maker perspective, this provides the rationale for intervening in the health labor market to influence demand, or supply, or both, to move employment outcomes closer to those that promote society's goals with respect to health outcomes.

The Developing Country Context

Several aspects of the health labor market that are, if not unique, at least particularly relevant in the developing country context are worth discussing.

Remuneration is Highly Regulated

In settings where health workers are employed as civil servants, remuneration levels are highly regulated and must be set within civil service regulations. Market forces do not exert a strong influence on health worker remuneration in such settings. Health worker salaries are rarely adjusted in response to actual or projected shortages or surpluses. Rather, they are set relative to other occupations (e.g., teachers) and relative to historical levels. For example, instead of being a function of market conditions, wages for one level of nurse are often set relative to a higher or lower level nurse or relative to another civil service worker with the same number of years of training and experience, such as a teacher. The empirical evidence suggests that remuneration regulations in developing countries – for both legal and political reasons – constrain health worker remuneration changes. In settings where health worker remuneration has been decentralized or removed from the overall civil service, there is much more autonomy for facilities to adjust remuneration in response to market signals.

Salary Is a Dominant Remuneration Method

The way doctors and nurses are paid can provide strong incentives for improving health worker productivity and quality of care. In many low-income countries, health workers in the public sector receive most of their compensation in the form of a salary. Along with weaknesses in governance, this is an important factor contributing to the significant level of health workforce absenteeism and low productivity many developing countries experience. Alternative types of payment mechanisms have the potential to provide stronger incentives to health workers and thereby improve performance and efficiency. Developed countries have a long history of alternative payment mechanisms, including fee-for-service, capitation, and performance-based pay, but only recently have developing countries experimented with innovative compensation policies. The benefit of performance-based pay is that it aligns the incentives and rewards to health workers with the particular objectives of the district or facility where health workers are employed, and several empirical studies have demonstrated this.

Remuneration is Fragmented

Allowances are often a significant component of health worker remuneration. For instance, allowances account for 45% of the overall health wage bill in Kenya and 14% of the overall wage bill in the Dominican Republic. However, allowances are often fragmented and are not used strategically. In Kenya, for instance, the more lucrative housing allowance that accrues to doctors in the Nairobi area has created a disincentive to locate in remote areas. In the Dominican Republic, after a health worker leaves a location, the geographic allowance he or she was receiving turns into a permanent component of the worker's wage. The allowance structure in many developing countries is often not designed to raise remuneration levels in less desirable work settings relative to remuneration in desirable settings. These compensating differentials in remuneration are necessary to recruit health workers to less desirable settings such as rural areas.

Donor Assistance for Health Is Significant

In many developing countries, particularly in subSaharan Africa, there are significant external resources devoted to investments in the health workforce. For example, approximately one-fifth of the UK's support for the health sector in developing countries is channeled to health workforce activities. Although most of these health workforce resources are used to finance in-service training of health workers, agencies such as the Global Fund to Fight AIDS, TB and Malaria and the Global Alliance for Vaccines and Immunization devote significant resources toward health worker remuneration. When there are significant levels of donor assistance for health that are not fully coordinated with the government (through a national health strategy), this poses a challenge for health workforce policy. Nongovernment organizations and other nonprofit organizations are not subject to the same regulations as the government and, as a result,

offer terms of employment that are very different than what is available to health workers in the public sector. This can generate significant movements of health workers across different sectors and can influence greatly the allocation of health workers across various priority programs.

There Are Administrative Inefficiencies in Key Management Functions

Owing to various reasons, including a centralized hiring process, the recruitment process in developing countries is often subject to significant delay and is not targeted to areas with the highest need for staff. For example, in Kenya in the late-2000s, it took an average of 10 months to fill a vacant position once a suitable candidate was found. With reforms to the hiring process, this has recently been reduced to an average of 3 months. In many developing countries, salaries follow people rather than remaining tied to a particular position. In other words, when health workers transfer or move, they often retain their remuneration level. This poses a significant challenge in that it limits the extent to which remuneration can be linked to a specific position (rather than person) and, therefore, the ability of policy makers to generate compensating differentials to attract health workers to less desirable settings. Decentralization, under certain conditions, has the potential to significantly reduce many of these inefficiencies in administrative procedures. For example, when Rwanda devolved remuneration authority to the local level, facilities were able to adjust payment levels to attract health workers to some of the hardest-to-fill positions.

Dual Job Holding Is Extremely Common

In developing countries, health workers often hold more than one job. For example, more than half of doctors in South Africa have additional employment outside of their primary practice. Often, the primary job is a salaried position within the public sector and the second job is after-hours in a private clinic. Although some governments explicitly allow dual job holding through part-time contracts (e.g., Dominican Republic), it is often poorly monitored and regulated. The challenge that dual job holding poses is that it limits the influence policy makers have on total remuneration and, therefore, the incentive structure health workers face within the entire health-care system. Vietnam is a useful illustrative example. In Vietnam, salaries of physicians working in the public sector are set according to Ministry of Health policy and are deliberately set higher in rural areas in order to make rural postings more attractive. This is a sound strategy on the part of the Ministry of Health. However, when all sources of income are taken into consideration, including earnings from dual job holding, the total remuneration in urban areas ends up being much higher than in rural areas. In fact, the effective hourly wage in the second job (in the private sector) is almost double the primary job in the public sector. As a result of dual job holding, there is a considerable earnings disadvantage to locating in rural areas of Vietnam that the Ministry of Health's salary structure did little to reverse.

Using a Labor Economics Perspectives to Guide Policy

The labor economics perspective suggests that to design effective health workforce policies, it is important to understand the overall labor market conditions in the health sector – namely, is the current employment level demand constrained, supply constrained, or at or near equilibrium? For example, when there are surpluses (i.e., few unfilled vacancies and unemployed health workers), it is necessary to stimulate demand in order to increase employment levels. In the public sector, this might be done through lowering wages or increasing resources available for hiring health workers. Negotiating lower wages in the public sector is difficult politically for the various reasons mentioned, but effective wages can be lowered through skill substitution (e.g., shifting tasks away from physicians toward nurses) or contracting with private agencies where total labor costs might be lower. Increasing the level of resources for salaries can be achieved through direct increases in Ministry of Health salary budgets or increased block transfers to districts or facilities (in a decentralized setting). Each strategy has its associated challenges, but there are several examples of countries that have successfully implemented these policies. Reducing the price of health services to households is also an effective way to stimulate demand for health-care and, therefore, for health workers. This can be achieved through reducing or removing user fees or other financial barriers to care. However, policies that aim to increase the supply of health workers are much less appropriate when there are labor surpluses. Increasing the number of graduates, for example, will likely increase health worker unemployment rates when employment levels are demand constrained.

When there are shortages of health workers (i.e., there are unfilled vacancies), a different set of policy options is required in order to change employment levels. In this case, the supply of health workers needs to be targeted. One option is to expand training capacity to increase the number of health workers, provided that graduates remain in the country. Higher wages, improved working conditions, and better continuing education opportunities are some of the interventions that will make jobs more attractive to health workers. Although wages tend to receive the most attention, evidence has shown that improving other job characteristics is often a more cost-effective way to attract workers to vacant posts.

Labor economics also offers some specific quantitative and qualitative analytic tools that can help generate empirical evidence to guide health workforce policy on specific issues. For example, qualitative analysis can be used to identify the critical job characteristics that influence health worker decisions to locate in rural areas and, more broadly, factors that influence health worker motivation and performance. A technique known as discrete choice analysis, in which potential workers are asked to rank jobs with different attributes (including, e.g., wage, location, and training) can be used to quantify the expected impact of alternative policies aimed at recruiting health workers to rural areas. The Government of Liberia recently implemented a rural area incentive program for nurses that directly incorporates findings from a discrete choice analysis. Labor force surveys can be used to

measure current health worker remuneration differentials between different levels of care, specialties, and geographic areas, and the remunerations differentials that would be necessary to entice health workers to change job locations.

Through a better understanding of the underlying behavior of health workers and those that employ them, and how they interact in the health labor market, policy makers can more effectively design health workforce policies. The labor economics paradigm can be an important tool to help address the many health workforce challenges in developing countries and, ultimately, to improve the health of the population.

See also: Dentistry, Economics of. Market for Professional Nurses in the US. Physician Labor Supply. Physician Market

Further Reading

Anand, S. and Barnighausen, T. (2004). Human resources and health outcomes: Cross country econometric study. *Lancet* **364**, 1603–1609.

Buerhaus, P. (2008). Current and future state of the US nursing workforce. *Journal of the American Medical Association* **300**(20), 2422–2424.

Buerhaus, P., Auerbach, D., Staiger, D. (2009). The recent surge in nurse employment: Causes and implications. *Health Affairs* w657–w668.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K. and Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives* **20**(1), 91–116.

Dussault, G. and Vujicic, M. (2009). The demand and supply of human resources for health. In Carrin, G., Buse, K., Heggenhougen, K. and Quah, S. (eds.) *Health systems policy, finance, and organization*, pp 296–303. New York: Elsevier.

Lagarde, M. and Blaauw, D. (2009). A review of the application and contribution of discrete choice experiments to inform human resources policy interventions. *Human Resources for Health* **7**(62), 1–10.

McCoy, D., Bennett, S., Pond, B., et al. (2008). Salaries and incomes of health workers in sub-Saharan Africa. *Lancet* **371**, 675–681.

Serneels, P., Lindelow, M. and Lievens, T. (2008). Qualitative research to inform quantitative analysis: Health workers' absenteeism in two countries. In Amin, S., Das, J. and Goldstein, M. (eds.) *Are you being served? New tools for measuring service delivery*, pp 271–298. Washington, DC: The World Bank.

Vujicic, M. and Zurn, P. (2006). The dynamics of the health labor market. *International Journal of Health Planning and Management* **21**(2), 1–15.

Vujicic, M., Zurn, P., Diallo, K., Adams, O. and Dal Poz, M. (2004). The role of wages in the migration of health care professionals from developing countries. *Human Resources for Health* **2**(3), 1–14.

Relevant Websites

http://www.worldbank.org/hrh/
The World Bank.

http://www.who.int/hrh/en/;
http://www.who.int/whr/2006/en/
World Health Organization.

Health Microinsurance Programs in Developing Countries

DM Dror, Micro Insurance Academy, New Delhi, India, and Erasmus University Rotterdam, Rotterdam, The Netherlands

What is Microinsurance?

Microinsurance does not have a single accepted definition. However, two well-known sources provide high-level definitions and describe salient traits that help establish what microinsurance is and what it is not. These are introduced in this section and used throughout this article to anchor the discussion.

1. Dror and Jacquier's seminal work coined the expression 'microinsurance' and defined it as voluntary, group-based, self-help insurance schemes for which the group designs the premium, benefits, and/or claims to be attractive, relevant, and affordable to excluded populations in the informal sector. This definition departs from classical demand-driven market theory which views the individual as formulating demand, whereas here the group takes that role, and group demand reflects its aptitude to pool both risks and resources in order to provide protection to all members. This definition can be viewed as applying the subsidiarity principle (that decisions should be taken at the lowest level where they can be taken). Hence, both the governance and the utility are mutually determined by those most concerned. And, because microinsurance is typically targeted at low income, poor people (even though this is not a necessary trait), it manifests atypical pooling and risk transfer.
2. The International Association of Insurance Supervisors (IAIS) defines microinsurance as insurance for low-income people provided by a variety of institutions, run in accordance with generally accepted Insurance Core Principles, and funded by premiums proportionate to the likelihood and cost of the risk involved. Microinsurance serves populations in the informal sector that are excluded from or not served by other insurance.

These definitions have much in common:

- microinsurance is insurance (as distinct from savings and credit) and applies principles of risk pooling;
- coverage is always contributory (i.e., schemes that are 100% subsidized would not qualify as 'microinsurance');
- microinsurance is independent of the size of the risk-carrier (can be a local, informal mutual-aid society, or a large national or multinational insurance company);
- microinsurance is independent of the scope of the risk (risks do not become 'micro' when coverage is partial or the insured that experience them are poor);
- microinsurance is independent of the delivery channel (the most common options are small community-based schemes, credit unions, microfinance institutions, or local agencies);
- microinsurance is independent of the class of risk (life, health, crop, livestock, assets, etc.);
- microinsurance targets people in the informal sector;
- microinsurance is suited to people on low incomes (in the second definition this is a defining trait);
- affiliation to microinsurance is voluntary, the first definition makes this explicit and it is generally consistent with the tenor of the second definition.

Although both definitions identify that microinsurance suits poor people with low incomes, and it is an intuitively appealing place to start a definition, including it as a defining trait in the second definition raises operational problems. Measuring low income is complex especially when accounting for the depth of poverty, length of time of being low income, the phase of life (e.g., childhood vs. mid-life), and the comparative deprivation level by reference to the society in which a person lives. Thus, insurers rarely have the knowledge or the motivation to synthesize such complex information that is per se not relevant for underwriting risks. Furthermore, it is also not simple to determine whether premiums are proportionate to the likelihood and cost of the risk involved without specifying the method of premium pricing; this information is not revealed by habitual insurance performance indicators.

There is at least one fundamental difference between the two definitions, vis the role of the group. Because this radically fundamental distinguishes the definitions, the author explores it further.

Communities might variously be area-based, trade-based, faith-based, gender-based, cause-based, ethnicity-based, and other. The core assumption underlying the first definition is that the group is the framework within which cultural, demographic, and general economic factors are shaped in an otherwise unstructured 'informal' environment. The community relies on profound information that is not known outside the community and may have a different logic to that on which commercial insurance decisions are based. Without this group engagement, the market for insurance continues to struggle to establish viable supply and solvent demand for insurance.

Those that accept the second definition believe that commercial or other service providers have the capacity to establish viable supply for which demand can be assumed to exist.

Under both definitions, it is clear that if the scheme is not customized to be relevant to the specific context and needs of the community, it cannot be classified as microinsurance. This has erroneously been taken to mean that national microinsurance programs are not possible. If national programs can be tailored to local needs, they can be described as microinsurance.

How Common is Microinsurance for Health?

Recent overviews of microinsurance activity in poor nations have shown a low penetration rate with some 3% of the poor

 doi:10.1016/B978-0-12-375678-7.00922-6

in the world's 100 poorest nations having some microinsurance in 2007, and only 0.3% having health microinsurance. These figures relate to insurance more consistent with the second definition than the first. Of the 78 million people covered by any microinsurance, only 3.2 million are covered by mutual or community-based organizations that would clearly fit the first definition. Although it is not clear what proportion of these 3.2 million have health microinsurance, it is manifestly apparent that health microinsurance consistent with the author's definition has not reached many of the at least 2.5 billion people who need it. Although a raft of barriers to the growth of microinsurance have been identified, the most fundamental include: poor distribution channels and poor business infrastructure; a history of mandated credit life insurance that builds antagonism among consumers; a prevalence of commercial microinsurance schemes that, although compliant with government requirements, often provide no benefits to the poor; ill-fated attempts to build universal coverage of health insurance where there are neither funds for such a scheme nor adequate available health services. Moreover, regulation remains poor with microinsurance sometimes being ignored by policy and sometimes included without distinction, and the best practice is yet to be identified.

Stated simply, the market (both supply and demand) for health microinsurance remains small or even nonexistent in most settings. If insurance were simply a risk-transfer tool, health microinsurance would theoretically be attractive to low-income populations who are exposed to many and various health risks. Moreover, Nyman's game-changing assertion that health insurance is demanded because it provides an income transfer from the well insured to the sick insured irrespective of risk management, suggests that microinsurance for health would be particularly attractive among the poor who practice reciprocity (rather than state-mandated cross subsidization or solidarity) for their burden of disease. And, notwithstanding the Prospect Theory's challenge to the validity of risk avoidance as a rational motive for insurance, there remains widespread acceptance that the market for insurance in general and health insurance in particular is a market for risk avoidance.

Yet, the typical situation observable everywhere is a dearth of both supply and demand for health insurance at the informal sector in low-income countries, which the private/commercial sector and government are ill-suited to resolve because the economic and behavioral choices in the informal economy differ fundamentally from those prevailing in rich and orderly/rule-based economies. These behaviors and decisions are often regulated or shaped by community interests in ways that may be inconsistent with the expected behavior of a single individual economic actor.

Given that poor excluded populations have great health risk and are in need of income transfer when ill and that they live in communities that can give structure to both the supply and demand side, why then is community-based microinsurance for health not more prevalent? The reason is that successful insurance, even at the community level, requires technical and actuarial knowledge as well as advanced financial literacy, which are sorely missing in the informal sector. Therefore, these barriers to success cannot typically be overcome without external drivers to develop capacity and drive institutional change that will enable markets to establish. In this sense, the lack of a market for micro health insurance is a failure of context rather than market failure.

Typology of Microinsurance Business Models

At least four basic operating models to deliver health microinsurance have been described as 'microinsurance':

1. The partner-agent model in which the role of the insurance company ('the partner') includes designing, pricing and underwriting of products, and responsibility for scheme solvency in the long-term. Distribution/marketing, premium collection, and product servicing are usually delegated to an intermediary ('the agent'), often a person or a for-profit legal entity. Insurance companies usually pay agents a commission on premiums sold. This remuneration method is effective in urban settings and among solvent populations, but less so in the informal sector, where reaching persons may cost much time, and lead to few closed sales. This is why, in Africa and in Asia, insurers have been keen to assign the agency role to bodies that interact frequently with rural and low-income populations, such as nongovernmental organizations (NGOs) or Micro Finance Institutions (MFIs). Where MFIs have identified a need for health insurance (e.g., when illness caused default on repayment of debts), they have sometimes approached insurers to design a suitable insurance product, and suggested the price range that would be acceptable, and pressured providers for better services and claim settlement.

   The partner-agent model could qualify as 'microinsurance' under the second definition if it offered insurance to low-income people, but may not qualify under the first definition as the design decisions are taken by insurers, not by the community. That said, when MFIs acting as agents also involve the community in the bidding process and in priority setting, one could argue that this is a borderline situation that could also meet the requirement of the first definition.
2. The provider-driven model, in which clients pay premiums to the healthcare provider (e.g., hospital, physician), which in turn enables them to consume health services without having to pay out of pocket at the point and time of service. The healthcare provider benefits from this arrangement by creating larger solvent demand for health services, sold mainly by the provider-insurer; and increasing and smoothing the cash flow as it is dissociated from incidence of illness.

   The healthcare providers are responsible for designing, pricing, and underwriting insurance products and for the long-term sustainability of insurance operations.

   The provider-driven model could qualify as microinsurance under the second definition if the client-base of the insurance is composed of poor people. However, it is unclear whether low-premium policies that include rare and expensive surgical procedures would still qualify as 'microinsurance.' Under the first definition, this arrangement would not qualify as 'microinsurance' as the

decisions on pricing, package design, and claims settlement are taken by the provider-insurer according to its commercial interests and capacity to provide, possibly without inputs from, or participation of the community in governance of the insurance. Although there are examples of the healthcare provider investigating need for and willingness to pay (WTP) for health insurance, in all cases the role of the community was limited to passive informant rather than meaningful engagement in decision making.

3. Charitable insurance model (a.k.a. 'full-service' model), in which an external charitable organization, acting as 'insurer,' assumes responsibility for the long-term sustainability of the scheme by supplementing the payment of premiums, because there is an assumption that contributions could never cover all costs of benefits provided. Many charitable insurers are run by NGOs, many are operating not-for-profit, and many may view the insurance as a suitable vehicle to promote their main development, or religious goals. The external donor retains much of the responsibility for product design, pricing, and administering the scheme, in ways that would align with the fundamental objectives of the organization. Thus, there are instances where the charitable insurer fixes premiums below the actuarial cost, or does not enforce the requirement that only paid-up insured can access benefits.

   The charitable insurance model could qualify as 'microinsurance' under the second definition when the financial arrangement protects low-income people against specific perils in exchange for regular premium payments proportionate to the likelihood and cost of the risk involved; it would not qualify as 'microinsurance' when the payment of premium is irregular, and/or when that premium is disproportionate to the risks involved. Under the first definition, charitable insurance would qualify as 'microinsurance' when the community of beneficiaries participates in key decisions and in governance of the scheme, and would not qualify as 'microinsurance' otherwise.

4. Mutual/cooperative insurance model, in which the insured is also the insurer, so that each member of the mutual (or cooperative), together with all other members, is simultaneously benefitting and underwriting at least part of the risk. The community of members is thus responsible for all aspects of the scheme including designing, pricing, and underwriting products and for the long-term solvency of the insurance. The mutual model finds its origins in nineteenth-century Europe, and has been launched, designed, implemented, and administered by and for groups of people without access to the resources and financial techniques of commercial insurance. Being directly in contact with its membership, this insurer can 'disintermediate' the agent role and save agent commissions. As the interests of mutual insurers are identical to those of its members, the first priority is to establish a good fit between the needs of members and the benefit package. Many mutual societies are not only insurance providers, as they function as broader mutual-interest organizations. Some mutual organizations have grown to be very large and have professional management, which can distance the operations from the members, resulting in less social cohesion in large mutuals than in community-based schemes.

   The mutual/cooperative insurance model could qualify as 'microinsurance' under the first and second definitions, provided that the insurance covers low-income people, and the community of beneficiaries participates in key decisions and in governance of the scheme. It is in fact the only model that could satisfy both definitions of 'microinsurance' (Table 1).

In reality, any health microinsurance scheme can have features of multiple models. For example, in Uganda, there are several mission hospitals that run provider-driven insurance schemes, yet they are heavily subsidized, and thus similar to a charitable scheme. Moreover, schemes can start as one model and change over time, as did the Yeshasvini Trust in India, which was initially founded by healthcare providers, and is currently run as a not-for-profit charitable model. The providers largely designed the current benefit package and the trust developed as a mixture between a charitable model and provider-driven model.

Insurance Failures under Microinsurance

There is no principal difference between health microinsurance and any other health insurance in terms of exposure to insurance failures, but there are, or can be, significant differences to exposure under the different business models. The phenomena usually considered as 'insurance failures' include adverse selection, cream skimming, moral hazard, free riding, and fraud.

Adverse selection describes the situation in which an insurer accepts offers of insurance of high-risk persons at rates that do not reflect the actuarial premium attached to their risk class because the insurer does not have full information about the risk that these individuals represent. In response, insurers increase premiums to fund higher costs, which lead to lower participation of people with lower risks, and could even lead to exit of higher risks due to unaffordable premiums. Adverse

Table 1 The fit between the definition of microinsurance and microinsurance business models

Type of business model	*First definition (Dror and Jacquier)*	*Second definition (International Association of Insurance Supervisors)*
Partner-agent model	No, unless motivated to include community decision making	Yes, if client is poor
Provider-driven insurance	No, unless motivated to include community decision making	Yes, if client is poor
Charitable insurance	No, unless motivated to include community decision making	Yes, if client is poor and they pay a premium proportional to risk
Mutual/Cooperative insurance	Yes	Yes

selection is more likely to occur when affiliation is based on individual contracts and is voluntary, and is least likely when affiliation is mandated for a large group. The effective countermeasure to adverse selection is 'en bloc affiliation' of an entire community, even when this is not obligatory.

In the context of health microinsurance, adverse selection is more likely to occur under partner-agent, provider-driven, and charitable models, when they allow voluntary and individual affiliation. En bloc affiliation occurs often under the mutual/cooperative model, and sometimes under the partner-agent model when the agent is a strong NGO that can in fact affiliate an entire community.

Cream skimming, also known as 'preferred risk selection' or 'cherry-picking' (and when it takes the form of nonrenewal it is called 'lemon-dropping'), occurs when an insurer selects only part of a large heterogeneous group which the insurer estimates as being lower-than-average risk (the preferred risks) without discounting the risk-rated premium they are required to pay. The purpose of cream skimming is to enable the insurer to retain profits by reducing the loss ratio.

In the context of health microinsurance, cream skimming occurs when the standard contract includes certain limitations by age (e.g., the insurance is valid only from age 3 to age 60), or by health status (e.g., excluding certain illnesses, chronic conditions etc.), or by benefit types (e.g., cover only a limited list of procedures requiring hospitalizations etc.). Clearly, exposure to this situation is more likely to prevail when the underwriter has free hand in determining the terms of the policy (this is frequently the case in the partner-agent model or the provider-driven insurance) and is less likely to occur when the insured can influence the terms of the policy (as in the mutual/cooperative model).

Moral hazard is the increase in healthcare utilization that occurs when a person becomes insured, which is an insurance failure because insurers pay out more in benefits than was expected when setting premiums. The conventional interpretation is that this additional healthcare utilization represents a loss to other insured people, as they ultimately bear the cost of additional demand. Nyman pointed out that this interpretation is based on the assumption that the extra healthcare is not clinically needed (e.g., cosmetic surgery), but where the extra demand is needed the extra care is a gain to society. Although increased utilization patterns cannot automatically be considered as bad outcomes (as suggested by the term 'hazard'), they are insurance failures when payout exceed actuarial estimates. The conventional remedy for moral hazard is to require consumers to pay part of the cost (i.e., copay) in every case, or that the insurer can legitimately control utilization.

In the context of health microinsurance, given that the target populations chronically underutilize health services, it would be reasonable to expect that health insurance would lead to increased utilization of services covered by the particular health microinsurance scheme. This theoretic welfare gain is borne out by empirical research in India and the Philippines, which indicate a transition from underutilization to normal utilization, as the income transfer overcomes the financial constraints on accessing healthcare. Under the mutual-aid model, where community of insured is simultaneously also the underwriter, the community has very good information on its members when the group size is small. Thus, it can exercise peer pressure to reduce moral hazard.

Moral hazard can also be induced by providers of care that benefit from overtreating insured persons. 'Supply-induced moral hazard'is not easily detected or limited in the context of health microinsurance, but low insurance caps obviously limit not just the cover but also the margin of providers to generate overconsumption. Under the provider-driven insurance model, the provider can exercise better control of supplier-induced moral hazard, but at the same time the provider might be in a situation of conflict of interest to do so.

Free riding arises when a person benefits from the health insurance scheme without paying premiums. This risk is due to imperfect monitoring of those drawing benefits; cashless delivery of services can increase the risk of free riding. The countermeasure for this is to improve monitoring of the system with the view to ensuring that only legitimate beneficiaries will draw benefits. Smart cards and similar electronic devices are increasingly popular aids to monitoring.

In the context of health microinsurance, where there is very little access to IT and where online/real-time access to an management information system (MIS) is rare, it may not be feasible to reduce free riding through automated checks of claims. Rather, the remedy to free riding would consist of creating a counteracting interest to disallow or adjust payments. The mutual/cooperative model has such inherent characteristic, in that all members are simultaneously insurers and insured, share the same informal, local information that circulates informally and free-of-charge (i.e., gossip), and have (or could have) a say in claims adjudication. Thus, they have both the incentive to reduce costs (as excessive payments would translate into higher premiums) and the responsibility to settle claims (and can therefore filter unjustified ones).

Fraud is when someone knowingly provides inaccurate or incomplete information to claim benefits or advantages to which they are not entitled, or someone knowingly denies a benefit to which someone else is entitled and that is due. This issue is similar to free riding (although there may also be provider-induced fraud).

In the context of health microinsurance, the two business models that can reduce the risk of fraud by narrowing the gap between the flow of information and the flow of funds are the provider-driven insurance model and the mutual/cooperative model. This is because the underwriter also has much information on the claimants that can be availed free-of-charge. Provider-insurers in the provider-driven insurance model are in a potentially strong position to undertake provider fraud, which they can avoid only when they refrain from acting on their incentive to maximize profits. Operators in the mutual/cooperative model have no particular access to information on provider fraud, and would have to rely on investigation as would underwriters operating the partner-agent model and the charitable model, which could be disproportionately costly relative to the small sums insured.

Clearly, operators in the mutual/cooperative model have greater access to information that reduces or removes insurance failures arising from asymmetric information, i.e., adverse selection, moral hazard, free riding, and fraud. The exception is information asymmetries on provider-induced moral hazard and provider fraud, for which the insurer-providers in the

provider-driven insurance model have an information advantage. Absence of such information exposes the other business models to greater risk of failure (Table 2).

Application of Specific Actuarial Issues to Microinsurance

It is often stated that insurance is a numbers game relying on the Law of Large Numbers, vis: that the larger the number of independent risks in a pool, the lower the variance of mean losses. Lower variance translates to lower pure risk premium. Yet, most health microinsurance schemes are small, their intrinsic capacity to diversify risks limited, and their exposure to covariate risk is high due to the homogeneity of their clients. Simulation studies have shown that capital loadings to secure solvency are exponentially higher for small schemes. Notwithstanding the financial advantages and potential of lower premiums, pooling of schemes on a voluntary basis has not occurred. Pooling small schemes would be relatively simple if they all had an identical risk profile and shared priorities. In reality, health microinsurance schemes usually cover location-specific risks and priorities, which make pooling schemes more complex because of the differences from scheme to scheme and from community to community. The potential for governments to put in place mechanisms to adjust risks across mandatorily pooled schemes is remote, given the voluntary nature of microinsurance, and the damage such regulatory intervention could have on the role of the community in designing premiums and benefits. A proposal to create reinsurance for microinsurance (labeled 'social reinsurance'), which would provide large pool efficiencies at the reinsurance level, has so far not been implemented.

A paucity of data and its quality with which to determine stochasticity and quantify risks is a perennial problem for microinsurance, particularly health microinsurance. This means that launching a new micro health insurance (MHI) scheme must be preceded by data collection to ensure that premiums reflect rigorous risk estimates, and benefits are customized to address the main risks. Some early movers in the health microinsurance market took a simpler approach to the problem of lack of data by downsizing commercial insurance products that they had developed for the entire country instead of designing specific products with accurate pricing for this market based on a deep understanding of the particular needs of potential customers of microinsurance. The low uptake of such low-cost-low-benefit packages indicates that this approach was not suitable.

Some MHI schemes have introduced innovations in coverage that have actuarial ramifications. For example, in India and Nepal, where entire families share one 'purse,' some schemes have introduced a 'family floater' condition (i.e., a capped benefit which can be used by one or more members of that household) which requires rather sophisticated actuarial calculations to triangulate the estimated loss ratios to the distribution of family size.

In addition to ensuring that the pure risk premium is commensurate with the risk covered, actuaries need to calculate loadings on the premium to cover administrative, operational, and other costs, and, in some business models, profit. Given the high transaction costs associated with business models other than the mutual/cooperative business model, there is rationale to increase premiums, which is at odds with the clients' apparent WTP. In the absence of an acceptable notion of an equitable price, setting the premiums is fraught with uncertainties.

A different approach to explaining the link between premium and coverage has been to say that in microinsurance the price determines the coverage, whereas in other insurance the product determines the price; this point is elaborated in the Section The 'Make-It-Or-Break-It' Factor of Microinsurance: Willingness to Pay.

The 'Make-It-Or-Break-It' Factor of Microinsurance: Willingness to Pay

Under all definitions and types of health microinsurance, prospective clients, who are mostly living and working in the informal economy of low-income countries, affiliate on a voluntary basis. These people cannot be obliged (by governments or others) to pay a premium, even when subsidies cover a share of the expected costs. This means that the WTP of the target population determines the insurance package, rather than the product determining the price, as is typical in insurance. Therefore, WTP is the make-it-or-break-it factor of health microinsurance.

This is why it is essential to estimate WTP before launching the insurance. The most common method for prelaunch estimation of WTP for health microinsurance is contingent valuation (CV), which surveys the target population's responses to hypothetical insurance products and premiums. Respondents are required to think about the contingency (or feasibility) of an actual market for the benefits, and state the maximum they would be willing to pay for them. Over the years, different methods have been developed for the presentation of scenarios and the analysis of the responses.

WTP for health microinsurance is positively associated with income and increases nominally as income rises, but when expressed as a proportion of income, WTP declines as income grows; education; the quality and availability of health services; and recent exposure to healthcare costs. Men are willing to pay higher amounts than women. However, empirical evidence from India and Nigeria show that notwithstanding these variables, WTP is highly location specific, meaning that any temptation to roll out a one-size-fits-all microinsurance (be it in order to capture economies of scale in administering policies, or to establish some kind of a prescribed minimum level of benefits, or to aggregate the risk of more insured persons) may be thwarted. As WTP is location specific, so health microinsurance should be context-relevant in order to succeed. The related question is whether people actually pay the expressed WTP; at this point in time there is not enough published evidence on this question in the context of microinsurance.

The Impact of Health Microinsurance, and Why it is Not More Common

Early attempts to assess the performance of microinsurance for health were limited to measuring several accounting ratios,

Table 2 Summary comparison of the basic features of four microinsurance business models

	Partner-agent model	*Provider-driven insurance*	*Charitable insurance*	*Mutual/Cooperative insurance*
Objective	For commercial underwriter and for-profit agent: main objective is profit. For nonprofit agent (NGO) main objective pooling of health risks. And if NGO is MFI, insurance helps ensure healthier clients – and better loan repayment	For-profit providers: profit. For nonprofit providers: disseminate services and increase access to own services	Provide insurance cover for selected risks at 'affordable' prices and donor's overall purpose	Mutual aid of the members only, by pooling their risks and resources
Product design and pricing	Mainly high-cost and low frequency events	Benefit package designed with the view to increasing utilization of own services	Responding to the donor's perception of prioritizes needs	Responding to the members' prioritized needs and willingness to pay
Rating method	Risk-rating if sold to individuals; community rating if sold through NGOs or MFIs	Experience rating	Community rating	Community rating
Sales	Exclusively through agents (with or without commission)	Direct marketing, through paid or unpaid referrals and/or agents	Use of existing community structures (e.g., SHG, NGOs, etc.)	Use of existing community structures (e.g., cooperatives, SHGs, NGOs, women's associations etc.) and involvement of members
Servicing	Direct claims processing (back office); or through third-party administrators (TPA); in rare cases, agents also contribute to servicing	Servicing through own facilities	NGOs may have a role in claims servicing; otherwise, through back-office functions at providers' facilities (especially when donor arranges cashless access to services)	Both front-office functions (collection of premiums, dissemination of information etc.) and back-office functions; claims processing are done by community members
Sustainability				
Who is responsible?	Responsibility with underwriter	Responsibility with provider	Responsibility with donor	Responsibility with the insured, collectively
Source of income	Income from premiums; income from investment of reserves	Income from premiums; income from other activities of the provider or from subsidy	Income from premiums; income from the donor	Income from premiums
Pricing of premiums	Pricing must ensure underwriter profits (i.e., cover claims costs, marketing and admin costs, operational risk costs, regulatory compliance costs, agent commissions, plus profit).			
What if NGO?	For-profit: Pricing must ensure provider profits for services rendered to the insured; nonprofit: prices must ensure cost of services rendered minus any costs covered by subsidies	Pricing must cover the cost of benefits and admin costs up to the level considered by the donor as 'affordable'	Pricing must cover the cost of benefits plus admin costs	

(*Continued*)

Table 2 Continued

	Partner-agent model	*Provider-driven insurance*	*Charitable insurance*	*Mutual/Cooperative insurance*
Risks to sustainability	Nonrenewal due to client attrition; error in accurate pricing; fraud; moral hazard; provider-induced moral hazard; adverse selection	Nonrenewal due to client attrition; error in accurate pricing; fraud; moral hazard; adverse selection	Donor attrition; low affiliation rate if benefits do not meet clients' priorities; nonrenewal (client attrition); fraud; moral hazard; adverse selection	Error in pricing of premium; exposure to higher-than-expected cost of benefits due to random aggregation of cost-generating events (higher fluctuations due to small risk pools in the absence of reinsurer; nonprofessional management of scheme
Factors contributing to sustainability	Large client-base, high renewal rate, low claims ratio	Contributory factors include: provider attracts the target population (perceived as good, offers the right services etc.)	Contributory factors include: premium perceived as affordable, good fit between clients' and donor's perceived priorities	Contributory factors include: involving insured in benefit-package design and priority setting; good fit between premiums and ability to pay; lower admin costs when front office and back office are managed by the mutual/cooperative locally, at local prices; en bloc affiliation; high renewal rate; low claims ratio; social capital reduces moral hazard
Main conflicts of interest client versus underwriter				
Premium	Yes: the underwriter wants to draw profit NOT IF NFP, and the client wants low premiums and high coverage	Yes: the underwriter wants to draw profit NOT IF NFP, and the client wants low premiums and high coverage	No: donor and client share the wish to have low premium and high coverage	No: the cooperative and the individual client share the wish to have low premium and high coverage
Benefit caps	Yes: insurer prefers a low cap to reduce underwriting exposure; clients prefer high caps to reduce risk of having to pay above-cap OOPS	Yes: insurer prefers a low cap, which reduces underwriting exposure; client would like high caps, which reduce risk of having to pay above-cap OOPS	Yes: insurer prefers a low cap to increase the number of patients served for a given subsidy. Clients prefer high caps, which reduce risk of having to pay above-cap OOPS	No: the cooperative and the individual client share the wish to have caps as low as feasible
Range of benefits	Yes: insurer prefers a narrow range (mainly low probability and high costs); client prefers a broad range of benefits, to enhance likelihood that insurance will cover any event	Yes: insurer prefers a narrow range, limited to the services that the provider can deliver; clients prefer a broad range of benefits, to enhance likelihood that insurance will cover any event	Yes: insurer prefers a range of benefits that are directly linked to the donor's prioritized services; clients prefer a broad range of benefits, to enhance likelihood that insurance will cover any event	No: the cooperative and the individual client share the wish to cover the priorities of the community members

Abbreviations: NFP, not for profit, SHG, self help group.

mostly reflecting financial performance of schemes. More recently, the product, access, cost, and experience (PACE) is used by practitioners to develop a better value-proposition for clients by comparing various microinsurance products to one another and to alternative means of protection from similar risks (including informal mechanisms and social security schemes). However, neither the performance indicators nor the PACE tool offer conclusive and robust insight to three fundamental issues: (1) what difference does the insurance have on utilization of healthcare services among the insured? (2) what difference does the insurance have on the financial exposure/protection of the insured? and, (3) what impact does insurance related improvements in healthcare utilization have on the health of the target populations? These are considered in order as follows:

1. What difference does the insurance have on utilization of healthcare services among the insured? A literature review aimed at answering the question, 'Do clients get value from microinsurance?' suggested that 'value' included three aspects: (1) Expected value – the value clients may get from a product through behavioral incentives and peace of mind, even if claims are not made; (2) Financial value – the value of the product when claims are made compared with other coping strategies; and (3) Service quality value – the externalities created by microinsurance providing access to product-related services of benefit to the client. Answers to these questions were sought in 83 studies on health microinsurance products. According to that report, some 43 studies found that health insurance positively influenced the use of health services. And some 33 studies generally found that insurance led to lower out of pocket spending (OOPS) in case of hospitalization. The major impact of insurance on increasing utilization of health services was confirmed by a different literature review using the randomized controlled trials (RCT) method of measuring impact, and the Cochrane Handbook's characteristics.

   These findings should be put in context. In high-income countries it is often assumed that increased utilization of health services among voluntarily insured persons suggests (or is evidence for) adverse selection (namely, higher propensity to insure among persons who are likely to have above-average healthcare utilization). This assumption is not supported by the findings of studies of healthcare utilization among clients of health microinsurance, where higher frequency of illness was not systematically associated with insurance status, suggesting that in these populations the assumption of adverse selection must be rejected. Rather, it seems that most of the target population for health microinsurance in low-income countries suffers from chronic underutilization of healthcare services, due to the inability to pay for more or better healthcare. Thus, improved utilization of health services among the insured population is an indicator of success in achieving a key objective of health microinsurance, of reducing the limiting factor of unaffordability.

   However, the utopian aspiration that health microinsurance would put in place both more utilization and more equitable distribution of that utilization may be too much to ask, considering the inherent limitation of microinsurance. The poorer the insured person, or the lower the coverage relative to the full cost of services, the more likely it is that insured persons would be unable to pay any copay required to access insured benefits. Thus, the effect of health microinsurance schemes on equality is ambiguous in theory, and in practice, it has been observed to be both positive and negative.
2. What difference does the insurance have on the financial exposure/protection of the insured?One possible indicator of the impact of microinsurance on financial protection is the total OOPS that the insured must bear when accessing insured benefits taking into account also indirect costs and premiums. Although no such analysis has been published, there has been discussion of 'catastrophic' healthcare costs, which have been defined in terms of percentages of household income or disposable income net of subsistence needs. Although such definitions have been used to show significant reductions in the incidence of catastrophic costs among members of health microinsurance schemes, the definitions are insensitive to relative levels of hardship and other healthcare cost-related factors that lead to hardship such as the need to get money quickly, which may necessitate borrowing at high rates and/or selling assets on unfavorable terms. Such 'hardship financing' can be more costly than the healthcare and can throw the sick into poverty. Therefore a more appropriate impact indicator might be the extent to which health microinsurance reduces the frequency or intensity of hardship financing. Studies using this alternative indicator are yet to be carried out.

   The impact on medical expenditure patterns (disregarding indirect costs and premiums) have been studied in various contexts with mixed findings. Some studies have found OOPS (defined as healthcare expenditures net of reimbursement by insurance, either per visit or over the course of an illness) decreasing whereas others found no effect. However, OOPS based measures ignore both premium payments and frequency of visits. Other studies assessed the impact on total annual healthcare spending, either per person or per household, and found an increase in annual health spending because although the cost of individual visits sometimes decreases under health microinsurance, the number of visits typically increases, potentially leading to increased overall expenditure. However, these findings are obscured by the failure of the studies to control for changes in the price of healthcare and changes in household income. Finally, three schemes have been evaluated for their effect on some measure of the socioeconomic status (SES) of insured households. Although two schemes reported a statistically significant increase in SES with higher levels of income growth among households which are insured, and/or reduced likelihood to sell off their food stocks to pay medical bill, the third found no effect of insurance on household income levels, assets levels or self-reports of food sufficiency.

   The literature on estimating or measuring financial protection of microinsurance must be considered with reserve, due to the challenge of obtaining a statistically valid comparison between insured and uninsured cohorts. Most

studies compare utilization of healthcare benefits by the insured with the utilization by persons residing in the same geographical area who are not insured. These comparisons were flawed on several counts: Firstly, they were usually one-off studies following implementation of the microinsurance that did not adequately examine whether the cohorts were different before implementation of the insurance and whether any difference was attributable to the intervention or to an inherent difference between the cohorts. Secondly, a simple comparison of the situation among the insured cohort before and after implementing health microinsurance could be misleading because one cannot exclude the possibility that several changes occurred that were unrelated to the intervention but which had an impact A recent publication explains the methodology elaborated to address such a challenge, following Cluster Randomized Controlled Trial protocol. The method is tested in three waves of microinsurance implementation, ensuring that at the end of the experiment the entire population is offered affiliation with a community-based health microinsurance, but through staggered affiliation. In each wave, villages are grouped into congruous preexisting social clusters; these clusters are randomly assigned to one of the waves of treatment. Before each wave, a baseline evaluation is conducted (using mixed methods, with quantitative, qualitative and spatial evidence collected on the situation). This method assures that the microinsurance schemes operate in an environment replicating a nonexperimental implementation and that all households are offered insurance.

3. What impact does microinsurance have on the health of the target populations? The few studies have explored the health outcomes of health microinsurance have generally found that although healthcare utilization has increased, it is too early to say if the insured have better health. In some cases, there has been a lack of baseline information on health status, making evaluation more difficult. The scant evidence available suggests that any improvement in health outcomes is typically skewed toward the wealthier members of the schemes, possibly because they are better informed about health and have better access to noninsured care and support.

Concluding Note

If health insurance is a 'numbers' game,' and if health microinsurance is the pro-poor variant of health insurance, then it should become the dominant model by virtue of the huge number of persons in the informal sector without health insurance, of whom many are poor. However, progress to develop both supply of and demand for health microinsurance is contingent on developing a workable business model. With most of the target population living and working in the 'informal sector' where governments cannot mandate payment of premiums or apply means-testing for partial subsidization, the implementation of MHI depends on WTP. At the current level of knowledge of how to estimate WTP, it seems that participatory, needs-based, context-relevant, partial, and complementary solutions offer more promise than supply-driven one-size-fits-all products or mandated dissemination models. The partner-agent model remains subject to acute risk of conflicts of interest between underwriter, agent and client, which is never eliminated. Provider-driven supply of health insurance has so far not offered a general formula for scaling. The charitable model, based on delivering subsidized health microinsurance is limited by the funds that the charitable donors can devote in the long term. The mutual/cooperative model, though typically small scale can overcome most barriers to establish a functioning market for health insurance among the poor.

The poor want health insurance and can pay premiums and health microinsurance can operate without subsidy, but possibly not at profit and not with extensive commercial intermediation. The low penetration of health microinsurance can be explained in terms of barriers to the establishment of a market with viable supply and solvent demand in the informal sector. This overall context failure cannot necessarily be solved by the insurance industry offering innovative products through various channels, and cannot necessarily be solved by government regulation, although innovation and regulation are essential if we are to put in place systemic (regulatory and financial) mechanisms to encourage the development of local health microinsurance schemes or to pool risks across schemes, or articulate the relations between local schemes and commercial underwriters and reinsurers. Communities have a central role to play in building capacity and awareness, providing information for actuarial calculations and scheme designs to suit local priorities and service availability, and building the institutional context in which viable supply and solvent demand can be established. That said, local grassroots initiatives are neither willing nor able to scale microinsurance over entire countries.

Preferred Definition of Microinsurance for Health

As a conclusion, the preferred definition of health microinsurance is as follows:

> Health microinsurance is insurance contextualized to the WTP, needs and priorities of people in the informal sector who are excluded from other forms of health insurance. The schemes are voluntary, with premiums suited to people with low incomes. Although health microinsurance is independent of the size of the insurer, the scope of the risk covered, and the delivery channel, it is essential that the scheme is designed to benefit the insured. For practical intents and purposes, this definition implies a central role for the community in at least the design of the scheme, and possibly its operation and governance.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Global Health Initiatives and Financing for Health. Measuring Health Inequalities Using the Concentration Index Approach. Modeling Cost and Expenditure for Healthcare. Moral Hazard. Public Health in Resource Poor Settings. Rationing of Demand. Risk Selection and Risk Adjustment. Willingness to Pay for Health

Further Reading

General introduction

Dror, D. M. and Jacquier, C. (1999). Micro-insurance: Extending health insurance to the excluded. *International Social Security Review* **52**(No. 1), 71–97. (Geneva), ISSA.

IAIS (2007). Issues in regulation and supervision of microinsurance. Available at: http://www.iaisweb.org/view/element_href.cfmsrc=1/2495.pdf (accessed on 21 June 2013).

Prevalence of Microinsurance and Health Microinsurance

Matul, M., McCord, M., Phily, C. and Harms, J. (2009). The landscape of microinsurance in Africa. Available at: http://www.ilo.org/employment/Whatwedo/Publications/WCMS_124365/lang–en/index.htm (accessed on 21 June 2013).

McCord, M., Tatin-Jaleran, C. and Ingram, M. (2012). The landscape of microinsurance in Latin America and the Carribean. Available at: http://www.munichre-foundation.org/dms/MRS/Documents/Microinsurance/2012_IMC/20121010_Landscape_Microinsurance_LAC.pdf (accessed on 21 June 2013).

Roth, J., McCord, M. J. and Liber, D. (2007). The landscape of microinsurance in the world's 100 poorest countries. MicroInsurance centre. Available at: http://www.microinsurancecentre.org/resources/documents/doc_details/634-the-landscape-of-microinsurance-in-the-worlds-100-poorest-countries-in-english.html (accessed on 21 June 2013)

MicroSave. (2012). Securing the silent: Microinsurance in India the story so far. Available at: http://www.microsave.net/resource/securing_the_silent_microinsurance_in_india_the_story_so_far#.UV1dvaJHLoI (accessed on 21 June 2013).

Willingness to Pay, Actuarial Issues and Theory

Binnendijk, E., Dror, D. M., Gerelle, E. and Koren, R. (2013). Estimating willingness-to-pay for health insurance among rural poor in India, by reference to Engel's law. *Social Science & Medicine* **76**, 67–73.

Dror, D. M. and Armstrong, J. (2006). Do micro health insurance units need capital or reinsurance? A simulated exercise to examine different alternatives. *The Geneva papers on risk and insurance* **31**, 739–761. Available at: http://ssrn.com/abstract=1017101 (accessed on 21 June 2013).

Dror, D. M. and Koren, R. (2011). The elusive quest for estimates of willingness to pay for health micro insurance among the poor in low-income countries. In Churchill, C. and Matul, M. (eds.) *Micro insurance compendium II, 2012*, pp 156–173. Geneva: ILO and Munich Re Foundation.

Dror, D. M., Koren, R., Ost, A., et al. (2007). Health insurance benefit packages prioritized by low-income clients in India: Three criteria to estimate effectiveness of choice. *Social Science & Medicine* **64**(4), 884–896.

Nyman, J. (2003). *The theory of the demand for health insurance*. Palo Alto, CA: Stanford University Press.

Impact of Health Insurance

Aggarwal, A. (2010). Impact evaluation of India's "Yeshasvini". *Community Based Health Insurance Program, Health Economics* **19**, 5–35.

Dror, D. M., Radermacher, R., Khadilkar, S. B., et al. (2009). Microinsurance: Innovations in low-cost health insurance. *Health Affairs (Millwood)* **28**(6), 1788–1798.

Magnoni, B. and Zimmerman, E. (2011). Do clients get value from microinsurance? A systematic review of recent and current research. The Microinsurance Centre MILK project. Available at: http://www.microinsurancecentre.org/milk-project/milk-docs/doc_details/811-do-clients-get-value-from-microinsurance-a-systematic-review-of-recent-and-current-research.html (accessed on 21 June 2013).

Wagstaff, A., Lindelow, M., Jun, G., Ling, X. and Juncheng, Q. (2009). Extending health insurance to the rural population: An impact evaluation of China's new cooperative medical scheme. *Journal of Health Economics* **28**(1), 1–19.

Relevant Websites

http://www.microinsuranceacademy.org/
Micro Insurance Academy.

http://www.microinsurancecentre.org/
MicroInsurance Center.

http://www.ilo.org/public/english/employment/mifacility/
Microinsurance Innovation Facility.

http://www.microinsurancenetwork.org/
Microinsurance Network.

Health Services in Low- and Middle-Income Countries: Financing, Payment, and Provision

A Mills and J Hsu, London School of Hygiene and Tropical Medicine, London, UK

Introduction

A total of 7.6 million children and 287 000 mothers (2010 data) die every year, and approximately 95% of these deaths occurred in 75 countries with the highest burden of maternal and child deaths. Of these, more than two-thirds could be avoided if everyone had access to known effective interventions. Making such interventions available is not just a matter of supplying a drug or vaccine; to ensure effective delivery of interventions, health systems need to be strengthened at all levels, from the community up to the national level.

Economists investigating how health systems can be strengthened in low- and middle-income countries have explored a myriad of issues. This article addresses three core questions:

- How are health services financed?
- What payment methods are used to purchase health services?
- Who are the health service providers?

In each case the concern is to establish the evidence and discuss the implications of current arrangements for efficiency and equity.

There is very active debate on some key policy issues relating to reform of financing, payment, and provision. The second part of this article addresses some of the most contentious issues, notably:

- The appropriate mix of financing sources as countries seek to expand financial protection and move toward universal coverage of health care.
- The role and impact of development assistance for health (DAH) in low- and middle-income countries.
- The desirability of incentive-based payments to health service users and health care providers.
- The role of private sector agencies in health system arrangements (insurance, payment, and provision).

In addressing all low- and middle-income countries, this article considers a very wide range of country circumstances. Health systems differ greatly across low- and middle-income countries, influenced not only by the level of national income but also by countries' colonial history (British, French, Dutch, etc.), political orientation post-independence, degree of openness to market forces both historically and up to the present, income distribution (existence of high income groups with considerable purchasing power), and of course health conditions. To avoid implying that one pattern fits all, it is crucial to recognize this diversity and its implications for appropriate solutions to the many challenging issues facing health systems in these countries.

How are Health Services Financed?

In general, health services in low- and middle-income countries are financed at a much lower level than in high-income countries, and smaller proportion of the total flows through organized sources (i.e., government and insurance intermediaries). Table 1 summarizes health expenditure per capita and the pattern of financing sources and agents by income group and geographical region.

The larger a country's economy, the more it spends on health – high-income countries spend on average US$4660 on health per person compared to US$356 in lower- and upper-middle-income countries combined and US$61 in low-income countries. The level of health expenditure per capita thus mirrors gross national income per capita and the evolution over time can be displayed on Gapminder (hyperlink embedded in Figure 1). Although spending more on health does not necessarily lead to improved health outcomes, a minimum amount of financial resources is required by a health system to deliver essential interventions. It is estimated that spending of US$60 per capita is needed by 2015 for low-income countries to provide the basic package of essential services required to reach the health-related Millennium Development Goals and strengthen underlying health systems. At present, 15 of the 35 low-income countries spend less than the 2015 mark on all health care, and all but two of these countries (Bangladesh and Afghanistan) are located in Sub-Saharan Africa. These figures highlight the challenges faced by low-income countries, particularly those in Africa, in financing essential services for their citizens.

Lower levels of health expenditure are combined with relatively low shares of financing pooled across population groups, indicative of a lack of organized financing arrangements. Low- and middle-income countries usually lack an adequate tax base or large formal employment sector and/or have weaker infrastructure and management capability; some are also suffering from conflict or are in the midst of a political transition such that they have a weak or nonfunctioning state. In contrast, financing agencies in high-income countries are better established, typically following a model where services are funded primarily from general tax or compulsory social insurance. Some middle-income countries have been more successful than others in expanding pooling arrangements: for example, in East Asia and Europe, social security makes up 56% and 47%, respectively, of general government health expenditure. This particularly reflects China, Indonesia, Philippines, and Vietnam where social health insurance is mandated and countries of the former Soviet Republics, which developed social insurance schemes following independence.

The counterpart to relatively low levels of pooling is that countries rely more heavily on private health expenditures,

 doi:10.1016/B978-0-12-375678-7.01008-7

Table 1 Health financing indicators by income group and by region (in current PPP, millions of international $)

	THE per capita	*GHE per capita*	*External resources on health as % of THE*	*GHE as a % of THE*	*PHE as a % of THE*	*Social security funds as % of GHE*	*Out-of-pocket expenditure as % of PHE*	*Private insurance as % of PHE*
Income group								
LICs	61	25	26.0	40.0	60.0	3.8	77.7	2.3
LMICs	148	57	2.5	38.3	61.7	15.7	87.7	4.5
UMICs	576	318	0.3	55.3	44.7	45.5	73.8	17.5
HICs	4660	2997	0.0	64.3	35.6	67.4	38.9	53.7
Geographical region								
East Asia and Pacific	317	169	0.4	53.2	46.8	56.4	78.9	7.5
Europe and Central Asia	799	514	0.6	64.4	35.6	47.4	83.7	6.8
Latin America and Caribbean	845	432	0.3	51.1	48.9	23.1	64.8	31.7
Middle East and North Africa	426	202	0.6	47.4	52.6	36.8	95.0	4.7
South Asia	115	35	2.1	30.1	69.9	14.8	86.4	4.1
Sub-Saharan Africa	143	64	12.6	44.7	55.3	3.4	61.5	29.9

Abbreviations: GHE, government health expenditure; HICs, high-income countries; LICs, low-income countries; LMICs, lower-middle-income countries; PHE, private health expenditure; PPP, purchasing power parity; THE, total health expenditure; and UMICs, upper-middle-income countries.

Source: WHO Global Health Expenditure Database. Available at: http://apps.who.int/nha/database/PreDataExplorer.aspx?d=1 (accessed 20.05.12). Aggregated based on the World Bank's income and regional classification.

Data are from 2010 and are country weighted, not population weighted.

especially paid out of pocket. Figure 2 shows this reliance to be especially high in low-income countries and in the Sub-Saharan Africa region, and this pattern is also evident over time as displayed in time series data presented in Gapminder (hyperlink embedded in Figure 2). Indeed, private expenditures make up 60% of total spending in low-income countries (compared to 36% in high-income countries) and, within this, out-of-pocket payments represent the majority (i.e., 78% of private health expenditures) and therefore almost half of total health expenditure (Figure 2). High levels of out-of-pocket payments reflect the lack of government ability to collect taxes and provide accessible and good quality health care. Some low-income countries rely heavily on external resources to supplement public financing with donors on average contributing more than a quarter of total health spending in low-income countries. Such high reliance on external funding raises concerns for sustainability of services should these contributions decrease, as well as many other concerns discussed in the Section Development Assistance for Health.

A common criticism of financing patterns in low- and middle-income countries is that financing incidence is regressive – i.e., payments for health care weigh more heavily on lower income households. Recent studies have shed light on this question and are summarized in Table 2.

The mix of health financing sources varies substantially across countries, so it is important to consider the incidence of not only overall health care financing but also the main sources: the incidence of a specific type of tax can vary considerably depending on its design and the country context. For example, indirect taxes (e.g., value-added tax (VAT), fuel levies, and excise duties) are regressive in South Africa but progressive in Ghana and Tanzania, a difference likely be explained by the fact that a larger proportion of the South African population are able to purchase goods and services liable to VAT. In addition, although mandatory insurance is progressive in most countries, it is slightly regressive in the three Asian countries studied with universal insurance systems (i.e., Japan, South Korea, and Taiwan). However, such indices need to be interpreted carefully as in systems with less than universal coverage, the progressive insurance schemes may cover only a select population group composed of formal workers or those who are less poor.

Much of the discussion around appropriate financing mechanisms revolves around the need to protect households from catastrophic payments (i.e., levels of expenditures that are high relative to the amount of resources available to the household to pay for their basic needs, which World Health Organization (WHO) defines catastrophic expenditure as equal to or greater than 40% of nonsubsistence spending). The consequences of lack of financial protection against the costs of health care are abundant. Households may forego expenditures on other necessities such as food, clothing, or education, or they may borrow money or sell valuable household assets in order to pay for health services. They may also choose to simply not seek care at all, potentially exacerbating their illness and risking further adverse effects on their earnings. Catastrophic payments can occur in countries at all levels of economic development, but the incidence is higher where out-of-pocket payments are more than 15% of total health expenditures. Households in 18 low- and middle-income countries are therefore at especially higher risk of facing such costs, and up

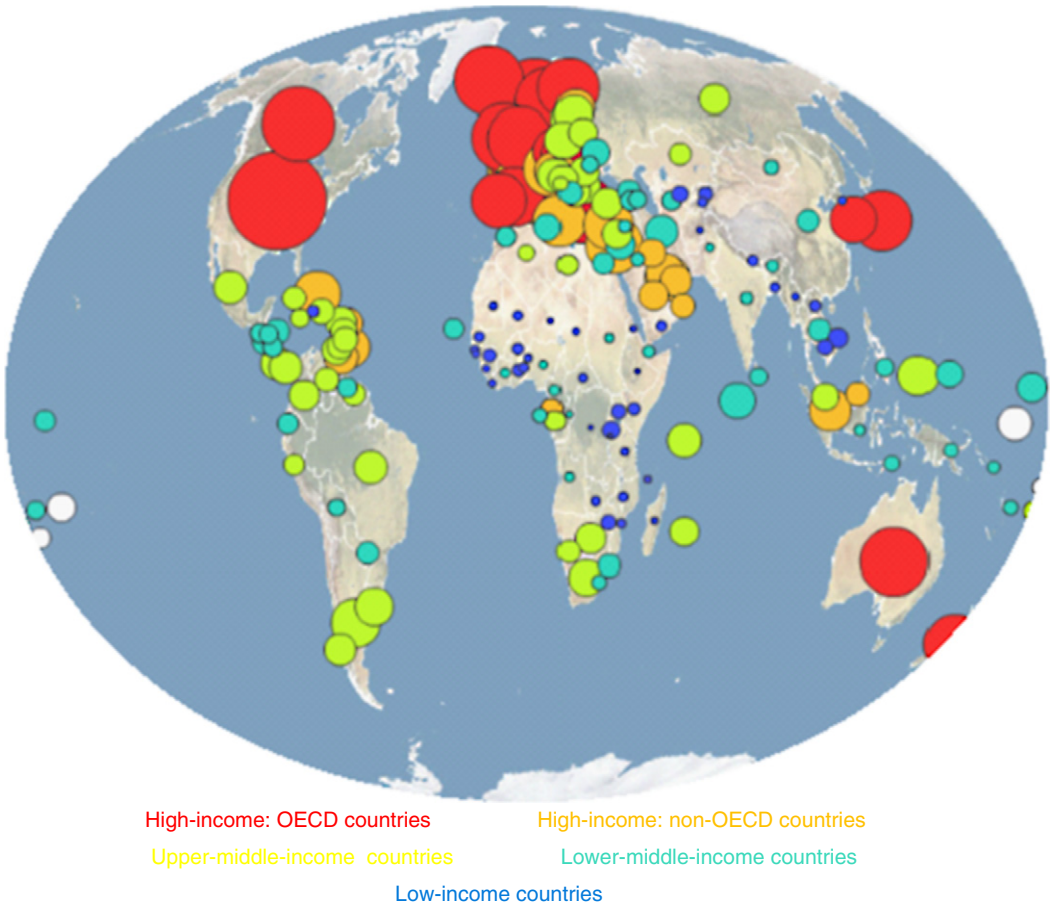


Figure 1 Total health expenditure (international $) per capita by country income group. Data presented by Gapminder; circles represent country data from the WHO such that the size represents total health expenditure per capita and the color represents an income group based on World Bank classification. To see an animated map showing the evolution of data over time, click on the map to visit Gapminder. OECD, Organisation for Economic Co-operation and Development. Reproduced from Gapminder. Available at: http://www.gapminder.org/data/ (accessed 30.05.12).

to 5% of households in those countries' risk being pushed into poverty by health care payments.

Low- and middle-income countries thus face major challenges in financing adequate health services and providing financial protection against catastrophic costs. Characterized by low levels of expenditure, fragmentation, and a reliance on out-of-pocket payments, the financing systems suffer from inequities and inefficiencies. Low- and middle-income countries need to expand forms of prepaid financing and reduce fragmentation in the flow and pooling of funds. Doing so will improve equity by cross-subsidizing risks between the rich and poor and the healthy and sick. It will further increase efficiency by decreasing administrative costs and duplicated coordination efforts required for multiple channels and pools.

However, the equity and efficiency of health financing systems are determined also by many other factors affecting both supply and demand. For example, the low status of women may affect their ability to leave the house to seek care; households may not be aware of the benefits of health care; local health services may lack drugs and qualified health workers or be staffed by health workers who are rarely present or who treat patients with disrespect. Equity and efficiency in financing health services further depend on how funds are used to pay for services and providers – issues covered in the following Sections How are Health Services Paid for? and Who are the Health Service Providers?.

How are Health Services Paid for?

Countries have a choice in deciding how to pay for health services and providers. These choices involve deciding how funding should be channeled from various funding pools (e.g., revenue generated by tax, insurance premiums, and DAH) and payments from individual payers to service providers. There are three principal methods for doing this:

- In relation to inputs (e.g., number of beds, facilities, staff, and items of service).
- In relation to services or outputs (e.g., outpatient numbers and inpatient cases or days).
- In relation to need (e.g., standardized mortality rates).

The payment method used tends to depend on the source of finance. Public funds have traditionally been allocated through hierarchical management structures down to the local

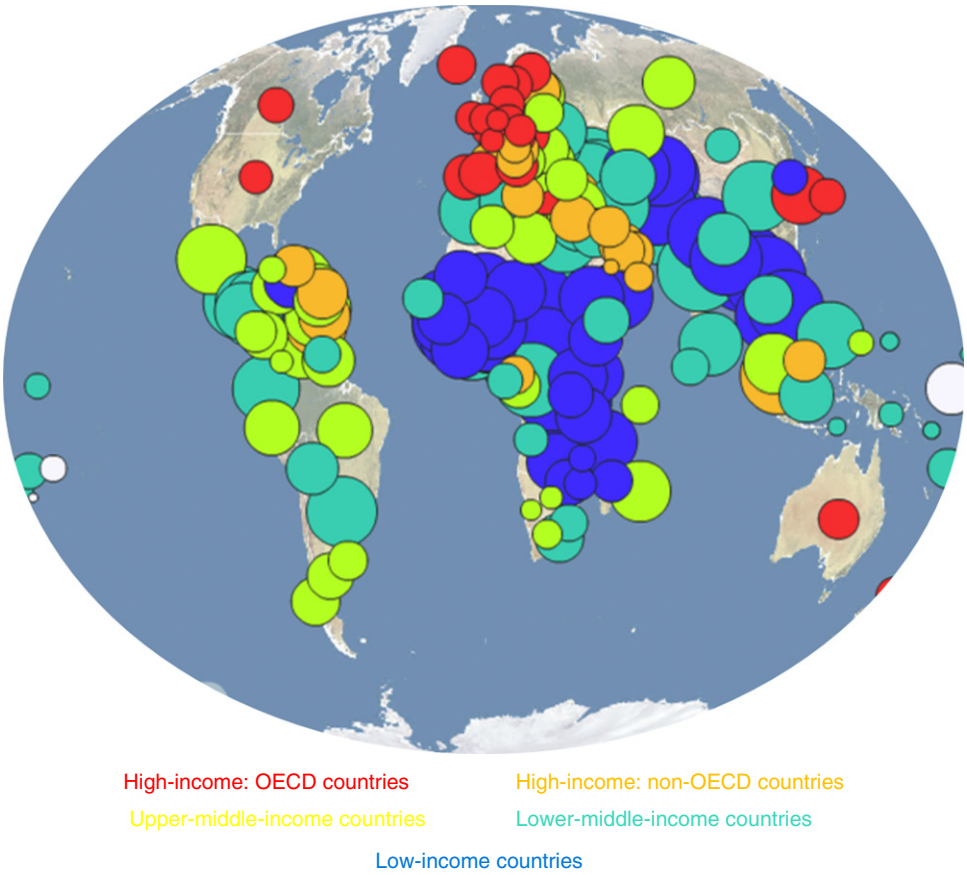


Figure 2 Out-of-pocket expenditure as a percentage of total health expenditure by country income group. Data presented by Gapminder; circles represent country data from the WHO such that the size represents out-of-pocket expenditure as a percentage of total health expenditure, and the color represents an income group based on World Bank classification. To see an animated map showing the evolution of data over time, click on the map to visit Gapminder. Reproduced from Gapminder. Available at: http://www.gapminder.org/data/ (accessed 30.05.12).

service delivery level and are frequently influenced by previous allocations, service or facility volumes and norms, capital developments and associated recurrent expenditure needs, and political influences. Such payment methods tend to assume the historical level of service inputs and outputs is optimal, or at least still appropriate, and does not especially consider efficiency or equity goals. Arrangements may be very inefficient as when more services are purchased than needed, or very fragmented, as in Indonesia where there are multiple channels through which funding flows to the district level. They may also be considered inequitable in that they may not adequately consider health needs, local costs, or income distribution of the recipients for those services.

Following the adoption of similar approaches in richer countries, many low- and middle-income countries have sought to introduce approaches that are population- and/or needs based. For example, Brazil, India, South Africa, India, Thailand, and Nigeria have all sought to improve equity in the allocation of public funds (including the health sector) across geographical areas through resource allocation formulae, which account for provincial variances in factors such as population, socioeconomic status, income levels, health needs, and/or membership in insurance schemes. Such approaches recognize that health services are geographically specific and purchasing should therefore be to some extent decentralized. These approaches are still evolving and commonly struggle to overcome both political influences and historical imbalances in the geographical distribution of the capital stock and related inputs. In Thailand, for example, there was a short-lived experiment with per capita allocation of total Ministry of Health funding; subsequently the salary element was removed and allocate separately, thus severely limiting the ability of the funding mechanism to improve inequality in the distribution of health workers.

Within the public health system, health providers are normally salaried and hospitals allocated an annual budget. More recently, contracts that link payments to the performance of health providers or facilities are increasingly found in the public sector and increasingly used to buy the services of private providers or facilities. These approaches, often termed 'results-based financing,' aim to increase efficiency in the purchasing of services, equity in access to priority services, and quality of service delivery, but evidence of their performance is sparse. Such issues are discussed further in the section Key Issues.

Insurance agencies (whether public or private) normally pay for health services using activity-related measures such as fee-for-service and case payment. The risk, especially with

Table 2 Kakwani indices for select African and Asian countries

	Direct taxes	*Indirect taxes*	*General taxes*	*Mandatory insurance*	*Total public*	*Private insurance*	*Direct payments*	*Total private*	*Total payments*
Asian countries									
Bangladesh	0.55	0.11	–	–	–	–	0.22	–	0.21
China	0.15	0.04	–	0.24	–	–	−0.02	–	0.04
Hong Kong SAR	0.39	0.11	–	–	–	0.04	0.01	–	0.17
Indonesia	0.20	0.07	–	0.31	–	–	0.18	–	0.17
Japan	0.10	−0.22	–	−0.04	–	–	−0.27	–	−0.07
Korea, Rep	0.27	0.04	–	−0.16	–	–	0.01	–	−0.02
Kyrgyz, Rep	0.24	0.05	–	0.14	–	–	−0.05	–	0.01
Nepal	0.14	0.11	–	–	–	–	0.05	–	0.06
Philippines	0.38	0.00	–	0.21	–	0.12	0.14	–	0.16
Sri Lanka	0.57	−0.01	–	–	–	With direct payments	0.07	–	0.09
Taiwan	0.26	0.03	–	−0.03	–	0.20	−0.10	–	−0.01
Thailand	0.51	0.18	–	0.18	–	0.00	0.09	–	0.20
African countries									
Tanzania	0.48	0.07	0.18	0.42	0.18	−0.49	−0.08	−0.08	0.05
South Africa	0.04	−0.02	0.01	–	0.01	0.14	−0.04	0.06	0.07
Ghana	0.20	0.06	0.10	0.26	0.14	−0.31	−0.07	−0.07	0.07

Note: The Kakwani index compares the distribution of health care payments across income groups such that a negative index indicates regressivity and a positive index indicates progressivity.

Source: Reprinted from Mills, A., Ataguba, J. E., Akazili, J., *et al.* (2012). Equity in financing and use of health care in Ghana, South Africa, and Tanzania: Implications for paths to universal coverage. *Lancet* **380**, 126–133.. doi:10.1016/S0140-6736(12)60357-2; table includes Asian data drawn from O'Donnell, O., van Doorslaer, E., Rannan-Eliya, R. P., *et al.* (2008). Who pays for health care in Asia? *Journal of Health Economics* **27**, 460–475.

fee-for-service payment, is that it encourages an unnecessary expansion in the volume of services and a subsequent increase in expenditure. For example, the fee-for-service payment system has been associated with a rapid increase in expenditure in Thailand (for the Civil Service Medical Benefit Scheme), South Africa (for private insurers), and Taiwan and South Korea (both associated with the implementation of universal health care coverage based on social health insurance). Such cost inflation has encouraged the introduction of payment methods, which do better at containing increases in expenditure. In 2002, the South Korean health system introduced a voluntary prospective payment method for inpatient care based on diagnosis-related groups, resulting in costs of care decreasing by an average of 8.3% in participating health facilities. Reform of the payment system, however, has not been comprehensive as plans to mandate the method were prevented by physician opposition. Thailand, in contrast, drawing on its own experience as well as that of other countries in the region, has had a very successful experience of payment reform with its universal coverage scheme. This pays for inpatient care based on diagnosis-related groups within a global budget and for outpatient care based on capitation payment. This has been relatively successful in extending financial protection while restraining costs: public health expenditure has increased steadily to compensate for increasing levels of utilization, but so far, the share of gross domestic product going to the health sector has not increased.

Household direct payments for care are made in response to fee schedules of providers. Although publicly levied fees may be quite simple in structure (e.g., a flat registration fee), private fees may be per item, with drugs charged separately and often with quite substantial markups. Indeed, practices in the procurement, prescribing and dispensing, and pricing of medicines account for three of the top ten causes of inefficiency identified by WHO in the 2010 World Health Report. In particular, drug dispensing is a major source of inefficiency when linked to prescribing functions as it can represent a significant source of income for private providers (and even public providers) – unofficial estimates indicate up to a 50% profit from drug charges in Taiwan. In response, some countries have sought to break the link between drug prescribing and provider income, a measure adopted some time ago in the rich world. These reforms have often been vehemently opposed with varying government responses and impact on expenditure. For example, Taiwan's 2002 reforms to separate purchasing and dispensing functions were met with strong resistance and a series of protests by the medical profession. To facilitate implementation of the policy, exceptions were made (e.g., rights to dispense were granted to clinics with on-site pharmacists). Such concessions dampened the impact on containment of total health expenditure, although it was successful in reducing drug expenditure. South Korea adopted a different, more rigorous approach in its 2000 pharmaceutical payment reform, breaking the link between prescribing and dispensing, removing all financial incentives, and eliminating profits earned by physicians from drugs. In reaction, however, physicians' fees increased by up to 44% and a greater proportion of brand-name drugs were prescribed.

Different payment methods thus provide different incentives to health providers and their implementation can sometimes have unexpected effects. Countries need to decide which arrangements to use for purchasing health services. These decisions will affect the efficiency, equity, and quality of services provided. For example, fee-for-service can not only

promote responsiveness and productivity but also can lead to inefficiency through supplier-induced demand and cost escalation; capitation and case-based payment can promote efficiency and affordability but may be problematic for quality. The performance of payment systems are determined by the incentives set and how much is being paid, what is being paid for, and who is being paid. In addition, in contexts where capacity to monitor is weak and data limited, there is greater risk of fraud and greater difficulty in fine-tuning payment systems to get the desired results.

Who are the Health Service Providers?

As countries grow richer, a greater share of total health expenditure is publicly financed, as discussed in the Section How are Health Services Financed?, and hence a greater proportion of health care provision is formally organized. The poorer the country, the greater the diversity of types of provider and greater the fragmentation of health services. In general, health service providers can be categorized into seven main groups:

- Government health services for the general public.
- Services run by social health insurance agencies (in countries where they are direct service providers).
- Services run by nongovernment organizations (NGOs) including church organizations.
- Occupational health service providers, both government (e.g., army) and private (e.g., mines and plantations).
- Private for-profit allopathic providers, both individuals and facilities.
- Traditional medicine providers ranging from the more formal (e.g., Ayurveda) to the somewhat less formal (e.g., traditional healers).
- Informal providers such as drug peddlers and unqualified providers (e.g., known as quacks in India).

Data on health providers are much more limited than on health financing. In particular, data on private provision are especially scanty, making it difficult to quantify the relative share of public and private provisions. The 2006 World Health Report stated that approximately 70% of physicians and 50% of other health workers cited their employment as within the public sector; however, the report pointed out that the actual distribution in the public sector is likely to be much lower as the data tend to reflect the health worker's primary employer rather than their main source of income, which in low- and middle-income countries can be significantly higher in the private sector. Evidence on health worker income from Ethiopia and Zambia underscores that the private sector offers much higher remuneration than in the public sector (Figure 3).

Data on utilization patterns can provide additional information on public and private health service providers. Figures 4(a) and (b) show the relative importance of the two sectors in providing health care to women and children in 25 low-income countries. Although there were high levels of variation across individual countries, the use of public health service providers more than half of the time was reported in only four

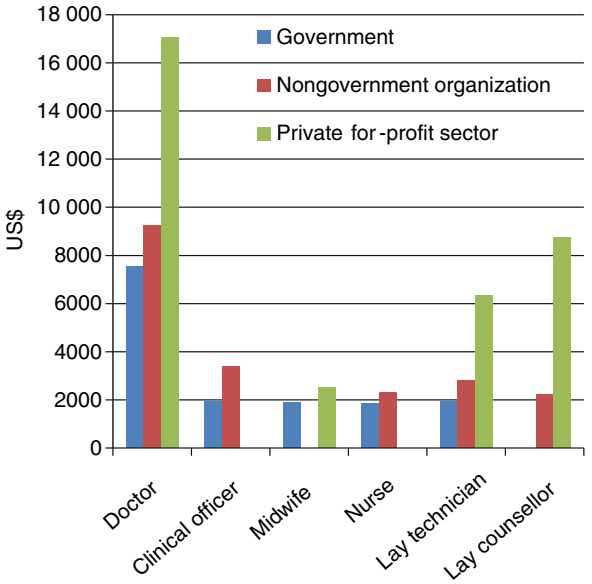


Figure 3 Average annual salaries for health providers in Zambia in 2004. Reproduced from McCoy, D., Bennett, S., Witter, S. *et al.* (2008). Salaries and incomes of health workers in sub-Saharan Africa. *Lancet* **371**, 675–681.

of the countries for deliveries and in only seven of the countries for child fever/cough. In general, adults, especially men, tend to use private facilities more than children, and the probability of using public facilities is higher for inpatient than outpatient care. For example, other cross-country analyses have found that public hospitals account for 73% of inpatient stays in 39 low- and lower-middle-income countries.

The distribution of health expenditures can also give an indication of balance of service provision. With regards to the level of care, hospitals account for approximately 60% of government health expenditures with tertiary hospitals absorbing as much as 45–69%. Such high levels raise efficiency concerns as hospital care tends not to be the most cost-effective when primary care coverage is incomplete. Indeed, inappropriate hospital admissions and excessive lengths of stay, as well as inappropriate hospital size, represent two of WHO's top 10 sources of health care inefficiency. The distribution of Official Development Assistance (ODA) for health indicates the priorities of donors: 40% of 2010 ODA disbursement went to providing HIV care and 19% to controlling infectious diseases with only 15% to basic health care and infrastructure (Figure 5).

Table 3 and Figures 6(a) and (b) provide data on various other dimensions of health service provision across income groups and regions. These show the relative lack of available service inputs and much lower coverage rates of essential interventions in low- and lower-middle-income countries – all of which carry implications for the equity and efficiency of service provision in developing countries. For example, low-income countries have five times fewer physicians per 10 000 individuals and approximately 50% fewer births attended by skilled health personnel and 16% lower coverage of child immunizations when compared to high-income countries. At the regional level, Sub-Saharan Africa has 30 times fewer

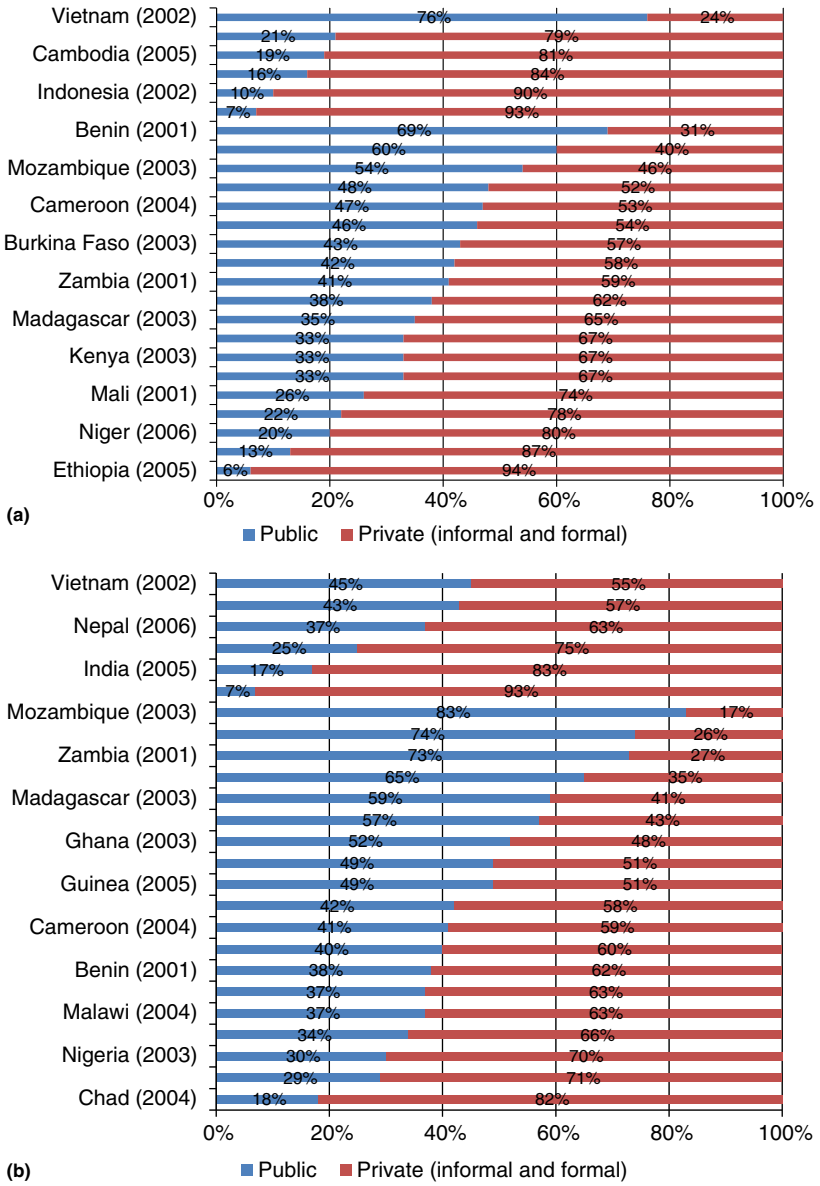


Figure 4 (a) Mothers giving birth in public or private facilities (%); (b) Children treated for fever/cough in public or private facilities (%): Data obtained from 2001 to 2006 demographic and health surveys (DHS). Public sector means health facilities and providers affiliated with the government. Private sector means formal private (e.g., commercial; for-profit hospitals, clinics, or pharmacies; facilities or providers that belong to NGOs or missions) and informal (e.g., traditional healers, drug peddlers or vendors, and shops as well as care provided by friends and relatives and other unspecified providers). Reproduced from Limwattananon (2008). Private–public mix in health care for women and children in low-income countries: an analysis of DHS. Thailand: International Health Policy Program.

physicians per 10 000 individuals and approximately 45% fewer births attended by skilled health personnel and 19% lower coverage of child immunizations when compared to Europe and Central Asia. The health worker shortage in these countries means the insufficient number of providers cannot adequately deliver the care needed in countries with major disease burdens. In the public sector, the mix of doctors and nurses and ratio of health providers to patients are suboptimal, with health workers frequently facing an overwhelming workload and hence delivering low quality of care. It is often for these reasons that the poor seek care in private facilities, which tend to be better staffed and provide more responsive care but often at a higher cost and not necessarily greater effectiveness.

There has been a long-standing debate over the relative efficiency of public and private providers, with claims that private providers are more efficient. However, evidence to support this is scanty and suffers from difficulties in standardizing for type of patients and service models. For example, a study of the provision of primary care in South Africa by various forms of providers (i.e., public clinics, private general practitioners (GPs) contracted to provide free care for poor

patients, private GPs practicing privately, a private clinic chain, and company clinic) found that two of the private sector models were delivering services at comparable cost to the public sector – the contracted GP model and clinic chain. However, the two other private sector models (i.e., independent GPs and company clinic) were delivering services at much higher cost, demonstrating the importance of examining the private sector model by model. Contextual influences, such as payment methods, practice styles, and traditions, also affect performance. Regardless of the type of ownership, investing resources in more efficient providers can result in substantial savings and a great potential to provide more health services within a fixed budget. In Namibia, savings from reducing hospital inefficiency could construct 50 clinics and, in South Africa, represent three times the value of user fee revenue.

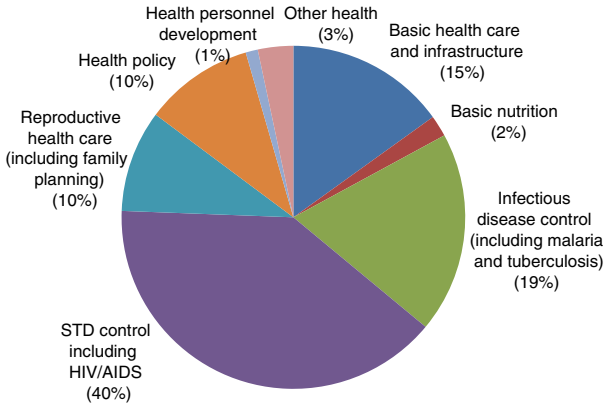


Figure 5 Distribution of 2010 ODA to health in low- and middle-income countries. STD, sexually transmitted diseases. Reproduced from OECD Creditor Reporting System. Available at: http://stats.oecd.org/Index.aspxdatasetcode=CRS1 (accessed 30.04.12).

Another major deficiency is strong evidence on the quality of health services. However, evidence is sufficient to confirm that quality in both public and private sectors is poor, with the private sector tending to perform better in drug availability and aspects of delivery of care, including responsiveness and effort, and possibly being more client orientated. In the case of the South African study referred to above, public clinics tended to offer better technical quality of care than private facilities, but quality as perceived by users was lower due to more crowded facilities and less responsive staff. But there is enormous variation. Many countries, for example, include at one end of the spectrum public and private hospitals offering care of international levels of quality, whereas at the other end of the spectrum are unlicensed and unqualified providers selling drugs, which should be prescription only. Arrangements may be agreed between hospitals and diagnostic laboratories, for example, to refer patients in return for a fee, and regulators may not be independent of the facilities they regulate.

There has been persistent criticism that the use of public services in low- and middle-income countries is inequitable, in that richer groups benefit more than poorer groups. A recent in-depth study of benefit incidence (and financing incidence) in Ghana, South Africa, and Tanzania confirmed this with respect to Ghana and South Africa, although public sector and faith-based organizations' health service benefits in Tanzania were more evenly distributed across the population. Inclusion of private sector services in this benefit incidence

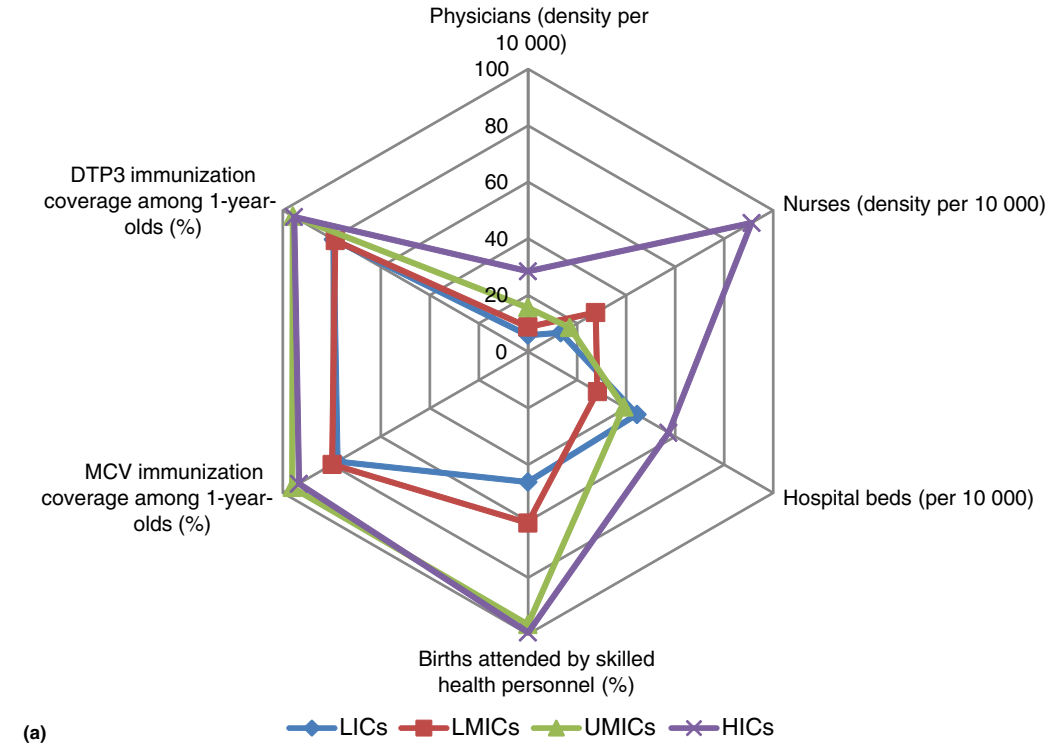
Table 3 Health service inputs and immunization coverage levels by country income group and by region

	Physicians (density per 10 000)	*Nurses (density per 10 000)*	*Hospital beds (per 10 000)*	*Births attended by skilled health personnel (%)*	*MCV immunization coverage among 1-year-olds (%)*	*DTP3 immunization coverage among 1-year-olds (%)*
Income group						
LICs	5.8	13.4	44.5	46.1	77.7	79.3
LMICs	8.7	27.6	28.3	60.7	79.8	78.6
UMICs	15.6	17.1	39.2	96.5	96.1	95.8
HICs	28.5	91.2	57.3	99.4	93.4	95.4
Geographical region						
East Asia and Pacific	14.2	13.8	39.1	92.6	95.3	94.2
Europe and Central Asia	26.8	73.2	55.2	99.5	96.2	95.2
Latin America and Caribbean	17.2	9.2	15.2	93.3	93.6	93.1
Middle East and North Africa	17.4	28.1	17.1	96.2	89.3	90.2
South Asia	6.4	4.3	43.0	57.7	77.3	76.2
Sub-Saharan Africa	0.9	9.9	8.1	47.3	75.5	76.6

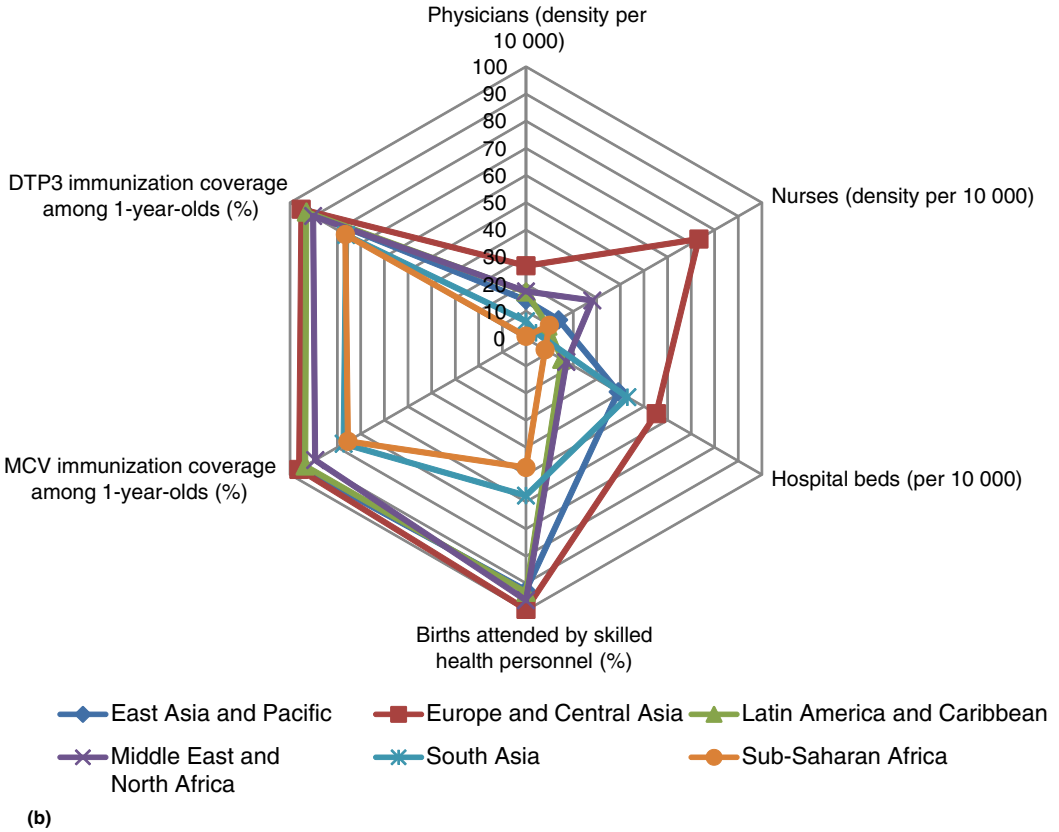
Abbreviations: DTP3, Diphtheria tetanus toxoid and pertussis; HICs, high-income countries; LICs, low-income countries; LMICs, lower-middle-income countries; MCV, Measles; and UMICs, upper-middle-income countries.

Note: Input data (i.e., physicians, nurses, and hospital beds) are from 2009. Coverage data (i.e., birth attendants and immunizations) are from 2010.

Source: WHO Global Health Observatory. Available at: http://apps.who.int/gho/data/ (accessed 23.05.12). Aggregated based on the World Bank's income and regional classification.



(a)



(b)

Figure 6 Health service inputs and immunization coverage levels by country: (a) income group and (b) regional group. DTP3, Diphtheria tetanus toxoid, and pertussis; MCV, Measles. Input data (i.e., physicians, nurses, and hospital beds) are from 2009. Coverage data (i.e., birth attendants and immunizations) are from 2010. Reproduced from WHO Global Health Observatory. Available at: http://apps.who.int/gho/data/ (accessed 23.05.12). Aggregated based on the World Bank's income and regional classification.

analysis showed even greater disparities in the distribution of benefits. Overall, health services benefited higher income groups despite the greater health needs of lower income groups. The key reasons constraining the access of poorer groups were problems in relation to the availability, affordability, and acceptability of services, particularly health care costs, transport costs, drug stock-outs, insufficient staff numbers, and poor staff attitudes. Such barriers need to be addressed to change the distribution of health services and move toward greater financial protection.

Key Issues

Financing Sources for Universal Coverage

Over the last few years, there has been growing momentum to expand financial protection and set universal coverage as a long-term goal. Evidence has been accumulating from countries such as Thailand that given willingness of governments to support the health care costs of the less well off and design features that constrain cost inflation, universal coverage of a benefit package of reasonable size is possible even for a lower- middle-income country. For example, Thailand achieved universal coverage in 2001 (at a per capita income of US$1900) by introducing a new scheme funded from general taxation to cover the 47 million people who fell outside the preexisting schemes for formal sector workers. Vietnam, Philippines, and Indonesia have now adopted universal coverage as a goal with a timetable for achievement. Both South Africa and India are actively debating plans for universal coverage.

It is clear that a mix of financing sources is needed for progress to be made toward universal coverage – general tax revenues are needed for those too poor to contribute; social health insurance arrangements are of value for enrolling formal sector workers; some degree of contributions from user fees is probably inevitable because even with offer of services free at the point of use, some people will still choose to purchase their care from the private sector. The critical question, over which there is considerable disagreement, is whether those in the informal sector who are not the poorest – often a very substantial number of people – should be covered by general tax funding or enrollment in contributory schemes (with or without government subsidy). Thailand, for example, has chosen general tax funding; Philippines and Indonesia have chosen to seek to extend their social health insurance scheme on a voluntary basis to encompass the informal sector; China has rolled out a massive and highly subsidized rural voluntary insurance scheme covering 835 million people by 2011. Key issues are willingness for the share of government funding to health to increase and the feasibility and management costs of encouraging a high proportion of the target population to enroll voluntarily. The latter concerns have led to a plan in Ghana, where a national health insurance scheme was introduced including voluntary enrollment into district insurance schemes, to move to a 'one time premium,' a largely nominal payment, thus recognizing the de facto situation that the great majority of funding for universal coverage is coming from direct and indirect taxes.

Development Assistance for Health

As shown in the Section How are Health Services Financed?, DAH is a substantial source of health financing in low-income countries – reaching more than a quarter of total health spending. Trend analysis further shows that the total amount of DAH has substantially increased over the last two decades, from an estimated US$5.8 billion in 1990 to US$27.7 billion in 2011 (in 2009 US$). DAH can have a number of economic consequences as well as political implications.

Development assistance has been criticized for fostering donor dependency and hindering economic growth in recipient countries. Indeed, a high reliance on external funding for health raises concerns over the ability of the government to deliver basic health services. Should these contributions decrease – and some recent estimates are showing a decreasing rate of growth of DAH flows since the global financial crisis – it would threaten the delivery of essential health care. Any gap in health financing would need to be covered by the government or private funding. In low- and middle-income country settings, where there are institutional, economic, and fiscal constraints hampering significant government funding increases, the outcome would most likely be higher out-of-pocket payments, further restricting access to health services by the poor.

However, development assistance has also been promoted as a means to empower countries to lead their own development by providing opportunities for strengthening the role of the state and for economic growth. Development assistance can help to build basic health infrastructure, especially in underserved areas or postconflict settings, which can be a visible and important indicator of a functioning state. It may also stimulate improved sector-level policies and strategies, especially when development assistance is channeled through mechanisms such as Sector-Wide Approaches (SWAps).

There has been controversy over whether DAH displaces domestic spending on health. A recent statistical analysis of expenditure data over the period 1995–2006 suggested that for every US$1 of DAH to governments, there was a decrease in government health expenditures by US$0.43–1.14. The analysis further found that when DAH was given to the nongovernmental sector, government health expenditures increased by US$0.58–1.72. However, the evidence for displacement is still inconclusive, not least because of data limitations. Data at the country level, especially in low-income countries, are often missing and estimates frequently vary across institutions (e.g., the degree of correlation between the WHO and International Monetary Fund estimates for government health expenditure is only 65%). In addition, the probability and extent of displacement is likely to vary greatly across countries. For example, in response to increases in DAH, the Democratic Republic of Congo appears to have decreased its domestic health spending by more than 30%, whereas its neighbor, the Central African Republic, increased spending by more than 30%. Factors specific to individual countries, such as donor behavior and domestic policy choices, are likely to be influential. Thus, firm conclusions cannot be drawn, and it is imperative to understand not only whether such effects are occurring but also why. Moreover, the debate underlines the need to understand broader issues such as how domestic spending responds to the volume and type of development assistance.

Additional issues relate to other aspects of the effectiveness of aid. There has been very long-standing concern that health aid flows through far too many channels, is fragmented and excessively tied to specific short-term projects rather than longer term programs and strategies, and is unpredictable. For example, the change in flows of funds from 1 year to the next can create difficulties in implementing sustainable health programs in recipient countries. The volatility in aid given to health over time is shown in time series data presented by Gapminder (hyperlink embedded in Figure 7). Furthermore, the individual reporting requirements of numerous development partners put pressure on already weak financing systems in recipient countries. All of these concerns are reflected in harmonization and alignment principles agreed in the Paris Declaration on Aid Effectiveness and subsequent Accra Agenda for Action. However, much remains to be done. For example, the proportion of ODA to maternal, newborn, and child health, which flows to projects (rather than sector-wide support, for example), has consistently stayed approximately 90% over the 2003–10 period. Various joint donor funding arrangements have sought to coordinate donor support, but many funding flows remain outside coordination mechanisms.

Results-Based Financing to Users and Providers

Results-based financing has recently attracted much attention as a way of implementing agreed priorities through stimulating demand, purchasing services, and encouraging improved health worker productivity and service quality. It is defined as 'a national-level tool for increasing the quantity and quality of health services used or provided based on cash or in-kind payments to providers, payers, and consumers after predetermined health results (outputs or outcomes) have been achieved' (http://www.rbfhealth.org/rbfhealth/about). It is a generic term for a number of different approaches, including:

- Provision of vouchers to enable individuals or households to obtain health care.
- Payment of cash to households conditional on use of specific services and other sorts of financial transfers, for example, to cover transport cost.
- Payment of financial incentives to providers (individual health workers, facilities, or organizations) to supply certain types of services or reach certain quantity or quality targets.
- Agreeing contracts for services with associated performance targets.
- Output-based aid where provision of aid is conditional on achievement of certain targets, such as a minimum immunization coverage level.

Interest in such approaches has grown rapidly over the last few years, with dedicated funding for such projects being provided by the World Bank and bilateral aid agencies. Results-based financing has been introduced in a number of

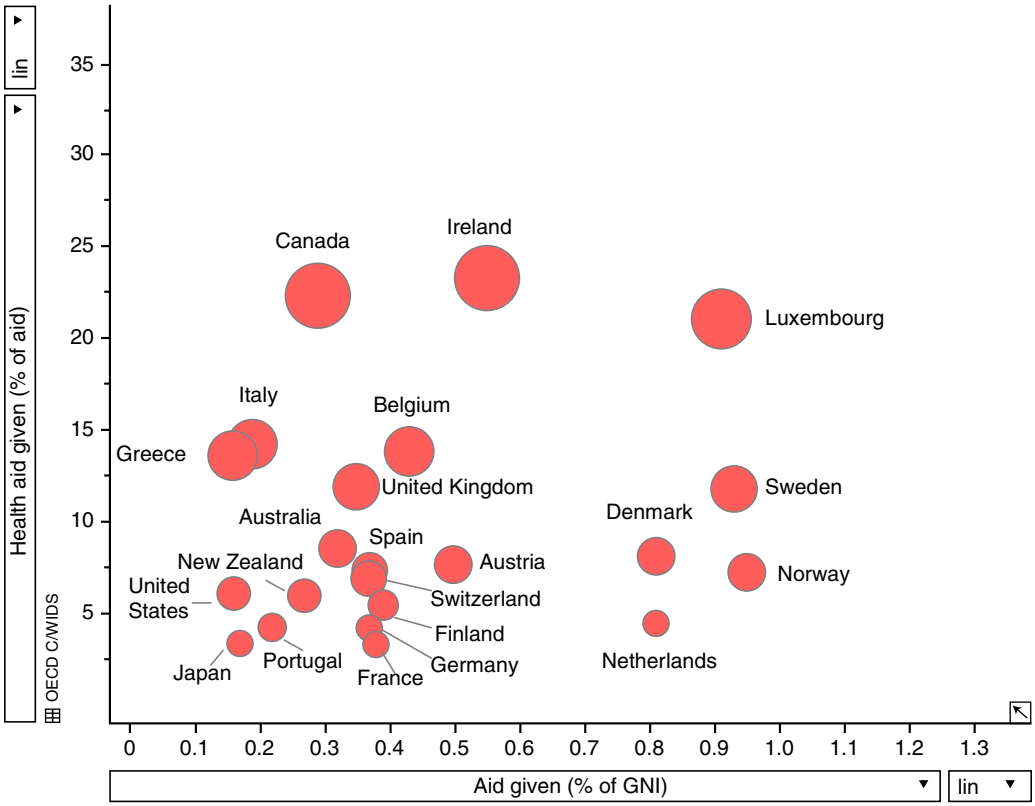


Figure 7 Volatility of aid to health. Data are from 2007 and presented by Gapminder; circles represent high-income country data from the OECD such that the size represents aid given to health as a percentage of total aid. To see an animated map showing the volatility of aid over time from 1971, click on the chart to visit Gapminder. Reproduced from Gapminder. Available at: http://www.gapminder.org/data/ (accessed 03.06.12).

developing countries, particularly in the Latin America and South Asia regions, and frequently for maternal and child health interventions. Although positive results have been reported in several schemes, reliable evidence on effectiveness, especially in low-income countries, is still fairly sparse. There is a gap in understanding how such mechanisms can improve performance and what are the necessary factors to ensure intended effects, and there is virtually no evidence on the cost-effectiveness of such approaches relative to other ways of improving provision of services and increasing the uptake of health care.

Proponents of results-based financing point to available evidence suggesting that such incentives have positively influenced various levels of the system – recipients of health care (individuals and households), providers of health care and facilities, as well as resulted in positive outcomes including higher coverage of key interventions, better service quality, increased efficiency, and/or improved health outcomes. Rwanda has often been cited as an example for its pay-for-performance scheme, and reports often identify increases in uptake of maternal and child health interventions. However, results-based financing may also increase inequities and produce undesirable effects (e.g., reducing the intrinsic motivation of health workers, gaming, cherry picking, neglect of other activities, and corruption). The effectiveness of performance contracts with the private sector has also been questioned as, while there is evidence they have improved access to services, little is known about its impact on the equity, efficiency, and quality of care of the wider health system. Finally, output-based aid can not only accelerate achievement of health targets but has also been criticized for being too narrow and short term in focus. The varied results are a function of the range of results-based financing instruments, their individual design, and implementation in diverse country contexts. Scheme design should be based on an understanding of the underlying problems the scheme is intended to address and on the country context (e.g., taking into account local managerial capacity), and performance indicators must be aligned with the goals of the health system. The impact of results-based financing also depends critically on the ability to implement the scheme effectively and monitor performance.

More broadly, results-based financing is not only argued to improve accountability (allowing for regular reviews of performance) and increase equity (in targeting certain population groups) and efficiency (in improving performance) but it also raises questions over the degree of involvement of donors in scheme initiation, design and implementation, and the sustainability of such arrangements beyond the initial donor funding.

The Role of Private Sector Agencies

Concern about the capacity and performance of governments in both low- and middle-income countries has led to considerable interest in how private sector agencies may perform some roles traditionally assigned to the state. Such roles may include:

- The provision of private insurance.
- The administration of insurance arrangements on behalf of the state.
- The management of drug distribution systems and other elements of public health service management.
- The provision of services.

Debates about private insurance mirror those in high-income countries – namely that it is likely to be neither an efficient nor equitable way of providing financial protection to significant numbers of people. Moreover, there are few countries, which have any sizeable private health insurance sector, given the very limited market of those who can afford to pay. The main potential role is to provide additional cover, to relieve the public health system of the pressure to cater for the highest income group.

A different role for private insurers is to administer state-sponsored financing arrangements. For example in India, the Rashtriya Swasthya Bima Yojana scheme, launched in 2008, targets households below the poverty line. Parastatal and private insurers bid to administer the scheme, which involves receiving a fixed sum of public money per household recruited to the scheme, providing them with a smart card, which is both the evidence of membership and records health care costs up to the allowable maximum per year, signing up hospitals to provide care, and managing payment arrangements. There is annual retendering of the contract, with competition focusing on the fixed sum per household that is requested. This design has permitted very rapid roll out of the scheme across India, with 40 million people covered by 2012. Concerns have focused on low rates of utilization of care by members in some states (hence increasing the profits for the company), fraudulent claims by providers, and in some states incremental creep year by year in the capitation sum.

The management strengths in the private sector have also been drawn on in other areas of health system management. For example, South Africa, which has some considerable private sector capacity, has experience of contracting out drug distribution to hospitals and clinics and also of contracting a private company to manage public hospitals. Evaluations of such arrangements have identified issues similar to those found in high-income countries – the challenges of managing the principal–agent relationship; difficulties of specifying contracts for clinical care; and difficulties public agencies can face in managing contracts well.

Private agencies can play two main roles in service provision. Private providers can directly be contracted to provide services on behalf of the state. Most experience of this model comes from contracts with NGOs, both the international NGOs and indigenous ones, and there is evidence that NGOs working under contract and managing district services have increased service delivery in underserved areas. A second approach is to use a variety of means to improve the quality and reduce the cost of the less formal part of the private sector that is extensively used by poorer groups. Approaches such as accreditation of clinics, franchizing outlets to provide contraception and sexually transmitted diseases treatment, and training of drug sellers can work successfully, although experience is very varied and most approaches have been tried only on a very small scale.

Effective engagement with the private sector is important, but a strong public primary care system has been shown to be

critical in bringing health services to communities and improving health outcomes. For example, the experiences of Ethiopia and Bangladesh in investing in human resources and innovative delivery methods in the public system have resulted in wide reaching and effective primary health care systems (see videos at http://ghlc.lshtm.ac.uk).

Conclusions

This article has covered a very wide canvas in terms of both countries and issues. Echoes are apparent with many of the issues facing high-income countries – the best mix of financing sources, role of out-of-pocket payments, best ways to pay providers, desirability of incentive-based arrangements, and relative roles of public and private sectors. However, the context of low- and middle-income countries means that policy lessons from high-income countries do not necessarily transfer well to all low- and middle-income country settings. Key features that affect the relevance of policies include the very widespread poverty; high proportion of the population in the informal sector; relative weakness of political; and social institutions including governance structures; limited management capacity in the public sector, and vulnerability to influence by agencies external to the country. Numerous studies show that the detailed ways in which policy reforms are designed and implemented in particular contexts play a key role in how they perform, alerting us to the need to be wary of seeking global solutions to health system challenges.

See also: Development Assistance in Health, Economics of. Global Health Initiatives and Financing for Health. Health Microinsurance Programs in Developing Countries

Further Reading

Berendes, S., Heywood, P., Oliver, S. and Garner, P. (2011). Quality of private and public ambulatory health care in low and middle income countries: Systematic review of comparative studies. *PLoS Medicine* **8**, e1000433, doi:10.1371/journal.pmed.1000433.

Gottret, P. and Schieber, G. (2006a). Financing health in low-income countries. In Gottret, P. and Schieber, G. (eds.) *Health financing revisited: A practioner's guide*, pp. 209–248. Washington, DC: The World Bank.

Gottret, P. and Schieber, G. (2006b). Financing health in middle-income countries. In Gottret, P. and Schieber, G. (eds.) *Health financing revisited: A practioner's guide*, pp. 249–278. Washington, DC: The World Bank.

Goudge, J., Russell, S., Gilson, L., Molyneux, C. and Hanson, K. (2009). Household experiences of ill-health and risk protection mechanisms. *Journal of International Development* **21**, 159–168.

Kalk, A. (2011). The costs of performance-based financing. *Bulletin of the World Health Organization* **89**, 319.

Lu, C., Schneider, M. T., Gubbins, P., et al. (2010). Public financing of health in developing countries: A cross-national systematic analysis. *Lancet* **375**, 1375–1387, doi:10.1016/S0140-6736(10)60233-4.

Meessen, B., Soucat, A. and Sekabaraga, S. (2011). Performance-based financing: Just a donor fad or a catalyst towards comprehensive health-care reform? *Bulletin of the World Health Organization* **89**, 153–156.

Mills, A., Ataguba, J. E., Akazili, J., et al. (2012). Equity in financing and use of health care in Ghana, South Africa, and Tanzania: Implications for paths to universal coverage. *Lancet* **380**, 126–133, doi:10.1016/S0140-6736(12)60357-2.

Mills, A. J. and Ranson, M. K. (2005). The design of health systems. In Merson, M. H., Black, R. E. and Mills, A. J. (eds.) *International public health: Diseases, programs, systems and policies*, 2nd ed., pp. 515–558. Boston: Jones and Bartlett Publishers.

Ooms, G., Decoster, K., Miti, K., et al. (2010). Crowding out: Are relations between international health aid and government health funding too complex to be captured in averages only? *Lancet* **375**, 1403–1405.

Tangcharoensathien, V., Patcharanarumol, W., Ir, P., et al. (2011). Health-financing reforms in Southeast Asia: Challenges in achieving universal coverage. *Lancet* **377**, 863–873.

WHO (2010). The World health report: Health systems financing. *The Path to Universal Coverage*. Geneva: World Health Organization.

Xu, K., Evans, D. B., Kawabata, K., et al. (2003). Household catastrophic expenditure: A multicountry analysis. *Lancet* **362**, 111–117, doi:10.1016/S0140-6736(03)13861-5.

Relevant Websites

http://www.gapminder.org/data/
Gapminder.

http://apps.who.int/nha/database/PreDataExplorer.aspxd=1
WHO Global Health Expenditure Database.

http://apps.who.int/ghodata/
WHO Global Health Observatory Data Repository.

Health Status in the Developing World, Determinants of

RR Soares, São Paulo School of Economics, FGV-SP, São Paulo, SP, Brazil

Glossary

Demographic transition Reduction in mortality and fertility rates experienced by most countries in a certain stage of their development process; it is first accompanied by accelerated population growth, then followed by declining rates of population growth.

Epidemiological transition Process of change in the causes of death that accompany the reduction in mortality observed during the demographic transition, from infectious diseases to other causes of death; it is also accompanied by a change in the age distribution of mortality, from early to older ages.

Germ theory Theory according to which certain diseases are caused by microorganisms; became widely accepted starting in the end of the nineteenth century.

Life expectancy Expected years of life if an individual were subject to the age-specific mortality rates observed at a point in time.

Malthusian mechanism/Malthusian response The Malthusian view of population behavior predicts that, in response to improvements in economic conditions, population growth is increased; population expansion, in turn, leads to a deterioration in living standards – through reduced availability of land per capita, wars, and disease – bringing economic conditions back to their original level; in the Malthusian mechanism, population expansions always perform the necessary adjustment, leaving no room for long-run improvements in living standards.

Introduction

Until the seventeenth century, world population behavior was governed by a straightforward Malthusian mechanism: sporadic technical advances and favorable climatic conditions reduced mortality via relaxation of the constraints imposed by the supply of goods; these would then lead to increased population, which would then reverse the movement, bringing standards of living back to the limits of simple reproduction. Mortality rates had great variability with no clear trend and, by the Year 1600, life expectancy was probably about the same as it had been 2000 years before. These Malthusian responses following positive permanent shocks explain a timid but persistent population growth, despite the trendless behaviors of mortality and fertility rates.

This pattern started to break down for some Western European and Scandinavian countries in the eighteenth century. Mortality rates fell (life expectancies increased) without any indication that a countervailing Malthusian mechanism was at work. Population growth for these countries increased, reaching a peak in the mid-nineteenth century, after which, as a consequence of fertility declines, growth rates started coming down. This pattern was followed closely by, among others, the United States and Canada and, by the beginning of the twentieth century, this group of countries had populations larger than they ever had before, together with health and life expectancy levels unprecedented in human history.

This transformation marked the onset of the demographic transition and was an essential part of the process of economic development that continued spreading unabated through most of the world until today. See article by Ebenstein in the same section of the Encyclopedia for a longer discussion of the demographic transition. This revolution, however, took some time to reach the developing countries. It was only after World War I that mortality levels began to decline in the poorer regions of the world. Nevertheless, in these areas, the process took place at a much faster pace and at much lower income levels than it had in Europe and North America. Renewed and persistent mortality reductions throughout most developing regions after World War II changed the face of human societies and led to the population explosion observed during the twentieth century.

These health improvements played a central role in the history of population growth. A strand of theoretical literature also argues that they were a potentially important force determining the reductions in fertility observed at later stages of the demographic transition, as well as the increases in human capital and growth registered thereafter. Nonetheless, the precise causes of the improvements in health and reductions in mortality in the developing world are not yet entirely understood.

In this article, the available evidence on the determinants of health and mortality in developing countries is reviewed. The next Section Patterns of Health and Mortality starts with a discussion of some historical patterns and aggregate studies. Following that, the results from a vast array of studies analyzing various dimensions of potential determinants of health and mortality are summarized. Finally, the Section Discussion concludes with a synthesis of what is known up to now and some general remarks.

Patterns of Health and Mortality

Perhaps the most striking feature of the improvements in health in the developing world is how they became increasingly dissociated from gains in income or overall improvements in individual living conditions. This is most clearly seen in the so-called Preston curve, which portrays the relationship between income per capita and life expectancy across countries. Figure 1

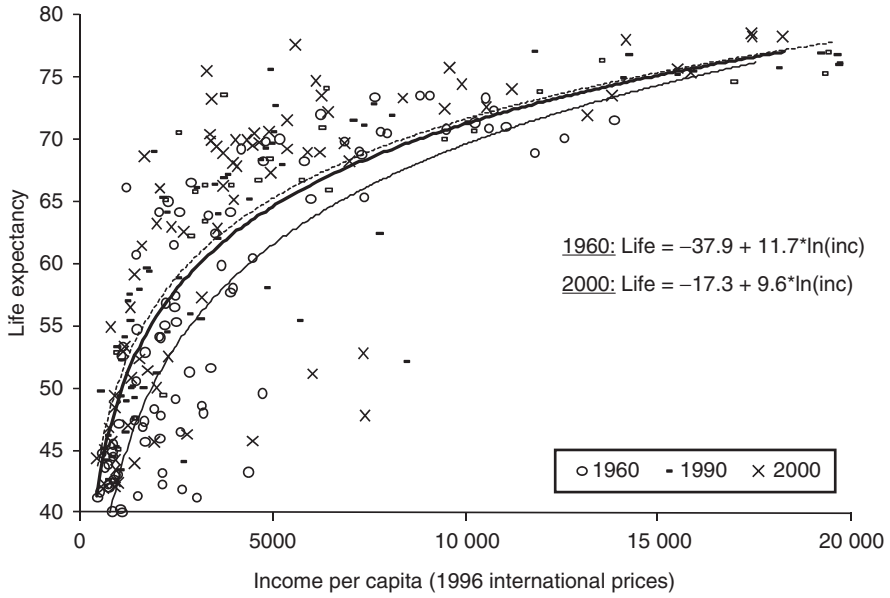


Figure 1 The changing relationship between income and life expectancy; 1960, 1990, and 2000 (Soares, 2007a).

reproduces this curve for the years 1960, 1990, and 2000. There is a positive correlation – close to logarithmic – between income per capita and life expectancy at each point in time. But this relationship has been shifting since the beginning of the twentieth century. This pattern was first noticed by Samuel Preston, who compared data between 1930 and 1960, and has persisted through several decades. In other words, countries at a given income level in 2000 experienced much higher life expectancies than countries at comparable income levels in 1960. From a historical perspective, this amounts to saying that a significant fraction of the gains in life expectancy over the last century were unrelated to changes in income.

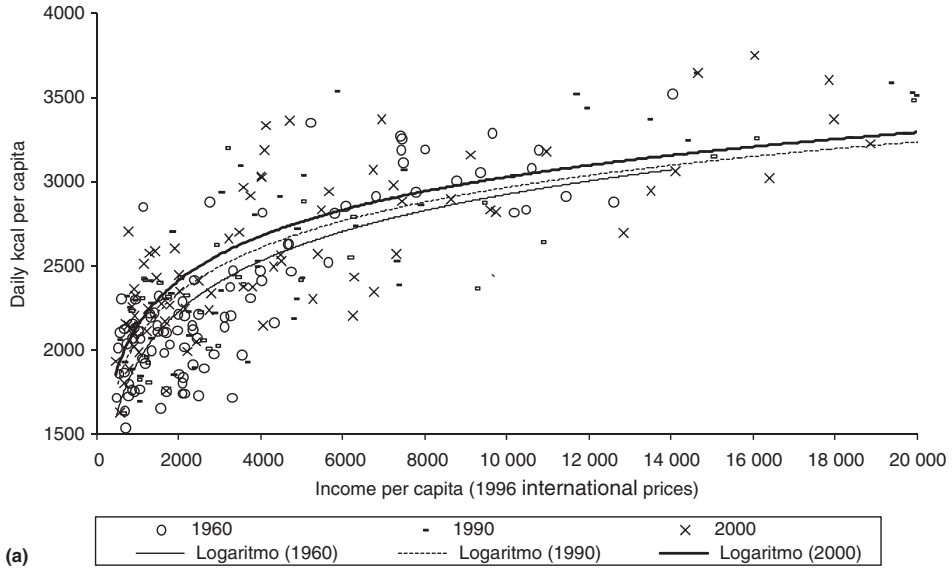
In addition, these gains have been particularly strong for countries at lower income levels. This pattern led to reductions in life expectancy inequality in the postwar period: by any measure, inequality in life expectancy declined substantially after 1960, apart from a mild increase after 1990 due to the arrival of HIV/AIDS. Despite different patterns of access to water, sanitation, education, income, and housing in developing countries, there was a surprising stability and homogeneity in this process of mortality reduction in the postwar period.

The evidence also shows that the shift of the Preston curve is not an artifact of a falling price of food and improved nutrition at constant levels of income. Preston classifies countries in different nutrition and income brackets and compares data from 1940 and 1970. He shows that life expectancy gains took place at constant levels of income and nutrition. Even for the lowest nutrition group (<2100 cal daily), he identifies an increase of 10 years in life expectancy at birth.

Figure 2 shows the same pattern. At constant levels of income, nutrition does seem to have improved slightly between 1960 and 2000. This may be the result of technological improvements and declines in the relative price of food. Nevertheless, it is far from enough to explain the shift in the income–life expectancy profile: the cross-sectional relationship between nutrition and life expectancy at birth shifted in much the same way as the cross-sectional relationship between income and life expectancy. Between 1960 and 1990, at constant nutritional levels, life expectancy at birth rose by as much as 8 years. In a cross-country econometric analysis relating life expectancy improvements to income and caloric consumption, Preston concludes that approximately 50% of the changes in life expectancy between 1940 and 1970 were due to 'structural factors,' unrelated to economic development or nutrition. Other research finds similar results for the period between 1960 and 2000.

The evidence also suggests that this is not an artificial result due to aggregation and within country changes in the distributions of these variables. In the case of Brazil, for example, municipality-level data between 1970 and 2000 show a within country shift in the cross-sectional relationship between income and life expectancy that is similar to that observed across countries. At constant levels of income, life expectancy typically rose by more than 5 years, meaning that at least 55% of the improvements in life expectancy in Brazil during these 30 years seemed to be unrelated to gains in income per capita. Similar evidence is also available for Mexican states.

Analogous conclusions were generated by other studies in very different settings. Mortality changes in Latin America between 1950 and 1990 show that mortality does respond to short-term economic crisis but that these responses are very small and quantitatively irrelevant when compared with historical changes (though morbidity changes may be substantial). The classic concept of 'mortality breakthroughs' itself was based on historical experiences of improvements in health that were not related to growth in income per capita. Several other researchers present various arguments and evidence indicating that the relationship between income, nutrition, and mortality is far from enough to explain the improvements in health and life expectancy observed during the twentieth century.



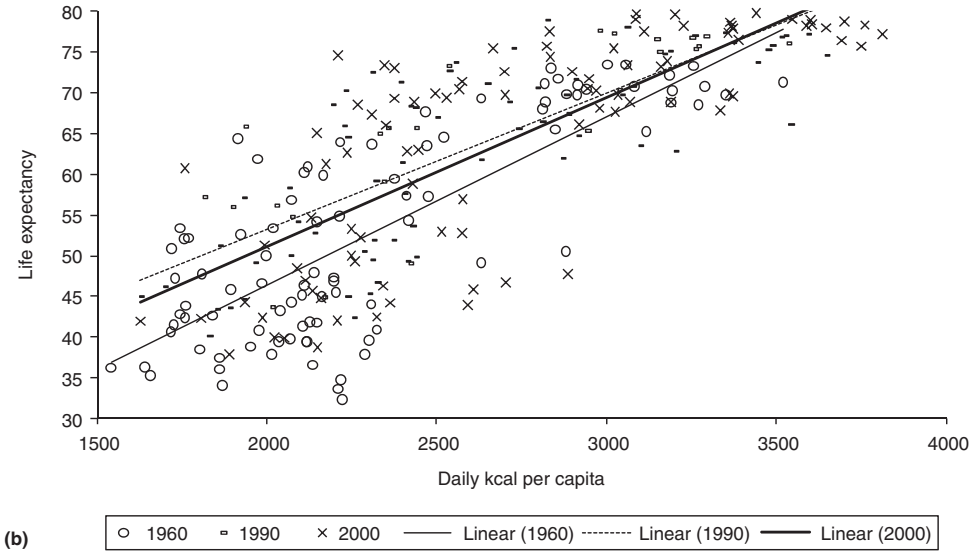


Figure 2 The relantionship between income, nutrition, and life expectancy; 1960, 1990, and 2000 (Soares, 2007a).

The question remains, therefore, as to what were the factors that determined these improvements in health, mostly independently of individual standards of living. Further insight in this matter can be obtained by looking into the profile of changes in the distribution of mortality by age and cause of death. This pattern of changes in the age and cause-distribution of mortality is usually referred to as the 'epidemiological transition,' a term first coined by Abdel Omran. It describes the process of change in leading causes of death, from infectious diseases to chronic nontransmissible diseases, that takes place as mortality reductions progress. There is also an accompanying shift in the age distribution of deaths, from younger to older ages, until child and infant mortalities converge to close to zero.

There is a wealth of information on the epidemiological transition experience of some developed countries. For the nineteenth century US, for example, infectious diseases were responsible for 45% of all deaths between the ages 0 and 4 years, with birth-related and childhood diseases accounting for an additional 30%. Improvements in the period were driven mainly by the acceptance of the germ theory, leading to the boiling of milk and sterilization of bottles, hand washing, isolation of the sick, etc. During the first half of the twentieth century, infectious diseases were still the leading cause of death, and nutrition and public-health infrastructure were the main determinants of improvements in health (reduced deaths from infectious diseases were responsible for three-quarters of the gains in life expectancy in the period). Between 1940 and 1960, infectious diseases continued to play a role, but medical innovations (antibiotics) became increasingly more important (health improvements concentrated on diseases for which new drugs became available). Finally, after

Table 1 Diseases responsible for mortality declines in less developed countries (LDCs) and methods that have been used against them, 1900–70

Dominant mode of transmission	*Diseases*	*Approximate % of morality decline in LDCs, accounted for by disease*	*Principal methods of prevention deployed*	*Principal methods of treatment deployed*
Airborne	Influenza/pneumonia/ bronchitis	30		Antibiotics
	Respiratory tuberculosis	10	Immunization; identification and isolation	Chemotherapy
	Smallpox	2	Immunization	Chemotherapy
	Measles	1	Immunization	Antibiotics
	Diphteria/whoopinh cough	2	Immunization	Antibiotics
	Subtotal	45		
Water, food, and fecesborne	Diarrhea, enteritis, gastroenteritis	7	Purification and increased supply of water; sewage disposal; personal sanitation	Rehydration
	Typhoid	1	Purification and increased supply of water; sewage disposal; personal sanitation, partially effective vaccine	Rehydration, antibiotics
	Cholera	1	Purification and increased supply of water; sewage disposal; personal sanitation; partially effective vaccine; quarantine	Rehydration
	Subtotal	9		
Insectborne	Malaria	13–33	Insecticides, drainage, larvicidesInsecticides, partially effective vaccines	Quinine drugs Antibiotics
	Typhus	1		
	Plague	1	Insecticides, rat control, quarantine	
	Sub-total	15–33		

Source: Preston (1975).

1960, mortality reductions shifted toward more sophisticated and technologically intensive medical advances, concentrated at old ages and on conditions such as heart and circulatory diseases.

The historical evidence from England shows a similar pattern. A relatively small number of infectious diseases account for the entire improvement in life expectancy observed in England and Wales between 1837 and 1900. Some interpretations argue that changes in nutrition were the main determinant of changes in susceptibility to these diseases, but others give more credit to public policy (mainly sanitary reforms, perhaps responsible for 25% of the reductions in mortality in the period). Infectious diseases accounted for 68% of the overall reductions in mortality in England up to the 1950s.

A similar path was followed by developing countries in the second half of the twentieth century. Preston was the first to try to map the reductions in mortality in the developing world between 1900 and 1970 into different causes of death. **Table 1** presents the approximate fraction of mortality reductions in less-developed countries accounted for by different diseases. Preston argues that preventive measures associated with public-health programs and infrastructure were probably the main determinants of the changes portrayed in the table (apart from the case of influenza, pneumonia, and bronchitis). Large-scale immunization, cleaning of water systems, and sewage disposal are examples of changes that took place in several less-developed countries throughout the period. This interpretation would suggest that approximately 50% of the life expectancy gains in the period were unrelated to simple improvements in material conditions.

Evidence for Latin America between 1955 and 1973 suggests that dimensions unrelated to living standards were more important in regions where malaria was endemic, and where other infectious diseases were more prevalent. According to this view, approximately 55% of the reductions in mortality would be attributable to factors not directly linked to improvements in living conditions.

The discussion from the previous paragraphs hints at a relationship between mortality by cause of death and available methods of prevention and treatment. Similarly, mortality by cause of death is intimately linked to mortality by age, and to the stage of a specific society in the process of epidemiological transition. At a given historical moment, both of these are associated with the health technologies available and employed in each particular case. For these reasons, the historical profile observed in developed countries is analogous to the cross-country gradient observed in the postwar period.

Analogously, mortality reductions experienced by developing regions in the past 40 years, for example, are very similar to those experienced by the US in the beginning of the twentieth century.

The pattern of cause and age-specific life expectancy gains across different development levels between 1965 and 1995 illustrates this point. In poorer regions (Middle East and North Africa), life expectancy gains are almost entirely concentrated on infectious diseases of the respiratory and digestive tract, and congenital anomalies and perinatal period conditions. As a result, 90% of the mortality reductions are concentrated at younger ages. As the development level increases, mortality shifts continuously from early to old ages (following, in sequence, Latin America and the Caribbean, East Asia and the Pacific, Europe and Central Asia, and North America). For the most developed regions, 60% of the life expectancy gains are due to heart and circulatory diseases and nervous systems and senses organs conditions, all concentrated in old ages.

Historical trends and cross-country profiles within countries suggest a specific process of health improvements and mortality reductions. This process mimics the movement of a country through the different stages of the epidemiological transition. Still, there is no consensus as to the specific factors that determined these improvements in health in each different circumstance. In the next Section Evidence on Determinants of Health Improvements, to shed some light on the issue, the evidence on the determinants of mortality reductions in specific contexts is discussed.

Evidence on Determinants of Health Improvements

The evidence discussed in the Section Patterns of Health and Mortality suggests that 'structural factors,' not directly related to economic development, were responsible for a substantial fraction of the recent reductions in mortality in developing countries. Substantial reductions in mortality were observed at very low income levels and with minimal expenditures on health, so it is believed that diffusion of new technologies must have played a role.

New technologies may come into play as determinants of health through various channels. First, in some dimensions, health is the outcome of household production (personal hygiene, handling and preparation of food, treatment of water, etc.). From this perspective, new technologies are incorporated through absorption of knowledge by individuals. This is probably particularly important at very low levels of development (or high levels of mortality).

Second, some health technologies have a major public good component. Ideas and knowledge are extreme examples of this. Once the germ theory became accepted, for example, its main implications became publicly available to all agents. In more specific health technologies, externalities and traditional public goods are also very important (development of new medicines, water and sewerage systems, vaccination campaigns, environmental regulations, etc.). Sometimes implementation involves large fixed costs and low marginal costs, other times adoption depends on the outcome of a centralized political process. Changes are, to a great extent, outside the control of any individual agent in society and, given its political and technological nature, may be even considered exogenous to the economic conditions faced by a country.

Therefore, the diffusion of health technologies in developing countries over the last century was most likely driven by the absorption of knowledge by agents and public provision, rather than by the same factors determining diffusion of technologies associated with the production of private goods. This is particularly important for changes in mortality observed at low levels of development, when improvements can take place even with minor expenditures on health. This logic points to particular candidates as main determinants of the health improvements discussed in the Section Patterns of Health and Mortality. These are associated with diffusion of pure nonrival and nonexcludable knowledge, public or international interventions related to public-health infrastructure and to particular diseases, and family and community health programs focused on health practices.

Perhaps the clearest example of the role of technology and public good provision is the United Nations' Expanded Program on Immunization (EPI). The program started in 1974 with the objective of extending worldwide access to vaccines against measles, diphtheria, pertussis, tetanus, tuberculosis, and polio, among others. In countries covered, the EPI led to major increases in immunization rates within few years, while infection rates dropped abruptly. Among other things, the program led to virtual eradication of polio from the Americas in 1994, and raised immunization for the six target diseases from 5% of the world's newborns in 1974 to approximately 80% in 2000.

Another example of a successful intervention against particular conditions is the case of Malaria. In Sri Lanka starting in 1945, dichlorodiphenyltrichloroethane (DDT) became available, leading to the elimination of mortality differentials between endemic and nonendemic areas, and to fast declines in mortality rates. Malaria control contributed with 23% of the observed reduction in death rates up until 1960. From 1946 to 1950, malaria is estimated to have contributed with one-third of the total reduction in mortality. Similar results from other malaria control programs have been documented in countries such as Guyana, Guatemala, Mexico, Venezuela, and Mauritius.

A very important coordinated effort was the World Health Organization (WHO) campaign launched in the 1950s to eradicate malaria. The campaign counted on WHO's technical support and was partially funded by USAID and UNICEF. It was based mostly on DDT spraying, with the objective of breaking up the transmission of malaria for long enough so that the pathogen would eventually die, coupled with some medical assistance. Analyses of the experiences of Brazil, Colombia, and Mexico indicate that in all three cases the campaign was followed by large declines in malaria prevalence. In Colombia, prevalence rates fell by approximately 80%. Overall, however, for Latin America as a whole, the campaign proved ineffective in eradicating malaria, with partial resurgence observed some decades after the initial intervention. Nonetheless, even in these cases, prevalence was never again comparable to the preintervention levels.

A view sometimes presented as a competing alternative in the demographic literature postulates that focused

interventions have limited effects, and that the main driver of good health in developing countries is a set of 'appropriate' social and political conditions. This has been argued to be the case, for example, in the three famous experiences of 'breakthroughs' in mortality reduction: Kerala (India, 1956–66), Sri Lanka (1946–53), and Costa Rica (1970–80). These three cases were also exceptional in their social and political environments, and in their effectiveness in providing inputs in the areas of education, health services, and nutrition. Female autonomy, open political systems (competition), large civil society without rigid class structure, and national consensus related to policies are highlighted as factors allowing the adoption of health inputs and the absorption of new technologies. In Sri Lanka, cholera was contained in the 1870s through quarantine measures and construction of water systems, whereas neonatal tetanus was cut down by the systematic use of midwives. From 1910 on, successful campaigns against diarrhea, respiratory infections, and hookworm stressed the need for public health, sanitation, and personal hygiene. Other important events included a malaria campaign started immediately after the war (using DDT) and the popularization of penicillin and sulfa (sulphonamide) drugs. Health expenditures were never more than 1.5% of gross domestic product, despite profound improvements in public health. In Kerala, the mortality breakthrough took place between 1956 and 1966, when deaths from cholera and smallpox were drastically reduced. Extensions of public-health programs and immunization – through provision of community level services – are identified as the proximate reasons behind these mortality reductions. Costa Rica, in turn, increased expenditure on health services leading to major health improvements between 1970 and 1980. Easy access to community-level services – coupled with immunization campaigns – were also identified in this case as important factors in the reduction in infant and child mortality. The case of Jamaica (which had life expectancy greater than 75 in 2000) also fits well in the above logic: women were historically more independent, schooling developed early, and there was a tradition of discussion of political issues. In Jamaica, school teachers were trained to be health educators, coaching people on how to recognize and treat themselves against specific diseases and vectors.

The important role of easy access to primary health care and family planning, sometimes combined with other interventions, is highlighted in various studies. Data from 16 years of operation of the International Centre for Diarrheal Disease Research (Matlab Thana, Bangladesh), between 1966 and 1981, provide evidence on the effect of family planning, tetanus vaccine, and oral rehydration therapy. The data suggest that tetanus vaccine (given to pregnant women) reduced newborn 4–14 day mortality by 68%. A broad program of family planning was estimated to be responsible for a 25% reduction in death rates, with rehydration therapy accounting for another 9%.

The Brazilian Family Health Program, implemented in the 1990s and expanded during the 2000s, provides additional evidence on the role of family and community based health interventions. The program was largely based on preventive care, but evidence shows that coverage also affected breastfeeding and immunization, and improved maternal management of diarrhea and respiratory infections. It was particularly effective in improving health at early ages and reducing deaths from perinatal period conditions and infectious diseases, and it was also associated with improved subjective assessments of health status.

The extreme experience of reduction in maternal mortality in Sri Lanka is also an important example. In Sri Lanka between 1946 and 1953, there was a reduction of 70% in maternal mortality rates, from 1.8% to 0.5%. This reduction is thought to have been the consequence of changes in various health policies associated with increased access to health centers, midwives, and hospitals (and possibly also with introduction of sulfa drugs and penicillin).

The historical experience of Cuba is yet another case supporting the role of community and family based interventions. US occupation of the island between 1898 and 1902 initiated a series of sanitary reforms, culminating in the virtual elimination of yellow fever, as well as reductions in mortality from tuberculosis and other infectious and parasitic diseases. In some cases, such as tuberculosis, health improvements seem to have been due to better economic conditions and nutrition, combined with the introduction of antibiotics after the 1940s. Other infectious and parasitic diseases – such as diphtheria, malaria, diarrhea, gastritis, and enteritis – were more directly affected by specific sanitary and public-health measures and efforts to teach proper infant care (supposedly accompanied by improvements in education). Nevertheless, some researchers point out that improvements in education, urbanization and targeted health programs occurred early in the twentieth century, whereas a major fraction of the progress in life expectancy was observed only long after that. Therefore, the authors suggest that the role of easy access to primary health care should be even larger than that initially suggested.

Also in the case of Costa Rica between 1968 and 1973, access to medical care (proportion of births under medical attention) had a substantial impact on child mortality. Still, as it relates to improvements in health overtime, education, and sanitation appear as important driving forces. One study shows that the same trend of health improvement continued in Costa Rica after 1970 and suggests that factors similar to those highlighted in the previous period played a role in this later experience. For rural India, data between 1973 and 1978 show that, together with mothers' literacy, type of birth attendant and triple vaccination were closely related to regional variations in child mortality. Poverty and medical care received at birth emerged as central for neonatal mortality, whereas availability of medical facilities and immunization coverage were the main correlates for postneonatal mortality.

Public-health infrastructure, combined with education, also appears as an important determinant of health improvements in various other contexts. Sanitation and women's education were the most important factors determining child mortality differences in Guatemala between 1959 and 1973. For the case of Brazil between 1970 and 2000, education and sanitation were also the key determinants of changes in child mortality, whereas access to clean water, in addition to education and sanitation, appeared as an important determinant of life expectancy at birth.

Access to clean water, again together with women's education, appears as an important determinant of health

outcomes in several papers. This is the case in the experience of Malaysia between 1946 and 1975, where mothers' education and piped water were the factors most closely associated with child mortality (sanitation also appears as marginally relevant), as well as for Brazil. In particular, data between 1970 and 1976 have been used to track down the effects of a program that targeted the improvement of urban environmental conditions (PLANASA), showing that parents' education and access to piped water were the factors most closely related to child mortality both in 1970 and 1976 (access to piped water explained one-fifth of regional differentials in child mortality). Some evidence on the importance of water quality comes from the Argentina, where researchers have explored improvements in the quality of water provision following the privatization of local water companies in approximately 30% of Argentina's municipalities. The results show a reduction of 8% in child mortality (mostly from infectious and parasitic diseases) in areas that had their water services privatized (the reduction increases to 26% in the poorest areas).

The evidence from the historical experience of the US also lends support to the potential role of clean water technologies in developing countries. It was estimated that clean water technologies were responsible for 43% of the reductions in mortality in major American cities during the early-twentieth century. For infant mortality, this share is estimated to rise to 74%, whereas for typhoid fever, clean water is thought to have led to virtual eradication.

For some other dimensions, there is no evidence available from developing countries. In some of these cases, the historical evidence from the developed world may also be informative. Regarding the role of new drugs, for example, there is evidence on the case of the introduction and diffusion of sulfa in the US after 1937. The prevailing view from the literature is that medical innovations played a small role in US mortality declines between 1900 and 1950, but the introduction of sulfa drugs in the mid-1930s represented the development of the first effective treatment of various bacterial infections, including scarlet fever, puerperal sepsis, erysipelas, pneumonia, and meningitis. The available literature suggest that the arrival of sulfa drugs was responsible for declines of 25% in maternal mortality, 13% in mortality from pneumonia and influenza, and 52% in mortality from scarlet fever, amounting to between 40% and 75% of the total decline in mortality from these causes of death during the period. Similarly, the episodes of eradication of hookworm diseases in the American South show how powerful the use of drugs (deworming medicines) coupled with educational campaigns (on how to recognize symptoms) can be. Infection rates among children, which were approximately 40% in 1910, dropped to nearly zero after an intervention sponsored by the Rockefeller Sanitation Commission.

Discussion

The evidence on the determinants of mortality and health in developing countries from the microliterature is very diverse in nature, focus, and methodology. Still, it does reveal some repeated patterns.

First, interventions targeted at particular conditions (malaria, tetanus, diarrhea, large-scale immunizations, etc.) have shown sustained success in improving health and reducing mortality. This debunks the once common argument that narrow approaches focused on specific technologies may end up simply increasing mortality from competing causes of death, and not lead to sustained improvements. The evidence suggests just the opposite: in the case of malaria and measles eradication in Guyana, Kenya, Sri Lanka, Tanzania, and Zaire, the implementation of targeted programs led to reductions in mortality systematically larger than the direct reduction in the cause of death that constituted the initial target. Reductions in mortality from one cause of death, in reality, seem to lead through synergistic links to reductions in mortality also from other causes. This should be expected when one type of disease increases individuals' susceptibility to infections and other diseases (due to weakened immune system or reduced capacity to absorb nutrients).

Still, family health programs and other broad-based community interventions, taking into account the scope of social specificities of local populations, also seem potentially relevant. This was the case with successful programs implemented in Bangladesh and Brazil, and also with some dimensions of the Jamaican experience. Disease-specific targeted interventions and broad programs focused on health practices and the cultural context, rather than being mutually exclusive alternatives to explain health improvements in the developing world, are likely to be both relevant in explaining the diversity of experiences observed. The ideal program in each particular case seems to be a function of the incidence of endemic conditions for which specific interventions are available, as compared to the incidence of conditions that can be minimized through improvements in individual or collective health practices.

Second, in relation to the role played by specific factors, there is an overwhelming amount of suggestive evidence pointing to the importance of education as a determinant of child health. Part of this relationship reflects the effect of income on health, but studies controlling for socioeconomic status still found robust correlations between mother's education and child mortality. Irrespectively, even if taken as causal, this relationship is not yet fully understood in the literature. Some suggest that parental education leads to more use of medical care and sanitary precautions, better understanding of nutritional information, and better recognition of serious health conditions. One study, for example, shows that mothers' literacy is associated with type of medical care during birth and in the postneonatal period. Still, the effect of parental schooling may be more related to modernization and indoctrination. Schooling could be a mechanism to familiarize the population with modern values, reducing resistance to formal medical attention and medicines. A review of a vast array of evidence concludes that educated mothers are better informed about and more likely to use medical facilities and other health technologies, are more likely to have their children immunized and to have received prenatal care, and are more likely to have their deliveries attended by trained personnel. At the same time, the social aspects in the relationship between education and child mortality were also present: educated mothers marry later, tend to have fewer children, and

to invest more in each child. Overall, the following channels linking mother's education to child mortality were identified: greater cleanliness, increased utilization of health services, greater emphasis on child quality, and enhanced female empowerment.

The role attributed to public-health infrastructure can be analyzed through the results related to access to clean water and sanitation. Some microstudies emphasize one of these dimensions in detriment of the other, maybe due to the high correlation between them, and few papers have been able to identify independent effects of each. But many of the analyses discussed here find a significant correlation between either sanitation, or access to clean water, and health (in most cases, mortality). Anecdotal evidence from Cuba and Kerala, among others, also supports the potential importance of factors linked to public-health infrastructure in triggering sustained improvements in health.

From a broad perspective, the evidence does not point to one specific factor as the main determinant of health status and mortality in developing countries. There is strong evidence on the success of targeted interventions in some contexts, such as malaria control, rehydration therapy, and immunization, whereas there are also various qualitative and quantitative studies indicating that family and community health programs can be effective, by reducing the probability of infections and improving health management. Finally, there is also evidence on importance of health infrastructure, through access to clean water and sanitation. Based on the evidence currently available, it is still impossible to isolate the specific role of each of these factors, or to identify their relative importance in different contexts. These would be important goals for future research in the area.

See also: Education and Health in Developing Economies. Fertility and Population in Developing Countries. Global Public Goods and Health. Infectious Disease Externalities. Nutrition, Health, and Economic Performance. Water Supply and Sanitation

References

Preston, S. H. (1980). Causes and consequences of mortality declines in less developed countries during the twentieth century. In Easterlin, R. S. (ed.) *Population and economic change in developing countries*, pp. 289–341. Chicago: National Bureau of Economic Research, The University of Chicago Press.

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies* **29**(2), 231–248.

Soares, R. R. (2007a). On the determinants of mortality reductions in the developing world. *Population and Development Review* **33**(2), 247–287.

Further Reading

Becker, G. S., Philipson, T. J. and Soares, R. R. (2005). The quantity and quality of life and the evolution of world inequality. *American Economic Review* **95**(1), 277–291.

Caldwell, J. C. (1986). Routes to low mortality in poor countries. *Population and Development Review* **12**(2), 171–220.

Fogel, R. W. (2004). *The escape from hunger and premature death, 1700–2100 – Europe, America, and the third World*. Cambridge: Cambridge University Press. 191 p.

Hill, K. and Pebley, A. R. (1989). Child mortality in the developing world. *Population and Development Review* **15**(4), 657–687.

Hobcraft, J. (1993). Women's education, child welfare and child survival: A review of the evidence. *Health Transition Review* **3**(2), 159–173.

Livi-Bacci, M. (2001). *A concise history of world population,* 3rd ed. 251 p. Malden: Blackwell Publishers.

Omran, A. (1971). The epidemiological transition: a theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* **49**, 509–538.

de Quadros, C. C. A., Marc Olivé, J., Nogueira, C., Carrasco, P. and Silveira, C. (1998). Expanded program on immunization. In Benguigui, Y., Land, S., María Paganini, J. and Yunes, J. (eds.) *Maternal and child health activities at the local level: Toward the goals of the world summit for children* 1998, pp. 141–170. Washington, DC: Pan American Health Organization.

Riley, J. C. (2001). *Rising life expectancy – A global history.* Cambridge, UK: Cambridge University Press.

Riley, J. C. (2005b). *Poverty and life expectancy.* Cambridge: Cambridge University Press.

Healthcare Safety Net in the US

PM Bernet and G Gumus, Florida Atlantic University, Boca Raton, FL, USA

Introduction

In most developed economies, universal health insurance coverage is standard and healthcare is paid for using insurance that is either mandated for those who can afford the premiums or subsidized through taxes. In the US, however, insurance purchase was not mandated through the 2000s, and almost 20% of the nonelderly had no coverage. People with no or inadequate health insurance often turn to safety net providers when they get sick.

The US does not have a formal safety net, but rather a patchwork of providers including hospitals, federally qualified health centers, local health departments, community health centers, and others. Some of these providers have an explicit mission to serve low-income, uninsured people whereas others fulfill this role as part of broader community benefit activities.

This article discusses the economic issues relating to safety net providers and the lower income population for whom they care. The most fundamental economic barrier faced by the poor is the lack of health insurance. Beyond that, however, the poor often live in rural areas, have language barriers, and often suffer from chronic conditions, making this population more difficult to treat. On the provider side, the need to remain financially viable is often at odds with charitable missions to care for the poor. The Affordable Care Act (ACA) of 2010 aims to make it easier for everyone to get health insurance, removing one of the major barriers to accessing care. Safety net providers, however, are expected to continue playing a vital role in the provision of care to the most vulnerable.

Special Needs of Lower Income Populations

Lower income populations have a number of attributes which can interfere with the efficient and effective delivery of healthcare services. First and foremost, they cannot afford adequate health insurance. They are uninsured, underinsured, or covered by Medicaid; and thus face problems with access and health outcomes. In addition to financial barriers, differences between patients and their providers can interfere with the provision of care. For lower income populations, such barriers include race, ethnicity, and language. Immigrants are especially prone to all three difficulties. Some groups with special needs are more likely to be living in poverty: children, pregnant women, and people with human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS). For the rural poor, geographic access barriers make it even harder to access care.

Insurance Barriers

The most effective safety net may be adequate insurance. Low income and uninsured people generally have poor access to medical care simply because they cannot afford to pay for services. There are areas that lack adequate primary care providers; however, more providers would be attracted to such areas if enough people had insurance. Unfortunately, it is hard to find affordable health insurance for those who do not work for large employers.

Health insurance coverage is associated with better access and better health outcomes. A lack of insurance often delays detection and can complicate treatment. The generosity of the insurance, measured by the physician compensation rates, may also help get patients seen in the right setting at the right time. Patients with insurance offering higher physician payments are less likely to go to hospitals for nonemergency conditions and are more likely to be seen in an ambulatory setting for conditions such as asthma and diabetes.

Even Medicaid, which generally pays providers much less than Medicare or commercial insurers, has improved its access to care for the poor. Although it would seem that expansions to Medicaid would help cover even more people, some research contends that public insurance reduces the demand for private insurance, whereby the more-expensive employer-based private options are crowded out of the market. This does not necessarily mean that the proportion covered by some form of health insurance changes; simply that the proportion covered by Medicaid increases as the proportion covered by private insurance decreases. Medicaid patients can wind up back with private insurers if the state decides to privatize care, whereby the government pays premiums to private insurers. Privatizing public insurance may not, however, save money. Some studies have observed that shifting recipients into Health Maintenance Organizations (HMOs) can result in a net increase in the overall Medicaid spending.

With the implementation of provisions of the ACA in 2010 access to insurance should improve. However, this is not projected to achieve universal coverage as some people may choose to remain uninsured because their income is too high for a subsidy but too low to afford insurance premiums. As higher take-up rates should improve system efficiencies, insurance premiums may drop as more enroll, making coverage even more affordable.

Special Medical Needs

People living in poverty frequently have special medical needs. Children are a significant portion of the poor and they require specialized care. Substance abusers and the homeless are also poor and generally require more mental healthcare. Sometimes, conditions such as pregnancy or HIV/AIDS precipitate a cash drain that leaves people unable to afford insurance in the first place. Maintaining a regiment of treatment can be difficult among lower income populations, hence further complicating care.

Even in an environment structured to meet the specific needs of the poor, the simple economic concepts of efficiency

and effectiveness are still important. Community health centers (CHCs) improve access to primary care for vulnerable populations. If it is easier for patients to get preventative and diagnostic care, then expensive complications are less likely to arise in the future. CHCs are preferred to more-expensive hospital outpatient departments, where services are more intense and it is more difficult to maintain continuity of care.

The consumers themselves are also rational economic agents. Those in need are not necessarily unsophisticated buyers and seem to have a similar propensity to use primary care *in lieu* of emergency care, where it is available. This reinforces the importance of access. Unfortunately, those in need may not always get the highest quality of care. When quality is measured by the ranking of medical training institutions, uninsured patients are treated disproportionately by physicians from lower ranked schools and residencies.

Other Barriers

Any differences between the physician and the patient – race, language, ethnicity, etc – can interfere with effective provision of care. The true nature of a patient's problems can literally be lost in translation, for example, potentially leading to missed diagnoses or delayed treatment. As many immigrants are poor and face such barriers, safety net providers must be capable of addressing a broad range of needs. As the languages supported by a private physician practice might be limited to English and Spanish, a safety net provider might have to offer Mandarin, Creole, Portuguese, etc. Economically, that raises the costs incurred by safety net providers relative to private practice.

There is much political rhetoric implying that immigrants are responsible for a significant share of uncompensated care or government-subsidized care. However, research shows that very little public tax money is spent on undocumented immigrants, who are less likely to use medical services and whose services cost less when used.

Geographic access barriers make it harder for anyone living in rural areas to get to providers. Simple transportation issues can present major logistic challenges for lower income people. Inadequate public transportation makes it hard for patients to keep appointments, increasing the difficulty and cost of executing a regiment of appropriate care. In addition to rural areas, living in an insurance dessert may also lead to bad health outcomes. Even for people who have health insurance, health service quality and access are worse in areas with higher proportions of uninsured people.

Challenges to Providers

Safety net providers face a number of challenges, both clinical and financial, in serving the needs of lower income populations. The patients often require more attention than the average patient, costing the providers more. Reimbursement, however, is often lower for these patients, compounding the financial strain on the safety net. Before going further, it is worth noting that there is no standard definition of 'safety net' providers; it varies from state to state. Many researchers classify safety net providers as those that provide a high ratio of uncompensated care. The financial challenges faced by the safety net providers start with the clinical aspects of care.

Difficult Clinical Care

In addition to problems in communication and transportation, lower income people are more likely to receive care in acute or urgent settings. As they are often uninsured or underinsured, many people living in poverty do not have a family physician. Medical problems are allowed to develop further because patients may hope the problem goes away before spending money to see a physician. Thus, by the time such patients do seek care, the condition is more complex and the severity of illness is greater. Although the poorer patients arrive sicker, safety net hospitals are still more efficient. Had the same mix of patients presented at for-profit hospitals, it may have cost the healthcare system even more.

Limited access to primary care services is not just the result of decisions by lower income patients on whether, where, or when to spend on healthcare. Managed care can indirectly make it harder for patients living in poor areas to access primary care physicians. HMO penetration is associated with limited access to primary care for poor patients. This may be the result of HMO patients crowding out poorer (possibly charity) patients, or it may be the result of HMOs not selling in primary care deserts. To the extent that the ACA reduces the proportion of the uninsured, it may mitigate complications resulting from delayed or forgone care. Once insured, a poor patient's decision to see a doctor is easier and less costly. If they see their primary care physician sooner, ailments can be addressed in a more timely manner, and thus with lower costs and better expected outcomes.

Low or No Reimbursement

In addition to having to care for patients suffering from more advanced conditions, safety net providers are generally paid less. Lower income patients are frequently uninsured or underinsured; either of which leaves the provider with the possibility of nonpayment. Or the patient might be covered by Medicaid, which normally pays less than any other payer. Providers with a disproportionate share of lower income patients will have limited ability to cross-subsidization or cost-shift to better-paying patients. Cost shifting occurs when hospitals use profits from more-generous payers to subsidize uncompensated care. As such, safety net provider cannot subsidize the more expensive care needed by poorer patients with profits from better-paying patients.

Even charitable and not-for-profit providers must obey the laws of economics; to stay in business, they have to at least break even. There is ample evidence that providers respond to financial incentives even when fulfilling their safety net missions. Safety net hospitals reduce their uncompensated care when insurer fees decrease. When Medicaid fees are cut, physician respond not only by seeing Medicaid patients less

often, but also by reducing the time spent when they do see the patient. In both cases, providers are simply responding to lower fees by offering less. Higher Medicaid fees are associated with increases in the number of services, the intensity of services, and the number of private physicians willing to care for Medicaid patients.

By making health insurance easier to obtain, one of the goals of the ACA is to move patients from self-pay to insured, removing reimbursement as a barrier to care.

Profit Motive and Access

The healthcare system in which providers operate does not give much incentive for providing care to uninsured and underinsured, exacerbating the access issues for lower income populations. Simple profit motives explain why for-profit hospitals provide significantly less uncompensated care than do public hospitals. Although for-profit hospitals are expected to provide some level of community benefit, their primary mission is to provide their investors with good returns, making charity care a lower priority. Many for-profit hospitals are affiliated to larger healthcare systems, which may further weaken the ties to one particular local community and their needs. For-profit status does not preclude a hospital from acting as a safety net provider, but it is more common in areas with less market pressure. Even when hospitals appear to be paying more attention to lower income patients, it often takes government financial incentives for charity care to illicit that reaction. Quite simply, for-profit hospitals are duty bound to provide a return for investors, and charity care cuts into profits.

Not-for-profit providers must also devise ways to survive financially. Here, too, it often involves trade-offs wherein market conditions put financial pressures on the providers to limit charity care. Hospitals provide significantly less uncompensated care in markets with higher HMO penetration. Even nonsafety net hospitals provide more uncompensated care in areas with lower levels of hospital competition, perhaps because of greater community expectations. One way that hospitals used to pay for uncompensated care was through cost shifting. However, insurer price pressures have reduced hospital revenues, leaving little surplus from private insurers to cover uncompensated care.

Disproportionate share payments provide an example of multiple financial incentives working at conflicting purposes. By improving reimbursement levels, it became easier for Medicaid patients to access better hospitals and doctors. However, this left safety net hospitals with fewer Medicaid patients, effectively increasing their relative share of uninsured and underinsured, putting them under further financial pressure. Disproportionate share payments are one possible remedy, providing relatively higher reimbursement to hospitals with a higher proportion of Medicaid patients. However, the allocation of such payments is left to state governments, resulting in multiple methods and unclear effectiveness.

The complexity of the healthcare system in the US can even result in unexpected problems associated with something as simple as a policy to expand Medicaid. On the positive side, this kind of broader access to insurance can reduce the need for safety net providers. However, some studies have found that expanding Medicaid resulted in decreased access for the uninsured because financial motives make hospitals more interested in Medicaid patients than charity patients. Furthermore, because higher reimbursements from Medicaid give poor patients access to a broader range of providers, for-profit hospitals seem to be skimming some of the more lucrative patients, such as Medicaid births. With safety net hospitals now losing Medicaid revenues that could have subsidized uncompensated care, what started as an attempt to help Medicaid patients can end up worsening the financial condition of safety net hospitals.

Taking a cue from insurer tools to avoid adverse selection, some hospitals alter their location or product mix to become less attractive to uninsured patients. By eliminating emergency rooms, AIDS units, maternity care, and substance abuse programs – all departments that attract a disproportionate share of nonpaying patients – hospitals can improve their profitability. For-profit hospitals are also located in better-insured areas, which naturally have less need for uncompensated care.

If uninsured patients still find their way to a provider, the latter can minimize losses by simply doing less. Public- and church-owned hospitals consistently provide more uncompensated care than for-profit hospitals, which may use the existence of a public hospital in the area as an excuse to provide less uncompensated care. For-profit hospitals skim profitable patients from all competitors, including safety net hospitals. This often leaves safety net hospitals under an increased financial pressure.

Precarious Future

Safety net providers are toiling under increasingly difficult financial conditions, making it impossible to provide as much care as needed. The safety net is currently inadequate and is increasingly weakening. State and local governments spending on health and hospitals is critical for providing care for the most disadvantaged populations. The recent economic recession has led to significant funding cuts, which generate serious concerns regarding the viability of the safety net systems.

Financial pressures have led many states to subcontract and privatize services. Medicaid HMOs have already been in use for years, yet have not demonstrated the ability to reduce costs. Privatization may not be the sensible financial decision because most commercial plans are not effective in targeting the special needs of the Medicaid population. Furthermore, their for-profit status gives for-profit Medicaid subcontractors conflicting incentives. For example, though it would improve the profitability of a privatized contract, insurer efforts to reduce service volumes could be extremely harmful to Medicaid recipients, many of whom suffer from chronic conditions.

The healthcare system in the US is extremely complex. Politicians, hospitals, physicians, and insurers often make decisions based on incomplete, incorrect, or misinterpreted information. One common belief is that doctors lose money on uninsured patients. In an irony borne out of the

convoluted machinations of a semimarket-based healthcare system, uninsured are likely to pay more for physician services. Virtually no insurer pays a provider's usual and customary fee, but that is what patients with no insurance are charged. Even after allowing for a share of uninsured patients who pay nothing, physicians actually make higher profits on uninsured than they do on insured patients. Put another way, physicians would have higher profits if they only accepted uninsured patients. Yet most physicians and policymakers believe the opposite to be true.

The expanded insurance availability under the ACA will bring many previously uninsured people into the traditional healthcare system, reducing the need for a safety net. Under the ACA, safety net providers are expected to continue playing a vital role as some people will still not be able to afford insurance; but they may be able to afford a reduced cost and a reduced benefit option. Some amount of insurance education will also be needed, perhaps giving safety net providers an expanded advocacy role. Many signing up for newly available insurance plans may be unfamiliar with how to get the most out of their coverage. Safety net providers are already familiar with these patients, so they may be best situated to help them navigate the healthcare system. As noted earlier, safety net providers are attuned to the specific needs of this population. Therefore, even if the ACA allows lower income populations to get care at their choice of providers, their best choice may still be a safety net provider.

See also: Access and Health Insurance. Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity

Further Reading

Baker, L. C. and Royalty, A. B. (2000). Medicaid policy, physician behavior, and health care for the low-income population. *The Journal of Human Resources* **35**, 480–502.

Bazzoli, G. J., Lindrooth, R. C., Kang, R. and Hasnain-Wynia, R. (2006). The influence of health policy and market factors on the hospital safety net. *Health Services Research* **41**, 1159–1180.

Bazzoli, G. J., Manheim, L. M. and Waters, T. M. (2003). U.S. hospital industry restructuring and the hospital safety net. *Inquiry* **40**, 6–24.

Cunningham, P. J., Bazzoli, G. J. and Katz, A. (2008). Caught in the competitive crossfire: Safety-net providers balance margin and mission in a profit-driven health care market. *Health Affairs* **27**, 374–382.

Davidoff, A. J., LoSasso, A. T., Bazzoli, G. J. and Zuckerman, S. (2000). The effect of changing state health policy on hospital uncompensated care. *Inquiry* **37**, 253–267.

Gaskin, D. J., Hadley, J. and Freeman, V. G. (2001). Are urban safety-net hospitals losing low-risk Medicaid maternity patients? *Health Services Research* **36**, 25–51.

Gresenz, C. R., Rogowski, J. and Escarce, J. J. (2007). Health care markets, the safety net, and utilization of care among the uninsured. *Health Services Research* **42**, 239–264.

Hadley, J. and Cunningham, P. (2004). Availability of safety net providers and access to care of uninsured persons. *Health Service Research* **39**, 1527–1546.

Lindrooth, R. C., Bazzoli, G. J., Needleman, J. and Hasnain-Wynia, R. (2006). The effect of changes in hospital reimbursement on nurse staffing decisions at safety net and nonsafety net hospitals. *Health Services Research* **41**, 701–720.

LoSasso, A. T. and Seamster, D. G. (2007). How federal and state policies affected hospital uncompensated care provision in the 1990s. *Medical Care Research and Review* **64**, 731–744.

Marquis, M. S., Rogowski, J. A. and Escarceo, J. J. (2004). Recent trends and geographic variation in the safety net. *Medical Care* **42**, 408–415.

Pauly, M. V. and Pagan, J. A. (2007). Spillovers and vulnerability: The case of community uninsurance. *Health Affairs* **26**, 1304–1314.

Volpp, K. G., Ketcham, J. D., Epstein, A. J. and Williams, S. V. (2005). The effects of price competition and reduced subsidies for uncompensated care on hospital mortality. *Health Services Research* **40**, 1056–1077.

Zwanziger, J. and Khan, N. (2008). Safety-net hospitals. *Medical Care Research and Review* **65**, 478–495.

Health-Insurer Market Power: Theory and Evidence

RE Santerre, University of Connecticut, Storrs, CT, USA

Introduction

The US, like the Netherlands and Switzerland, among other nations, relies primarily on private health insurance to finance and reimburse for medical care. In fact, approximately 64% of the nonelderly US population enrolled in private health insurance plans in 2011. This figure is down dramatically from its height of 76% in the mid-1970s. Some researchers point out that private insurance coverage fell over time because premium hikes have vastly outweighed raises in consumer income even though the aggregate premium elasticity of demand is slightly lower than the corresponding income elasticity. Others claim the Medicaid program crowded out some private health insurance coverage. Still others propose that occupational shifts from traditionally higher coverage manufacturing jobs to lower coverage service sector jobs in the US led to some of the reduction.

Although private health insurance enrollment has declined in the past in the US, many health policy analysts expect it to increase in the future because of the recently passed Patient Protection and Affordable Care Act of 2010. The Act mandates that most US citizens purchase private health insurance, if they are not eligible for public health coverage, or pay penalties. By 2019, nearly 8 million more nonelderly citizens are expected to purchase private insurance directly from health insurers because of the mandate. As a result, a sound understanding of the health insurance product and the current operation and performance of the health insurance industry will take on even more importance in the future.

At its most basic level, health insurance is no different than any other product sold by firms and purchased by consumers. Health insurance policies are sold indirectly to consumers in the form of employer-sponsored health insurance (ESI) or are directly purchased by consumers (DPI). Of those covered by private health insurance in 2011, approximately 88% received their coverage through employers. The ensuing transaction involving the health insurance product boils down to a potential win-win situation where both market participants stand to gain.

In particular, because of the irregularity and infrequency of health-care spending, consumers typically value health insurance because it offers financial security against unexpected losses and thereby moderates swings in their income. Additionally, consumers value health insurance because it provides them with access to expensive medical treatments which they might not otherwise be able to afford out of pocket. Hence, many consumers are willing to give up their premium dollars, even when feeling quite healthy, because that initial cost pales in comparison with the dollar benefits which they expect to receive from their health insurance companies when they unexpectedly enter into a state of sickness.

Health insurers also stand to gain from the market transaction as long as the health insurance premiums charged, at least cover the costs of providing health insurance during the policy period. Costs include the expected medical benefits to be paid out and the expense load that includes claims processing, underwriting, and marketing expenditures, taxes, and profits, less any interest income earned on invested premiums. Expected medical benefits, in turn, capture the dollar amount that health insurance companies expect to reimburse medical care providers, such as hospitals, physician clinics, and drug companies, for treating patients throughout the policy period. Thus, health insurance companies can be viewed as organizations that negotiate medical care contracts with providers; mark them up to reflect expenses, profits, and risk; and then sell those policies to employers and individuals. Within that perspective, health insurance companies are paid for negotiating health-care provider contracts, reimbursing claims, and managing the associated risks, with profits as the reward for successful performance.

It is evidenced from the preceding discussion that health insurance companies simultaneously operate on different sides of two highly intertwined markets – as buyers in the market for medical services and as sellers in the market for health insurance. It is in these important roles as buyers and sellers that health insurers potentially shape the manner in which these two markets operate and perform. As discussed in the Section 'Theoretical Aspects of Health-Insurer Market Power', economic theory generally suggests that markets operate more efficiently when structured in a competitive manner such that individual buyers and sellers act as price takers and possess no market power. But when markets are structured noncompetitively, sellers may wield market power to the detriment of buyers, or vice versa, with inefficiencies potentially arising in either case.

In the case of health insurers, some interesting market dynamics may be involved when markets are noncompetitively structured because of the simultaneous functioning on opposite sides of the medical care (input) and health insurance (product or output) marketplaces. Indeed, against the backdrop of a baseline case where both markets are reasonably competitive, a number of different scenarios can be imagined where either the medical care provider or health insurer possesses market power and the other does not, or both possess market power in the medical services input market.

With these possible market scenarios in mind, the next section of this article reviews the theoretical aspects of market power within the context of the health insurance industry. Once the basic theory is developed, Section 'Empirical Aspects of Health-Insurer Market Power' discusses the empirical aspects of testing for market power effects. Section 'Empirical

Findings Regarding Health-Insurer Market Power' reviews the empirical literature concerning market power effects in the health insurance and health-care industries. Section 'Summary and Conclusion' is the final section for this article.

Theoretical Aspects of Health-Insurer Market Power

To an economist, market power means that a single seller (buyer) can, individually or with a group of other sellers (buyers), raise (lower) the product's (input's) price without losing all of its sales (purchases). Sellers or buyers generally attain some market power when they are few in number and possess relatively large market shares. It must also be the case that some type of industry barrier prevents new sellers or buyers from entering the market because new entrants heighten competition and typically cause an offsetting price adjustment. If these market conditions hold, a few buyers or sellers will account for a dominant share of the industry purchases or sales and hence the seller side or buyer side of the market is considered to be highly concentrated.

In the limit, a single seller of a product or input is labeled as a monopoly, whereas a single buyer of a product or input is considered a monopsony. Given that health insurers simultaneously operate in the medical services input market and health insurance output market, five potential scenarios can be imagined:

1. Both medical care providers and health insurers do not possess market power (competitive case);
2. Medical care providers, as sellers, possess market power, but health insurers do not in the medical services input market (monopoly case);
3. Health insurers possess market power, but buyers (employers or individual consumers) do not in the health insurance output market (another monopoly case) (two other cases are possible (either buyers possess market power but health insurers do not or both buyers and health insurers possess market power in the output market), but their relevancy is questioned, so they are not covered in the following discussion. However, the extension of the analysis to these two cases should be evident);
4. Health insurers possess market power, as buyers, but medical care providers do not in the medical services input market (monopsony case); and
5. Both health insurers and medical care providers possess market power in the medical services input market (monopoly vs. monopsony or bilateral monopoly case).

Figure 1 provides a graphical illustration showing how the various market outcomes compare with the competitive outcome. (See Pauly (1988) and Scherer (1980) for a similar graphical model, although here the monopsonistic buyer also holds monopoly power in the product market.) Also, Table 1 provides a descriptive summary concerning how each of the scenarios compare to the competitive case in terms of price and quantity. In general terms, the positively sloped supply curve reflects that a higher price is necessary to attract increasing amounts of a particular type of medical service or more health insurance coverage into the marketplace. Also in general terms, the downward-sloping demand curve shows that the buyers' maximum willingness-to-pay declines for increasing units of an input such as medical services or output such as health insurance. If the graph represents the product market for health insurance, the demand curve captures how much additional utility consumers receive from increasing amounts of health insurance coverage. If an input market, the demand curve reflects how valuable increasing amounts of the medical services are to a health insurer, which is referred to as the value of the marginal product (VMP). The demand curve declines because of the law of diminishing marginal utility and productivity.

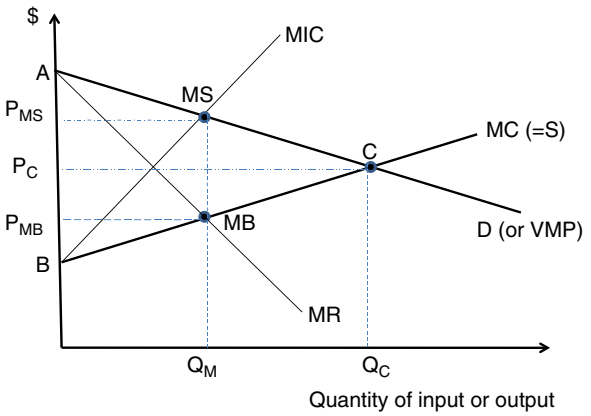


Figure 1 D (or VMP), demand for output in either type of market or demand for an input (value of the marginal product) in a competitive market; MC (=S), marginal cost or perfectly competitive supply curve; MR, marginal revenue of a monopolist in the output market or demand for an input by a monopolist (marginal revenue product); MIC, incremental cost of purchasing faced by a monopsonist; C = outcome when both the buyer and seller sides of the market are perfectly competitive in both the input and output markets; MS, outcome when the seller is a monopoly; MB, outcome when the buyer is a monopsony.

Note that the perfectly competitive equilibrium occurs at point C, where the supply and demand curves intersect. Because both individual buyers (e.g., health insurers or consumers) and sellers (e.g., medical care providers or health insurers) are assumed to be price takers in a competitive market, they each treat the good's price as a parameter – something outside their control. Thus, to maximize net returns – the difference between benefits and costs (net benefits represent profits to firms and consumer surplus to consumers) – sellers match up price to marginal cost (MC), whereas buyers match up price to demand (D) with price serving as the coordination device to equate supply and demand. In equilibrium, price and quantity equal P_C and Q_C, respectively. Buyers receive the triangular area A–P_C–C as 'consumer surplus' and sellers gain triangular area B–P_C–C as 'producer surplus.' Note the win-win aspect of the market transaction.

The two monopoly situations are scenarios (2) and (3). In these two scenarios, the sellers (either medical care providers or health insurers) possess monopoly power but the buyers (health insurers or employers/consumers) do not in the respective market. For a monopolist, theory suggests that the marginal revenue curve (MR) lies below the corresponding downward-sloping demand curve. Marginal revenue lies

Table 1 Summary of scenarios involving buyer and seller market power

Scenario	*Relevant market*	*Market power on buyer side of market*	*Market power on seller side of market*	*Label*	*Equilibrium outcome in Figure 1*	*Implication regarding price and quantity*
1	Output or input market	None	None	Perfect competition on buyer and supplier sides of the market	P_C and Q_C	Competitive price and quantity
2	Medical services input market (e.g., hospital or physician services)	None	Full	Monopoly supplier of medical services	P_{MS} and Q_M	Price of medical services higher and quantity lower than the competitive levels
3	Output market for insurance	Full	None	Monopoly supplier of insurance	P_{MS} and Q_M	Price of insurance higher and quantity lower than the competitive levels
4	Medical services input market (hospital or physician services)	None	Full	Monopsony buyer of medical services	P_{MB} and Q_M	Price and quantity of medical services lower than the competitive level
5	Medical services input market	Full	Full	Monopoly supplier and monopsony buyer	Indeterminate	Price and quantity determined by relative bargaining power

below demand because price must be continually lowered to sell additional units and the revenues from the increased volume fail to compensate for the lower revenues associated with the reduced prices on the previous units. (It is supposed that the demand curve in Figure 1 is captured by the equation $P = a - bQ$, where Q represents quantity and P stands for price. Total revenues equal P times Q or $(a - bQ)Q = aQ - bQ^2$. Taking the first derivative of this revenue function with respect to Q to get dTR/dQ gives marginal revenue equal to $a - 2bQ$. It should be noticed that MR has the same intercept as demand but twice its negative slope.)

To maximize economic profits, the monopolist-seller produces output or supplies an input up to the point where marginal profits are no longer positive (where MR equals MC) and charges the maximum price that buyers are willing to pay for that amount as indicated by the demand curve. Thus, the monopoly equilibrium occurs at MS with a price of P_{MS} and output of Q_M. Note that the monopoly outcome results in a higher price and lower quantity than those predicted by the competitive outcome, *C*. Also, note that consumer surplus shrinks to area P_{MS}–A–MS, whereas producer surplus expands to area B–P_{MS}–MS–MB. The triangular area MS–C–MB represents the competitive winnings that are lost because of the monopolistic restriction of quantity.

Scenario (3) represents a situation where a monopsonist engages in negotiations with a competitive seller side. As a single buyer, the only way a monopsonist can attract additional products or inputs into the market is by paying an increasingly higher price. As a result, if all units are similarly reimbursed when finally purchased, the actual incremental costs of purchasing a particular level of inputs will be greater than the marginal cost, which assumes a price independent of the units purchased. Thus, a monopsonist's incremental cost curve of purchasing inputs or outputs (MIC) lies above the corresponding marginal cost curve (MC) associated with a group of price-taking input buyers. (It is supposed that the supply curve in Figure 1 is captured by the equation $P = c + dQ$. Total costs equal P times Q or $(c + dQ)Q = cQ + dQ^2$. Taking the first derivative of this cost function with respect to Q to get dTC/dQ gives marginal incremental cost, which equals $c + 2dQ$. It should be noticed that MIC has the same intercept but twice the slope of the supply curve.)

To maximize economic profits, the monopsonistic health insurer continues to purchase medical services, as an input, as long as the added revenues, as reflected in *D* (or VMP), compensate for the added costs, as captured by MIC. Thus, in Figure 1, the health insurer purchases inputs up to the point where the MIC and *D* curves intersect. To attract that amount of medical services, the health insurer must pay the price indicated by point MB on the supply curve, S. Compared with the competitive case at point C, it should be noticed that the monopsonistic health insurer pays less for the medical services and purchases fewer units. Thus, in this case, the producer surplus shrinks to B–P_{MB}–MB and consumer surplus expands to P_{MB}–A–MS–MB. Once again some of the social winnings are lost, but this time because of a monopsonistic distortion.

Scenario (5), the bilateral monopoly situation, offers the most intriguing case. Here, a single buyer and a single seller haggle over the terms of the sale. The single seller prefers the MS outcome where seller profits are maximized but the single buyer prefers the MB outcome because buyer profits are maximized. However, it should be noticed that joint net benefits are maximized at the competitive outcome with a quantity of Q_C, that is, both the buyer and the seller can receive more net benefits than at their preferred outcome if they agree on the competitive output and then arrive at a mutually satisfying price to split the resulting winnings. Because neither the buyer nor the seller is able to play off the other by threatening to deal with other buyers or sellers, the resulting price depends on which party possesses a comparative advantage at bargaining or which party brings to the bargaining table something more than the other. For example, one of the

parties may be operating with greater excess capacity, so the increased volume associated with the transaction is relatively attractive and therefore that party is more willing to compromise on the deal.

The exact price that evolves from the negotiation is indeterminate without knowing more about the negotiating skills of the two parties bargaining. It is not known that the upper limit would be the price that forces the buyer's profit to zero and the lower limit would be the price that forces the seller's profit to zero because negative profits would cause one of the firms to drop out of the deal. Alternatively stated, the price must be high enough to make the seller at least as well-off with no sale and low enough to make the buyer at least as well-off with no sale. It should be noted that the alternative to bargaining is no sale because neither the monopoly nor the monopsony outcome is relevant because each entails competitive behavior on one side of the market, which is not a characteristic of bilateral monopoly.

In the real world, often markets are never perfectly competitive and a pure monopoly or monopsony situation, where only one seller or buyer exists, is also rare. A more likely scenario is when a few dominant sellers or buyers exist in some markets and thus these markets are said to be oligopolistic or oligopsonistic. Whether the few buyers or few sellers behave like the preceding models predict depends on whether each individual buyer or seller behaves independently or cooperates with others to extract more favorable prices from the other side of the market. Economic theory suggests that a host of factors influence if a group of sellers (or buyers) act independently or cooperatively. Among these factors are the exact number and relative size distribution of firms, height of any entry barriers, and the availability of close substitute products. These conditions are discussed in detail in the next section.

Empirical Aspects of Health-Insurer Market Power

Researchers have employed various methods when testing for market power effects, but here the reduced-form, structure-conduct-performance (SCP) approach is discussed. Although the SCP approach possesses several empirical shortcomings, it remains the most popular method when testing for market power effects in the health insurance industry. (Other techniques include structural modeling and stock market event analysis.) If suitable data exist, the following estimation equation would be specified to test for market power effects, where X stands for either price (P) or quantity (Q), MCS and MCB represent the market concentration of sellers and buyers, and D and C capture a vector of demand and costs factors, respectively.

$$X = f(MCS,\ MCB,\ MCS \cdot MSB;\ D, C) \qquad [1]$$

According to this monopoly theory, a direct relationship is expected between MCS and P, assuming buyer concentration is negligible and therefore has no separate impact on the market outcome. Under those same conditions, an inverse relationship is anticipated between MCS and Q. Moreover, monopsony theories predict an inverse relationship between MCB and both P and Q, based on low seller concentration. The bilateral monopoly situation, as characterized by the interaction term between the two types of concentration, $MCS \cdot MSB$, is anticipated to be directly related to Q but will have an ambiguous effect on P. Recall that the latter effect depends on the relative bargaining power of the two sides of the market. Finally, the vectors D and C simply act as control variables in eqn [1], so the independent effects of market structure on P or Q can be properly isolated. Variables in D might include buyer income and the price of substitutes and complements, whereas variables in C might include any entry barriers and the state of technology. Thus, this article is not necessarily focused on the impact of those control variables on the dependent variables.

The most basic way to estimate eqn [1] is with the ordinary least squares procedure. (The interested reader will have to consult an econometric text for specifics regarding ordinary least square estimation.) For two reasons, however, ordinary least squares estimation of eqn [1] may result in biased parameter estimates. Both of the reasons deal with some right-hand side variable, or variables, in this case market concentration, being endogenously rather than exogenously determined. First, reverse causality may hold between the dependent and market concentration variables. For instance, more firms may enter the market over time and dilute seller concentration when the market price is high. Or expectations of output, as indicated by Q, may influence seller concentration. Similar examples can be cited for how the magnitude of the dependent variables may influence buyer concentration.

The other problem is that some immeasurable and therefore omitted demand or cost factor may influence both the degree of seller or buyer concentration and the price or quantity. If so, any observed statistical correlation between market concentration and price (or quantity) may only reflect an association rather than a causal relationship because of this third-variable problem. For example, the baseline health of the population may be difficult to measure. Baseline health may influence both the number of hospitals and health insurers within an area as well as the price and quantity of medical care.

Because of the potential for reverse causality or a third-variable problem, estimation of eqn [1] typically requires a panel data set and/or an instrumental variables approach. (A social or natural experiment, which allows for a control group and random assignment of participants, is preferred but the first is expensive to design and the latter is often unavailable to the researcher. See the Appendix to Article 1 in Santerre and Neun (2013) for an elementary explanation of these two approaches.) A panel data set, which covers a number of repeating cross-sections (of individuals, household, states, etc.), allows the analyst to control for unobservable heterogeneity or any omitted variables that remain constant over time. This can be accomplished by including in the estimation equation a 0/1 binary or dummy variable to represent each of the repeating observations. If all omitted variables remain fairly constant over time, the set of dummy variables does a reasonably good job of capturing the fixed differences across observations and thereby corrects for the third-variable problem.

However, the analyst still may have to be concerned with the possibility of reverse causation and any omitted variables that do change over time. For example, the baseline health of

the population may be systematically worsening or improving because of some confounding factor that cannot be easily observed and measured. In this case, an instrumental variable approach should be employed and either implemented on a cross-sectional basis or incorporated within a fixed effects framework. A good instrumental variable is one that is highly correlated with the suspected endogenous right-hand side variable but uncorrelated with the dependent variable.

For example, suppose that the impact of health-insurer buyer concentration on the price of hospital services is empirically examined and assume that the seller side is fairly competitive in all of the hospital services markets under investigation. A good instrument, in this case, is highly correlated with health-insurer concentration but not correlated with the price of hospital services. With that in mind, some researchers have used the size distribution of employers in the market area as an instrument. The reasoning is that health insurance companies may be attracted to areas with more medium- and large-sized employers and employer size is unlikely to directly influence the price of hospital services.

This section has briefly reviewed the technique used by most researchers to test for the market power effects of health insurers as a way of providing some context to the next section that describes the empirical findings. The instrumental variables technique, although econometrically fairly powerful, is often difficult to implement in practice because suitable instruments are hard to find. This is particularly true for studies relating to health care where many variables, such as health status, health insurance coverage, and medical care utilization, are highly interrelated. The researcher must typically be ingenious with respect to uncovering an instrumental variable that influences the suspected endogenous variable but not the dependent variable in the estimation equation. It should be noticed in eqn [1] that at least three instruments may be necessary because both concentration measures as well as their interaction are likely endogenous.

Empirical Findings Regarding Health-Insurer Market Power

To estimate eqn [1], the analyst must identify the degree of market concentration in a particular market. Thus, defining the relevant market area is an important consideration. A relevant market area contains both a product and a geographical dimension. In an output market, the relevant product market considers all of the substitute products that buyers might switch to if any one product's price is raised by a nontrivial amount for a nontemporary period of time. These substitute products may satisfy similar needs or fulfill similar functions. For example, with respect to health insurance, analysts must consider if indemnity plans, health maintenance organizations (HMOs), and preferred provider organizations (PPOs) are substitutes or not. (In the past, researchers treated indemnity, HMO, and PPO plans as separate markets. More recently, the distinction between these plans have become blurred in practice, in part because most health insurers offer multiple products and buyers are willing to switch among products depending on relative prices. Also, many of these health insurance products now contain many features of the others.) In addition, for larger employer/firms, the analysts may consider if self-insured plans are reasonable substitutes for fully insured plans that are purchased from health insurance companies.

Similarly, the relevant geographical output market considers all other locations that buyers might switch to if the price of the product is increased by a significant amount for a meaningful period of time. For some products, the market may be very local in nature, but for others, the relevant geographical market may be regional, national, or even international in scope. Although many health insurers such as Aetna and Cigna operate nationally, most experts agree that the market for health insurance is local in nature because employers and consumers want access to a local network of providers. For example, consumers in Philadelphia wish access to a network of providers in that city so they likely are unwilling to purchase their insurance from a health insurer with provider network established in Boston. Consequently, the geographical market for health insurance is often defined as the metropolitan statistical area (MSA) for research and policy purposes. The important take-away for defining the relevant market area is that current purchasing patterns may not properly reflect the relevant market area because the switching of buyers to new products and locations will not take place until the change in the product's price actually occurs. Thus, one must consider potential substitute products and locations when defining the relevant market area.

Once the relevant market is identified, the degree of market concentration must be assessed. Customary measures of market concentration are the Herfindahl–Hirschman Index (HHI) of market concentration and the number of firms in the market. The HHI is computed by the squaring and adding, in percentage terms, the market shares of all firms in the industry. It ranges from 0 to 10 000 with the latter reflecting only one firm in the market.

The HHI is preferred to other measures such as the concentration ratio, which is an indicator of the percentage of output produced by the industry leaders, because it captures the relative size distribution of output among the leading firms. The value of the HHI decreases with a larger number of equally sized firms, so values closer to zero indicate a less concentrated market. The Federal Trade Commission and Department of Justice considers an HHI more than 2500 as representing a highly concentrated market or a market characterized as a tight oligopoly. In contrast, a market with an HHI more than 1500 but less than 2500 is interpreted as being mildly concentrated or a loose oligopoly. To put these numbers in some perspective, the American Medical Association (2011) reports that the health-insurer HHI is greater than 2500 in most MSAs of the US.

Theoretically, the HHI works best as a measure of market concentration when the products sold by the various firms are reasonably similar. However, when firms sell differentiated products, the HHI loses some of its appeal because niche markets may develop with some firms potentially establishing varying degrees of market power in the various niches. For example, local HMOs may not have a substantial competitive effect on those HMOs possessing a national geographic scope. In this case, the number of firms may provide a better measure

of the degree of market competition because the market takes on features similar to the economist's notion of monopolistic competition. Monopolistic competition holds when a large number of firms offering differentiated products coexist in a market and entry barriers are low or nonexistent. As a point of reference, greater than 200 health insurance companies operate in the typical US state.

Table 2 lists chronologically 17 empirical studies in the economics literature to date regarding the market power effects of health insurers on health-care provider behavior. Note in the table that information is provided for the unit of analysis and method used in each study followed by some abbreviated findings for each article. A number of caveats should be noted. First, although the author(s) may have used an

Table 2 Effect of health-insurer market concentration on provider behavior

Authors	*Unit of analysis*	*Method*	*Findings*
Feldman and Greenberg (1981)	59 BC plans in 1979	IV	Market share of BC plan does not affect hospital discount Discount directly affects market share
Adamache and Sloan (1983)	66 BC plans in 1979	IV	Greater market share directly affects hospital discount
Staten *et al.* (1987)	95 Indiana hospitals in 1983	OLS	BC market share does not affect hospital discount
Staten *et al.* (1988)	110 Indiana hospitals in 1984	OLS	Greater BC market share leads to higher hospital bid price to join PPO More hospital competition leads to greater hospital discounts
Melnick *et al.* (1992)	190 BC of California Network hospitals in 1987	IV	Greater importance of insurer lowers hospital price. Higher hospital prices are observed in more concentrated markets Greater importance of hospital raises hospital price
Foreman *et al.* (1996)	47 individual BC/BS plans during 1986–88	IV	Greater BC/BS market share lowers payments to providers
Brooks *et al.* (1997)	Random sample of more than 290 000 inpatient episodes for over 70 self-insured FFS plans during 1988–92	OLS based on bargaining model	Self-insured firms with a greater presence in a market have greater bargaining power. Greater hospital concentration leads to greater hospital bargaining power
Feldman and Wholey (2001)	Panel data set of all HMOs during 1985–97	IV hospital-FE	Greater HMO buyer power leads to lower hospital prices and greater hospital output HMO buyer power has no effect on the price or output of ambulatory services
Sorensen (2003)	31 hospitals in Connecticut from 1995–98 involving 94 payers (2010 agreements)	OLS hospital-FE based on bargaining model	Increased payer size raises hospital discount. Greater patient channeling of insurers raises discount. Hospitals with fewer rivals lower discount
Dor *et al.* (2004)	Claims data from approximately 80 large, self-insured employers in the 10 largest states of the US in 1995–96	OLS with state-FE based on bargaining model	HMO and PPOs obtain higher discounts than FFS plans for specific treatments and procedures More concentrated hospital services markets result in higher prices for specific treatments and procedures
Younis *et al.* (2005)	1967 hospitals in 1991	OLS	HMO competition has no effect on hospital costs
Bates *et al.* (2006)	306 MSAs in 1999	OLS with MSA-FE	Greater state-wide health insurer concentration leads to increased efficiency of the hospital industry
Bates and Santerre (2008)	Panel data set of 86 MSAs during 2001–4	IV with MSA-FE	Greater HMO concentration leads to more hospital inpatient care Greater PPO concentration leads to more hospital outpatient care
Schneider *et al.* (2008)	42 California counties in 2002	OLS	Health plan concentration has no effect on outpatient prices Physician organization concentration leads to higher physician prices
Dafny *et al.* (2012)	Panel data set of ESI plans enrolling more than 10 million people during 1998–2006	IV with plan-FE	Greater health-insurer concentration leads to a reduction in physician employment and relative earnings
Moriya *et al.* (2010)	National data set of 11 million insured Americans during 2001–3	OLS hospital-FE	Higher state-wide health insurance concentration leads to lower hospital prices. Hospital concentration at the health service area level did not affect hospital prices
Halsersma *et al.* (2010)	1235 unique hospital-insurer pairs during 2005–6 in the Netherlands	OLS based partly on bargaining model	Market shares and concentration of insurers (hospitals) have an inverse (a direct) impact on the hospital price-cost margin

Abbreviations: BC, Blue Cross; BS, Blue Shield; ESI, employer-sponsored insurance; FE, fixed effects; HMO, health maintenance organization; IV, instrumental variables approach; MSA, metropolitan statistical area; OLS, ordinary least squares method; PPO, preferred provider organization.

instrumental variables (IV) approach rather than ordinary least squares (OLS), the actual instrument or instruments used may have been weak in a theoretical or statistical sense. Recall that a good instrument must be correlated with the suspected endogenous independent variable but uncorrelated with the dependent variable. But in practice, some instruments are better at achieving that result than others. As a result, some statistical bias from reverse causality or a third-variable problem may still remain even though an IV procedure is employed if a weak instrument is used.

Second, notice that most of the earlier papers deal with Blue Cross (BC) plans. That early focus likely reflects that BC plans dominated many areas and data were available because most plans were organized on a nonprofit basis at the time. However, since the late 1980s, many BC plans have converted to for-profit status to gain access to equity capital so data have become more proprietary in nature. Third, only a few studies simultaneously control for both insurer and provider market concentration and none allow for an interaction term. Finally, it should be pointed out that some studies are conducted using national data for the US, whereas others are performed with data from particular states or areas.

With these caveats in mind, it appears to be the case that a majority of the relevant studies, reported in **Table 2**, find that a greater dominance of health insurers, as reflected in a higher market share or greater market concentration, results in a lower negotiated hospital price. Thus, it might appear that ample statistical evidence exists to suggest that health insurers possess and exercise market power in the hospital services market (i.e., a movement from point C to MB in **Figure 1**). However, an inverse relation between health-insurer market power and provider prices may not necessarily reflect monopsonistic exploitation, that is, instead of greater health-insurer market power resulting in a movement from point C to MB in **Figure 1**, it may actually be the case that the provider market adjusts from MS to C in response to greater health-insurer buyer pressure. If so, health insurers may actually be exercising monopoly-busting power by forcing dominant hospitals to lower price and produce more services. It follows that empirical evidence is required for both the change in price and the quantity to assess whether health insurers exercise monopsony power in provider markets.

With this perspective in mind, several articles analyze the quantity aspect of health-insurer market power effects. The first study, by Feldman and Wholey (2001), finds that greater HMO market power leads to a lower hospital price but also causes increased hospital output. Bates and Santerre (2008) extend the Feldman and Wholey study by examining the effects of both HMO and PPO market concentration on various measures of hospital output at the MSA level. They find that increased HMO and PPO market concentration leads to a more inpatient and outpatient care, respectively. Finally, Bates *et al.* (2006) find that greater health-insurer market concentration is associated with the hospital services industry using its resources in a more technically efficient manner (i.e., getting more output from the same inputs). These three papers, especially when considered together with the other studies finding lower negotiated hospital prices in response to greater health-insurer market concentration, imply fairly strongly that health insurers exercise monopoly busting rather than monopsony power in the hospital services industry.

However, some limited evidence suggests that the situation may be different in the physician services market. More specifically, although Feldman and Wholey (2001) and Schneider *et al.* (2008) find no relationship between health-insurer market power and physician pricing and output, Dafny *et al.* (2012) show that greater health-insurer market concentration is related to a reduction in both physician earnings and employment as a monopsony model suggests. The study by Dafny *et al.* (2012) comes across as being particularly persuasive because it uses a data set of 11 million people in various employer-sponsored health insurance plans across the nation over an 8-year period and specifies plan-fixed effects along with using a plausible instrumental variables approach. Dafny *et al.* (2012) findings also agree with basic intuition because physician markets are much less concentrated than hospital services markets and, unlike nurses, physicians are not unionized. Given these two conditions, health insurers may be able to exploit physicians. It will be interesting to see if future studies offer collaborative evidence.

The literature on the relationship between health-insurer market concentration and insurer behavior pales in comparison with the previous literature. It should be noted in **Table 3** that only six studies to date have focused on this particular topic and that these studies are relatively recent in comparison with the research on the previous topic. All but one study suggest that health insurers exercise market power by raising premiums and/or lowering output when the market for health insurance is more concentrated.

Dafny's (2010) study is particularly convincing because it shows that health insurers charge higher premiums to more profitable employers. Economic theory suggests that only firms with market power can practice price discrimination of that kind. In addition, Dafny *et al.* (2012) find that health insurance premiums spiked upward in areas where the health-insurer market concentration suddenly shot up because of a merger between Aetna and Prudential in 1998. Finally, Bates *et al.* (2012) show that the number of people with individually purchased health insurance (but not ESI) is lower in states where health-insurer market concentration is greater, particularly when no state rate review regulations exist. All in all, the evidence, although relatively limited, seems to suggest that health insurers are able to exercise market power in their output market. (Empirically examining the impact of mergers on premiums and profits provides another way of observing whether health insurers possess market power. Feldman *et al.* (1996) find that premiums increase in the most competitive market areas 1 year after mergers among HMOs. Hilliard *et al.* (2011) show that rivals' returns increase in response to a merger in market areas where the premerger HHI is high and the postmerger change in the HHI is large. Thus, both of these papers suggest health insurers engage in anticompetitive behavior.)

Summary and Conclusion

Whether health insurers possess and exercise market power remains an important issue for the US because the recently

Table 3 Effect of health-insurer concentration on insurer behavior

Author(s)	*Unit of analysis*	*Method*	*Findings*
Wholey *et al.* (1995)	1730 HMO market areas during 1988–91	OLS	Greater number of HMOs leads to lower HMO premiums
Foreman *et al.* (1996)	47 individual BC/BS plans during 1986–88	IV	Larger BC/BS market share leads to lower premiums
Pauly *et al.* (2002)	262 MSAs in 1994	IV	Greater competition among HMOs is associated with a lower industry profit rate
Dafny *et al.* (2012)	Panel data set of ESI plans enrolling more than 10 million people during 1998–2006	IV with plan-FE	Greater (merger-induced) health insurer concentration leads to higher employer premiums
Dafny (2010)	Panel data set of fully insured plan observations from 776 employers in 139 geographical markets during 1998–2005	IV with plan-FE	More profitable employers pay higher premiums in more concentrated markets
Bates *et al.* (2012)	Panel data set of 50 states and DC during 2001–7.	IV with state-FE	Greater health-insurer concentration leads to fewer people with individually purchased health insurance particularly in states without rate review regulations

Abbreviations: BC/BS, Blue Cross/Blue Shield; ESI, employer-sponsored insurance; FE, fixed effects; HMO, health maintenance organization; IV, instrumental variables approach; MSA, metropolitan statistical area; OLS, ordinary least squares method; PPO, preferred provider organization.

passed health insurance reform continues to rely heavily on a private health insurance industry. As discussed in this article, economic theory suggests that sellers and buyers may exploit their situation by raising prices above and lowering prices below the competitive level in the output and input markets, respectively, when the relevant market is highly concentrated. In both cases, these price distortions can lead to allocative inefficiency and large firms gaining at the expense of consumers or suppliers. The health insurance industry simultaneously plays critical roles as an important buyer of health- care provider services and as a health insurance provider to the public. Consequently, firms in the health insurance industry potentially can exercise both monopoly and monopsony power.

The empirical evidence to date suggests that health insurers may possess monopsony power in many physician services markets of the US. At least, one highly credible study finds that physicians are paid less and fewer physicians are employed when health insurers possess more market power in their area. However, studies focusing on the hospital services industry suggest the opposite. These studies find that health insurers, when they attain more market power, are able to bust the monopoly power of hospitals, thereby creating lower prices and more hospital services. Further complicating the analysis, recent research seems to have concluded that health insurers possess and exercise market power in their output market, that is, premiums are higher and fewer people are insured in areas where the health insurance industry is more highly concentrated. Consequently, it appears that health insurers, when they possess market power, are not passing along any cost savings from the hospital or physician services markets to buyers of health insurance.

Normally, reducing the market power of an industry, such as health insurance, would mean that suppliers and buyers, in this case physicians and consumer/patients, will unambiguously benefit. However, reducing market power also mean health insurers will be less able to hold the market power of hospitals in check. Given this trade-off, it is unclear how health policy authorities should craft public policies affecting the health insurance industry. For example, should public authorities level the playing field by allowing physicians to join unions so they can negotiate collectively to countervail the market power of health insurers? Or, should antitrust laws be enforced more aggressively toward health insurers or hospitals, or toward both? Or, would a profit tax on health insurers (and hospitals) be a better idea? How about a public health insurance option? According to the existing empirical literature, health policy analysts may have to confront these sorts of questions if economic efficiency is desired and a private health insurance system continues to be relied on in the US.

See also: Competition on the Hospital Sector. Empirical Market Models. Instrumental Variables: Informing Policy. Instrumental Variables: Methods. Markets in Health Care. Switching Costs in Competitive Health Insurance Markets

References

Adamache, K. W. and Sloan, F. A. (1983). Competition between non-profit and for-profit health insurers. *Journal of Health Economics* **2**, 225–243.

American Medical Association (2011). *Competition in Health Insurance: A Comprehensive Study of U.S. Markets*. Chicago, IL: American Medical Association.

Bates, L. J., Hilliard, J. I. and Santerre, R. E. (2012). Do health insurers possess market power? *Southern Economic Journal* **78**, 1289–1304.

Bates, L. J., Mukherjee, K. and Santerre, R. E. (2006). Market structure and technical efficiency in the hospital services industry: A DEA approach. *Medical Care Research and Review* **63**, 499–524.

Bates, L. J. and Santerre, R. E. (2008). Do health insurers possess monopsony power? *International Journal of Health Care Finance and Economics* **8**, 1–11.

Brooks, J. M., Dor, A. and Wong, H. S. (1997). Hospital-insurer bargaining: An empirical investigation of appendectomy pricing. *Journal of Health Economics* **16**, 417–434.

Dafny, L. (2010). Are health insurance markets competitive? *American Economic Review* **100**, 1399–1431.
Dafny, L., Duggan, M. and Ramanarayanan, S. (2012). Paying a premium on your premium? Consolidation in the health insurance industry. *American Economic Review* **102**, 1161–1185.
Dor, A., Grossman, M. and Koroukian, S. M. (2004). Hospital transaction prices and managed-care discounting for selected medical technologies. *American Economic Review* **94**, 352–356.
Feldman, R. and Greenberg, W. (1981). The relation between the Blue Cross share and the Blue Cross 'discount' on hospital charges. *Journal of Risk and Insurance* **48**, 235–246.
Feldman, R. and Wholey, D. (2001). Do HMOs have monopsony power? *International Journal of Health Care Finance and Economics* **1**, 7–22.
Feldman, R., Wholey, D. and Christianson, J. (1996). Effect of mergers on health maintenance organization premiums. *Health Care Financing Review* **17**, 171–189.
Foreman, S. E., Wilson, J. A. and Scheffler, R. M. (1996). Monopoly, monopsony, and contestability in health insurance: A study of Blue Cross plans. *Economic Inquiry* **34**, 662–677.
Halsersma, R. S., Mikkers, M. C., Motchenkova, E. and Seinen, I. (2010). Market structure and hospital-insurer bargaining in the Netherlands. *European Journal of Health Economics* **12**, 589–603.
Hilliard, J. I., Ghosh, C. and Santerre, R. E. (2011). *Changing competition in the health insurance industry: Are mergers anticompetitive?* Mimeo: University of Connecticut.
Melnick, G. A., Zwanziger, J., Bamezai, A. and Pattison, R. (1992). The effect of market structure and bargaining position on hospital prices. *Journal of Health Economics* **11**, 217–233.
Moriya, A. S., Gaynor, M. S. and Vogt, W. B. (2010). Hospital prices and market structure in the hospital and insurance industries. *Health Economics, Policy and Law* **5**, 459–479.
Pauly, M. V. (1998). Managed care, market power, and monopsony. *Health Services Research* **33**, 1439–1460.
Pauly, M. V., Hillman, A. L., Kim, M. S. and Brown, D. R. (2002). Competitive behavior in the HMO marketplace. *Health Affairs* **21**, 194–202.
Santerre, R. E. and Neun, S. P. (2013). *Health economics: Theories, insights, and industry studies*. Mason, Ohio: Cengage/Southwestern Publishers.
Scherer, F. M. (1980). *Industrial organization and market structure*. Chicago, IL: Rand-McNally.
Schneider, J. E., Li, P., Klepser, D. G., et al. (2008). The effect of physician and health plan market concentration on prices in commercial health insurance markets. *International Journal of Health Care Finance and Economics* **8**(1), 13–26.
Sorensen, A. T. (2003). Insurer-hospital bargaining: Negotiated discounts in post-deregulation Connecticut. *Journal of Industrial Economics* **51**(4), 469–490.
Staten, M., Dunkelberg, W. and Umbeck, J. (1987). Market share and the illusion of power: Can Blue Cross force hospitals to discount? *Journal of Health Economics* **6**, 43–58.
Staten, M., Umbeck, J. and Dunkelberg, W. (1988). Market share/market power revisited, a new test for an old theory. *Review of Economic Studies* **44**, 407–430.
Wholey, D., Feldman, R. and Christianson, J. B. (1995). The effect of market structure on HMO premiums. *Journal of Health Economics* **14**, 81–105.
Younis, M. Z., Rivers, P. A. and Fottler, M. D. (2005). The impact of HMO and hospital competition on hospital costs. *Journal of Health Care Finance* **31**(4), 60–74.

Heterogeneity of Hospitals

B Dormont, PSL, Université Paris Dauphine, Paris, France

Glossary

Economies of scale This is a result of increasing returns to scale: the amount of resource used per unit of output falls at higher output rates. It implies a falling unit cost as output rates increase, as long as input prices do not increase so as to offset the scale effect.

Economies of scope This enables a firm to produce several goods or services jointly more cheaply than producing them separately. The simultaneous production of hospital care and medical teaching is an example.

Exogenous source of cost variability In the context of a health care organization, a source is called exogenous if the hospital management cannot influence its level.

Legitimate source of cost variability Legitimacy is based on citizens' preferences. For instance, a particular location for a hospital, which corresponds to citizens' preferences for access to care, may be costlier than other locations.

Long-term moral hazard A time-invariant moral hazard implying that the hospital management is permanently inefficient.

Moral hazard In context with hospital payments, moral hazard refers to the fact that hospital managers can undertake more or less effort to minimize costs.

Prospective payment system A system that pays hospitals a fixed price per stay in a given diagnosis-related groups (DRG), irrespective of each hospital's actual cost. This provides a powerful incentive for managers to minimize costs.

Retrospective payment A payment representing reimbursement of the actual cost of treatment per stay.

Transitory moral hazard The effect on a hospital manager's transitory cost-reducing efforts.

Yardstick competition An industrial regulatory procedure under which the regulated price is set at the average of the estimated marginal costs of the firms in the industry. If differences in costs between hospitals are caused only by moral hazard, a yardstick competition rule of payment is to offer each hospital a lump sum payment per stay defined on the basis of average costs observed in other hospitals for stays in the same diagnosis-related groups (DRG). This system mimics competition on a free market in order to provide incentives for efficiency gains.

Introduction

Variability in hospital costs has often been used to convince citizens and policy makers of the extent of inefficiency in hospital care provision. Classification of hospital stays into diagnosis related groups (DRGs) has made it possible to place patients into groups that are supposed to be medically homogenous and to compare the cost of stays for similar cases in different hospitals. In a paper devoted to the political history of Medicare's transition to prospective payments per DRG, Mayes (2006) cited the unbelievably rapid growth rates of hospital costs in the USA: approximately 15% per year during the 1970s. However, there was still doubt about the contribution of inefficiency to such growth rates. Once DRGs were defined, policy makers finally reacted to differences in costs between hospitals for the same procedures. The introduction of prospective payments per DRG was decided on for Medicare, with the goal of forcing hospitals to increase efficiency. Similarly, in France, the debate on reforming hospital payments advanced in 1997, when the Ministry of Health decided to make public the differences in costs between French hospitals. Large differences in costs that were difficult to justify pointed to large differences in efficiency across hospitals, and showed that some of them were quite inefficient.

Nowadays, there is a general trend in all developed countries toward improving efficiency in hospital care through implementation of prospective payment systems (PPSs). Following the example of Medicare in 1983, other payers in the USA adopted PPS for inpatient care. European countries first adopted a global budget system to contain hospital costs during the 1980s, before turning to PPSs per DRG.

The Basic Inspiration of Prospective Payment Systems

The assumption at the root of a PPS is that any deviation in cost for a stay in a given DRG is because of inefficiency. Economists use the term 'moral hazard' to refer to the idea that the payer (the insurer or the regulator) cannot observe, much less monitor the efforts undertaken by hospital managers to minimize costs. Paying hospitals a fixed price per stay in a given DRG provides a powerful incentive for managers to minimize costs. Indeed, hospitals are supposed to keep the rent earned when their costs are lower than the fixed price. Conversely, they risk running operating losses if their costs are above DRG payment rates.

This payment scheme provides a perfect incentive for cost reduction because the payment is a lump sum per stay defined irrespective of a given hospital's actual cost. Yet, the regulator has an informational problem: she does not know how much care costs when the hospital is fully efficient (i.e., the 'true' minimal cost for a stay in a given DRG). The level of the lump sum defined by the regulator can lead the hospital to

 doi:10.1016/B978-0-12-375678-7.01314-6

(a)

(b)

Figure 1 Two hospitals in Paris. Both provide high-tech acute care. (a) *Hôpital Européen Georges Pompidou (HEPG)*, a very large hospital located in the center of Paris, has approximately 60 care units. Four older hospitals were closed and their care units grouped together at *HEPG*. Opened in 2001, this hospital was built following the latest standards of hospital architecture. It is reputed to be one of the best hospitals in Europe for cardiac surgery and cardiology. (b) *Groupe hospitalier La Pitié Salpétrière*, a very large hospital located in the center of Paris, has more than 70 care units. King Louis XIV ordered its creation. The hospital was designed by the architect Louis Le Vau, who was also in charge of the palace of Versailles. The hospital was built in the 17th century. Today, it is composed of many separate buildings, some of which date from the 17th century and some of which are modern. The old architecture of *La Pitié Salpétrière* is likely to induce higher costs because of difficulties in spatial organization. These extra costs are exogenous in the medium run. If the regulator does not pay for the consequences of this unfavorable architecture, she exposes *La Pitié Salpétrière* to operating losses, or provides incentives to reduce care quality, or to select patients, in short to cut costs in ways that run contrary to the general interest. Everybody is convinced that an old architecture induces higher costs. The question is: how much higher? It is not easy to answer this question because the extra costs are not observable: what can be observed is the impact of extra costs because of infrastructure difficulties combined with extra costs (or savings) owing to bad (or good) management.

bankruptcy or generate rents that are costly for tax payers (or the insured). This informational problem is solved by assuming that hospitals are homogeneous. In that case, differences in costs are caused only by moral hazard. Hence, an appropriate rule of payment is to offer each hospital a lump sum payment per stay defined on the basis of average costs observed in other hospitals for stays in the same DRG.

Shleifer's yardstick competition model provides the theoretical foundation for a PPS. This model is based on rather unrealistic assumptions: homogeneity of hospitals, homogeneity of patients for the same pathology, and fixed quality of care. Many studies have underscored the great diversity in hospitals' conditions of care delivery (teaching status, share of low-income patients, local wage level, etc.). Input prices can differ depending on location; care quality may vary, as may the severity of illnesses of admitted patients. These studies highlight the risks of such a PPS, namely selection of patients and a lowering of care quality. Indeed, hospitals which are subject to exogenous factors that lead to higher costs have to find ways to lower costs in order to avoid bankruptcy.

Sources of Heterogeneity in Hospital Costs

To avoid such problems, the regulator must design payments that allow for exogenous and legitimate sources of cost heterogeneity. This idea was formalized early on by Schleifer in his paper, published in 1985, one year after the beginning of Medicare's payment reform. He considered the case where the regulator can allow for the predicted impact on costs of observable characteristics that cannot be altered by the hospital. At first, Medicare adjusted its payments by a regional cost-of-labor index and gave extra payment to teaching hospitals. Currently, Medicare payments are adjusted for teaching hospitals, for a disproportionate share of indigent patients, and for local wage rates. In England, the national price per HRG (the English DRG) is adjusted for unavoidable differences in factor prices for staff, land, and building construction. More generally, in European countries, payment rates are adjusted for structural variables such as teaching, status, and region.

There is a theoretical debate on how observable causes of cost differences between hospitals should be allowed for in a PPS. Mougeot and Naegelen (2005) pointed out that most theoreticians implicitly assume that prospective payments are combined with a lump sum transfer. They show that this transfer should generally take the form of a tax paid by the hospital. Indeed, in a PPS hospitals whose costs are lower than the price per DRG receive a surplus, called a rent, which is costly to the tax payer. Social welfare will be maximized if this rent is extracted through a tax. But such a tax is not feasible in practice, given that most health care agencies do not have the power to 'fine' hospitals. If lump sum transfers are not feasible, it is possible to adjust fixed prices per DRG in order to reflect exogenous cost differences between hospitals. In this case, price adjustment should not necessarily be proportional to the extra cost; it can be optimal to discriminate against low-cost or high-cost hospitals by setting the price adjustment above or below marginal cost.

Table 1 Sources of cost variability between hospitals

Source of cost variability between hospitals	*Exogenous/endogenous*	*Are extra costs because of this factor legitimate or not?*	*Observable/ unobservable*	*Impact on costs*
Local wages, local input prices	Exogenous (for a given location)	Legitimate if the regulator wants a hospital located in the area	Observable	Can be estimated
Scale and scope economies	Exogenous or endogenous, depending on hospital's autonomy in developing supply strategies (institutional context)	Legitimate if they are because of regulations concerning the supply of care	Only rough indicators of specialization are available	Difficult to estimate (feasible for a restricted number of outputs only)
Other hospital characteristics that raise costs: poor architecture, etc.	Endogenous for private-for-profit hospitals Mostly exogenous for public hospitals	Legitimate if the regulator wants the hospital to continue to function and thinks rebuilding would be too costly	Observable	Cannot be estimated separately from other sources of cost differences
Care quality	Endogenous	Legitimate if the regulator wants to promote a quality level above the minimum	Partially observable; not all dimensions of quality are observed	Difficult to estimate
Patient characteristics	Endogenous if the hospital is allowed to select patients, exogenous otherwise	Legitimate if the regulator wants to prevent patient selection	Observable[a]	Can be estimated
Inefficiency because of moral hazard	Endogenous	Not legitimate	Unobservable	Cannot be estimated separately from other sources of cost differences

[a]It is supposed that the patient characteristics are observable for the most part, an assumption which ignores the possibility of moral hazard in reporting severity.

However, the main difficulty is that many sources of cost variability are not observable by the regulator, or the regulator cannot measure their impact on hospital costs. **Figure 1** concerns two hospitals in Paris. The *Hôpital Européen Georges Pompidou* was built recently, whereas most of *La Pitié Salpétrière* is very old and bears all the weight of its long history. Even if the regulator is convinced that *La Pitié Salpétrière* has extra costs because of the age of its buildings, the magnitude of these extra costs cannot be measured. At best, the regulator can observe the impact of additional costs because of poor infrastructure combined with extra costs (or savings) due to bad (or good) management, or combined with many other sources of cost variability: care quality, scale and scope economies, other hospital characteristics.

How can unobservable sources of cost heterogeneity be dealt with? How can we distinguish between differences in cost because of cost containment efforts and differences that cannot be reduced because they are a result of exogenous unobserved sources of hospital heterogeneity? Before turning to this question, the possible sources of cost heterogeneity are characterized by splitting them into six large categories (see **Table 1**). This classification is rather simplistic and debatable, but it may help in understanding what is at stake in the question of hospital heterogeneity.

For each source of cost variability, it is essential to know whether it is exogenous or endogenous, legitimate or illegitimate, and whether its impact on costs can be evaluated. A factor is considered to be exogenous if the hospital manager cannot influence its level. Legitimacy is based on citizens' willingness to pay (preferences). Consider for instance a hospital located in an area with limited road access. This induces higher transportation costs and possibly higher wages. Are people willing to pay an extra amount for a hospital located in this area? If the extra cost is considered illegitimate, the regulator will not adjust the DRG rate and the hospital must either reorganize or close down. Similarly, indigent patients induce higher costs because their hospital stays are generally longer. If the care system is supposed to offer similar access to care to every citizen, adjusting payments to avoid selection of patients is legitimate. The exogeneity of a cost factor may depend on hospital status: in France, patient characteristics are exogenous for public and private nonprofit hospitals, which are not allowed to select patients, whereas patient characteristics can be considered endogenous for private-for-profit hospitals.

Economies of scale are obtained when a lot of activity in one type of care service results in a lower cost per stay. Economies of scope arise when an appropriate mix of care services

results in a lower cost per stay. Very narrow specialization is generally linked to scale economies, combined with scope diseconomies. Scale and scope economies may be exogenous or endogenous, depending on the hospital's autonomy in developing supply strategies. The institutional context plays an important role: in the National Health Service of England, hospitals that are run by Foundation Trusts (FTs) have more freedom to shape their supply of care than other hospitals. In France, scale and scope economies are endogenous for private-for-profit hospitals but exogenous for public hospitals. The latter have a given capacity and their mandate obliges them to offer a broad mix of services in order to meet needs. Hence, extra costs because of diseconomies of scope for a private-for-profit hospital can be deemed illegitimate, if the hospital is not constrained by a public mandate.

The fact that a source of cost heterogeneity is observable does not imply that its impact on costs can be evaluated. As in the example of *La Pitié Salpétrière*, the regulator cannot measure extra costs associated with the age of the hospital buildings separately from other sources of extra costs. The factors considered in the table are shown in blue cells when their impacts on costs are likely to be difficult to estimate: they include moral hazard, of course, as well as some hospital characteristics, but also care quality and scale or scope economies. Indeed, quality is multidimensional and rather difficult to observe. (Concerning information on quality and the impact of competition on quality, see the contributions of Sutton and McGuire in this encyclopedia.) Scope economies are not easy to detect because currently available econometric tests are feasible only for a very small number of types of care services, which is not satisfactory, given the number of DRGs (at least several hundred) or Major Diagnosis Categories (several dozen).

How to Pay for Unobservable Heterogeneity?

Fixed payments per DRG put pressure on hospitals to compete. Because payments levels are set at average cost, hospitals which are affected by exogenous factors that induce higher than average costs risk losses. If they are already operating at full efficiency, they cannot realize further savings through efficiency gains. Hence, careless implementation of a PPS is likely to create undesirable incentives for selecting patients and lowering care quality. A regulator who aims at maximizing social welfare must design a payment system that creates virtuous incentives for enhancing hospital efficiency, without providing deleterious incentives for patient selection and quality reduction.

To address this question, many theoretical papers have tried to improve the basic model by lifting assumptions relative to patient and hospital homogeneity, and by allowing for endogenous levels in the number of procedures and quality of treatment. Using various theoretical frameworks and hypotheses, these papers show that social welfare can be improved through a mixed payment system that combines a fixed price with partial reimbursement of the actual cost of treatment per stay. To deal with unobserved sources of heterogeneity in costs, the regulator can construct a menu of contracts that combine a lump-sum transfer with partial reimbursement of actual costs. When the hospital chooses a contract, it reveals its unobserved cost component. Currently, however, such a payment scheme is not implemented in any health system. In fact, the theoretical design of the contracts often relies on unobservable variables or functions, such as, for instance, the function describing the disutility of the hospital manager's cost reduction efforts. Hence, such theoretical designs are hardly used in reality.

Another strategy is to use econometrics to evaluate unobservable sources of cost heterogeneity. The sources of hospital cost heterogeneity are summarized in **Table 1**. A hospital's activity is more or less costly, depending on its infrastructure, the existence of economies of scale or of scope, the quality of care and the cost reduction effort provided by the hospital manager (moral hazard). Moral hazard can be split into two components: long-term moral hazard and transitory moral hazard. Long-term moral hazard is supposed to be time invariant: the hospital management can be permanently inefficient. An example of permanent inefficiency would be an obsolete elevator which is very slow and subject to frequent breakdowns and which is not replaced for several years. Transitory moral hazard is linked to the manager's transitory cost reduction efforts. For instance, the manager can be more or less rigorous, each year, when negotiating prices for supplies or for services provided to the hospital by outside firms. It would be optimal for social welfare to eliminate long-term moral hazard as well as *transitory* moral hazard. However, it is very difficult to separate long-term moral hazard from other sources of cost heterogeneity which are legitimate.

The use of a three-dimensional nested database makes it possible to identify *transitory* moral hazard. It is then possible to design a payment that allows for hospital heterogeneity in costs, while still providing incentives to increase efficiency because it does not reimburse costs due to *transitory* moral hazard (see the technical appendix).

A fully PPS reimburses each stay with a fixed price regardless of the actual cost of the stay. The payment systems currently implemented in most countries take some observable sources of cost heterogeneity, such as local input prices, into account. A preferable method of payment would be to allow for observable and *some unobservable* sources of cost heterogeneity, provided they are *time invariant*. With such a payment rule, the regulator reimburses each hospital for extra costs that might correspond to undesirable long-term moral hazard, but which can as well correspond to legitimate heterogeneity. Nevertheless, this method of payment creates incentives to increase efficiency because it does not reimburse extra costs that are a result of *transitory* moral hazard.

The general idea is that the regulator has no means to disentangle legitimate and illegitimate sources of time-invariant cost heterogeneity, i.e., to separate the wheat from the chaff. In this context, it might be preferable to accept to pay for long-term moral hazard in order not to penalize hospitals which have legitimate sources of cost heterogeneity. Is this view unreasonable? The question becomes an empirical one: if transitory moral hazard has a substantial impact on cost variability, it would be possible to achieve large gains in

efficiency even while paying for permanent sources of hospital cost variability.

An empirical estimation has been carried out by Dormont and Milcent (2005) on a sample of stays for acute myocardial infarction in French public hospitals. It appears that the cost variability because of transitory moral hazard was quite sizeable. Simulations show that substantial budget savings – at least 20% – could be expected from implementation of a payment rule that takes all unobservable hospital heterogeneity into account, provided that it is time invariant. This payment rule is easy to implement if the regulator has information about costs of hospital stays. A drawback is that it gives higher reimbursements to hospitals which are costlier because of permanently inefficient management. However, it has the great advantage of reimbursing high quality care. Moreover, it can lead to substantial savings, because it provides incentives to reduce costs linked to transitory moral hazard, whose influence on cost variability can be sizeable

Technical Appendix: Designing Payments That Allow for Cost Heterogeneity between Hospitals

The use of a three-dimensional nested database, with information recorded at three levels (stays–hospitals–years), makes it possible to identify transitory moral hazard and to estimate its effect on hospital cost variability. For a given DRG, we can observe the cost $C_{i,h,t}$ of the stay i, which occurred in hospital h in year t. This cost can be decomposed as follows: $C_{i,h,t} = \tilde{C}_{i,h,t} + a + \eta_h + \varepsilon_{h,t} + u_{i,h,t}$. If stays for the same DRG always had the same cost, the cost would always be equal to the constant a. A fully PPS is based on this assumption, which implies that the other terms of the right-hand side of the equation would be equal to zero.

As stated above, there are some legitimate sources of cost variability, some of which are observable: patient characteristics, local input prices. The impact of these characteristics on costs can be estimated: we denote this cost heterogeneity as $\tilde{C}_{i,h,t}$. Given the observable characteristics, cost variability then depends on the sum of three random variables: $\eta_h + \varepsilon_{h,t} + u_{i,h,t}$. The term $u_{i,h,t}$ represents unobservable heterogeneity between patients: its average is equal to zero at the hospital level. Hence, for given observable hospital characteristics, hospital costs are affected by the terms η_h and $\varepsilon_{h,t}$. These random variables are not observed but can be estimated with a three-dimensional database.

By definition, the term η_h specifies time-constant unobservable hospital heterogeneity. It can be seen as the result of several components summarized in Table 1. In short, a hospital's activity is more or less costly, depending on several factors: its infrastructure, the existence of economies of scale or of scope, the quality of care and the cost reduction effort provided by the hospital manager (moral hazard). As a component of a time-invariant term (η_h), the moral hazard involved here is long term: the hospital management can be permanently inefficient. It would be optimal for social welfare to eliminate long-term moral hazard as well as transitory moral hazard. However, long-term moral hazard cannot be separated from the other components of η_h, which are legitimate sources of cost heterogeneity.

The term $\varepsilon_{h,t}$ is defined as the deviation, *ceteris paribus*, for a given year t, of hospital h's cost in relation to its average cost. It can be seen as the result of transitory moral hazard, measurement errors and unobserved transitory shocks affecting hospital costs. Actually, measurement errors and unobserved transitory shocks are likely to be of slight importance. Indeed, a measurement error belonging to $\varepsilon_{h,t}$ would be patient invariant by definition. In other words, it would be replicated for each stay in the same hospital during the same year, which is unlikely. As for transitory shocks, they should be observable if they are justifiable. It is true that any hospital can be affected by a shock in a given year: an electrical failure, for example. However, the regulator would be well advised to classify *a priori* these incidents as moral hazard, in order to give hospitals incentives to declare them, when the extra costs they induce are justifiable and exceptional. Hence $\varepsilon_{h,t}$ is mostly made of moral hazard: an econometric test run on French data by Dormont and Milcent (2005) gave empirical support to this conjecture. More precisely, $\varepsilon_{h,t}$ is an indicator of transitory moral hazard (indeed, all time-invariant components ofunobserved hospital heterogeneity are represented in the term η_h).

A fully PPS reimburses each stay with a fixed price $P_{i,h,t}=a$, whatever the actual cost of the stay $C_{i,h,t}$. The payment systems currently implemented in most countries take some observable sources of cost heterogeneity into account. With our notation, the reimbursement then equals: $P_{i,h,t} = \tilde{C}_{i,h,t} + a$. A preferable method of payment would be to allow for observable and some unobservable sources of cost heterogeneity, provided they are time invariant. The payment would be equal to: $P_{i,h,t} = \tilde{C}_{i,h,t} + a + \eta_h$. With such a payment rule, the regulator in effect tailors reimbursement to each hospital. Indeed, the component η_h is specific to hospital h. It might correspond to undesirable long-term moral hazard, but it can also correspond to legitimate heterogeneity. This method of payment can nevertheless create incentives to increase efficiency because it does not reimburse extra costs that are due to transitory moral hazard ($\varepsilon_{h,t}$ is not a component of payment $P_{i,h,t}$).

See also: Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Markets in Health Care

References

Dormont, B. and Milcent, C. (2005). How to regulate heterogeneous hospitals? *Journal of Economics and Management Strategy* **4**(3), 591–621.

Mayes, R. (2006). The origins, development, and passage of Medicare's revolutionary prospective payment system. *Journal of the History of Medicine and Allied Sciences* **62**(1), 21–55.

Mougeot, M. and Naegelen, F. (2005). Hospital price regulation and expenditure cap policy. *Journal of Health Economics* **24**, 55–72.

Further Reading

Chalkley, M. and Malcomson, J. M. (2000). Government purchasing of health services. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1A, pp 847–890. Amsterdam: Elsevier.

Laffont, J. J. and Tirole, J. (1993). *A theory of incentives in procurement and regulation*. Cambridge, MA: MIT Press.
Ma, A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy* **3**(1), 93–112.
Miraldo, M., Siciliani, L. and Street, A. (2011). Price adjustment in the hospital sector. *Journal of Health Economics* **30**, 112–125.
Mougeot, M. and Naegelen, F. (2012). Price adjustment in the hospital sector: How should the NHS discriminate between providers? A comment on Miraldo, Siciliani and Street. *Journal of Health Economics* **31**, 319–322.
Shleifer, A. (1985). A theory of yardstick competition. *RAND Journal of Economics* **16**, 319–327.

HIV/AIDS, Macroeconomic Effect of

M Haacker, London School of Hygiene and Tropical Medicine, London, England, UK

Introduction

Concerns about the macroeconomic consequences of the human immunodeficiency virus, and the associated acquired immunodeficiency syndrome (HIV/AIDS) have been fueled by several factors. Most obviously, the epidemic has a devastating impact on life expectancy in a number of countries. In the empirical literature on economic growth (not dealing specifically with HIV/AIDS), such a decline is associated with a steep drop in the growth of gross domestic product (GDP). More informally, there are concerns that the epidemic could affect long-term development aspects by destroying human capital and the incentives to invest in education, disrupt the social fabric of a society, and result in an increasing number of disadvantaged young people (mainly orphans).

Second, there have been concerns that the impact of the epidemic is tied up with and exacerbates the challenges of economic development. For example, the 2006 Political Declaration issued by the United Nations (UN) states "that in many parts of the world, the spread of HIV/AIDS is a cause and consequence of poverty, and that effectively combating HIV/AIDS is essential to the achievement of internationally agreed development goals and objectives."

Third, the response to HIV/AIDS in many countries has become a macroeconomic factor in its own right, not only because it partially reverses the adverse direct consequences of the epidemic but also because of the additional demand for (health) services, and the challenges of financing HIV programs.

Against this background, the article focuses on three areas. It sets out with a discussion of the state of the epidemic across countries and its correlation with the state of economic development. This is followed by a discussion of the literature and evidence on the macroeconomic impacts of HIV/AIDS. Finally, the article highlights macroeconomic aspects of the financing of HIV programs, including the role external assistance has played in this.

HIV/AIDS and the State of Economic Development

The macroeconomic implications of HIV/AIDS depend on the economic context, as well as the state of the epidemic. For example, the impact of HIV/AIDS on affected households depends on available health services and the availability of health and social insurance, companies with high value added per employee have higher stakes in investments to minimize the impact of HIV/AIDS on their staff and operations, and the government's capabilities in meeting the demand for HIV/AIDS-related services are constrained by its fiscal resources.

Moreover, the state of the epidemic is partly endogenous, and the quality of the policy response to the epidemic in turn reflects the quality of a country's institutions and its economic and public policy capacities. From the perspective of global development policy, where HIV/AIDS competes with other causes for external assistance, it is also useful to place the epidemic in an economic context.

HIV/AIDS-related deaths are concentrated in low-income countries, similar to infectious diseases more generally. According to the 'Causes of Death 2008' data published by the World Health Organisation, 41% of HIV/AIDS-related deaths and 36% of deaths from infectious diseases occurred in low-income countries in 2008 (which accounted for 12% of the global population). In terms of its association with economic development challenges, HIV/AIDS thus resembles infectious diseases in general, but it is not correlated as closely with basic economic development challenges as malaria deaths are, of which 58% occurred in low-income countries.

However, HIV/AIDS mortality has been declining, reflecting increased access to treatment. According to the data from the 2012 report on the Global AIDS Epidemic by the Joint United Nations Programme on HIV/AIDS (UNAIDS), 542,000 AIDS deaths (32% of global AIDS deaths) occurred in low-income countries in 2011.

The broad distribution of HIV deaths by income group, however, gives a misleading picture of the challenges posed by HIV/AIDS, as HIV/AIDS is distributed across countries very unevenly. Taking, for example, the global distribution of income (Gini coefficient: 0.64) as a reference point, the burden of HIV/AIDS is distributed much more unevenly (Gini coefficient: 0.74). This point is illustrated in Figure 1, which orders the global population by GDP per capita and adds a curve describing HIV prevalence in the respective countries. Indeed, HIV prevalence tends to be higher in countries with lower income. This is evident from the negative coefficient of correlation between HIV prevalence and GDP per capita (-0.09), substantial HIV prevalence in a number of low-income countries (broadly, those to the left of the 700-million population mark in Figure 1) and an absence of high HIV prevalence among high-income countries (broadly, the rightmost billion in Figure 1). The most striking feature of the distribution of people living with HIV, however, is the high concentration of HIV/AIDS in a few countries with HIV prevalence over 10% of the total population. In this regard, it also differs from malaria, which correlated more strongly with the state of economic development in general (with higher prevalence in low-income countries) but is less concentrated in specific countries.

Although the correlation between HIV prevalence and GDP per capita is not very strong, the consequences of an HIV infection differ substantially across countries. While mortality among people living with HIV typically was between 1% and 1.5% for high-income countries like France, Spain, or the US, it averaged 4.8% in 34 low-income countries in 2011, and exceeded 8% in Liberia, Nepal, and Somalia (according to estimates from the 2012 UNAIDS Report on the Global AIDS Epidemic). These estimates also illustrate the large impact of increased access to treatment – mortality among people living

 doi:10.1016/B978-0-12-375678-7.00610-6

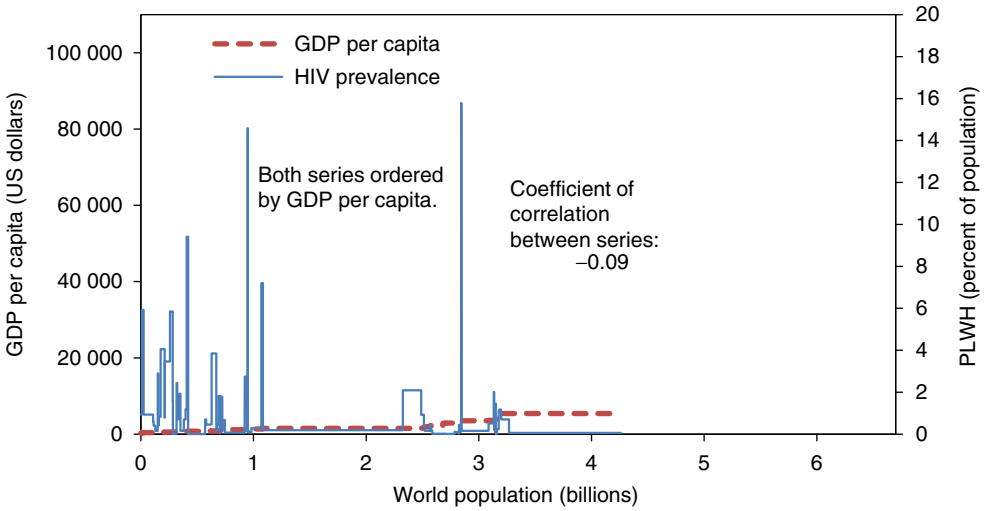


Figure 1 HIV prevalence and GDP per capita (2011). Data sources: UNAIDS (2012). *Report on the global AIDS epidemic 2012*. Geneva: UNAIDS, International Monetary Fund, world economic outlook database, October 2012 edition (2012), and United Nations Population Division, world population prospects: The 2010 revision (2011).

with HIV in the 34 low-income countries has declined by almost one-half (from 8.6%) since 2005. However, very large differences in the health consequences of an HIV infection across countries with different levels of economic development appear to persist.

Within countries, the correlation between HIV/AIDS and income (or other socioeconomic characteristics) is less straightforward. One of the most important data sources are demographic and health surveys also covering HIV prevalence. Most of these surveys suggest that HIV prevalence tends to be higher for wealthier population groups, but there is no consistent pattern across countries.

In summary, as is the case with infectious diseases more generally, HIV/AIDS deaths occur predominantly in developing countries. However, HIV/AIDS is unusual as it is distributed highly uneven across countries. These observations have implications for the macroeconomic significance of the epidemic. Because the health impact of HIV/AIDS has been so disruptive in specific countries, and because this health shock has emerged as a development threat only over the last 20 odd years, it is plausible that the epidemic has economic consequences (e.g., for GDP growth), which cannot easily be detected for more common and chronic health conditions (e.g., malaria). At the same time, the impact of HIV/AIDS provides a testing ground for theories on health and economic development.

Macroeconomic Impact of HIV/AIDS

In spite of its devastating impact on health outcomes such as life expectancy in a number of countries, the impact of HIV/AIDS on economic growth is not obvious. This point is illustrated by Figures 2 and 3, which contrast trends in life expectancy in 10 countries facing the highest HIV prevalence worldwide and the recent growth experience in these countries. (As all of these countries are located in sub-Saharan Africa, the figures also provide averages for the region for comparison.)

According to Figure 2, the impact of HIV/AIDS on life expectancy has been very large, ranging from a loss of 7 years (Uganda) to a loss of 21 years (Zimbabwe) in 2000–05. Moreover, the adverse impact was so strong that life expectancy declined in absolute terms in 8 of the 9 countries, and collapsed to a level last observed in the 1950s or 1960s in Botswana, Lesotho, South Africa, Zambia, and Zimbabwe. In some countries, the negative trend was started to reverse in 2005–10, partly because the HIV epidemic had matured (and the number of AIDS cases was no longer escalating) and partly as a consequence of increased access to treatment.

As evident from Figure 3, the large decline in life expectancy has not resulted in a steep drop in GDP growth. The rate of growth of GDP per capita in 9 countries with high HIV prevalence slowed down, somewhat relative to the rest of sub-Saharan Africa since the mid-1990s. However, the timing of the slowdown precedes or is less persistent than the increase in HIV/AIDS-related mortality. By 2010, the countries with high prevalence can be divided in two groups: (1) Low-income countries like Malawi, Mozambique, Zambia, and Zimbabwe, experiencing large swings in growth rates arguably not caused by HIV/AIDS (this applies especially to the economic crisis in Zimbabwe); (2) South Africa and the enclosed or neighboring middle-income countries Botswana, Lesotho, Namibia, and Swaziland, all experiencing growth rates below the average for sub-Saharan Africa, but which also differ from most countries in sub-Saharan Africa in many regards other than the state of HIV/AIDS.

The empirical evidence is also ambiguous. Studies including HIV prevalence or AIDS-related deaths directly in regressions find no or very small impacts of HIV/AIDS on growth. In contrast, studies identifying a large impact of HIV/AIDS usually build on established findings of the empirical growth literature, notably the positive correlation of growth and life expectancy and then link the variable of interest to HIV/AIDS. In light of the strong impact of HIV/AIDS on life

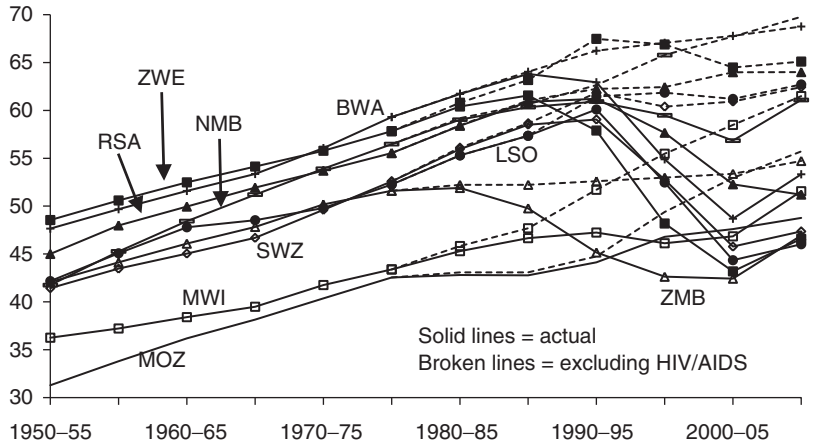


Figure 2 Life expectancy at birth, 9 countries with high HIV prevalence, 1950–2010 (years). Data sources: United Nations Population Division, world population prospects: The 2010 revision (2011). Figure covers BTW, Botswana; LSO, Lesotho; MWI, Malawi; MOZ, Mozambique; NMB, Namibia; RSA, South Africa; SWZ, Swaziland; ZMB, Zambia; and ZWE, Zimbabwe.

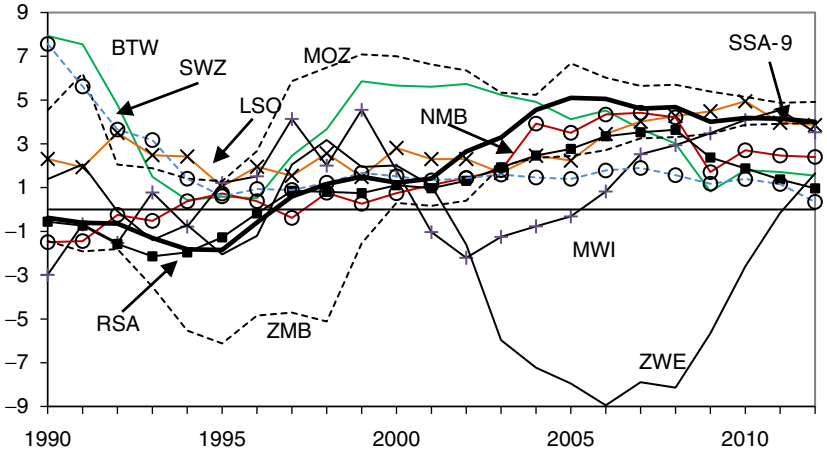


Figure 3 Growth of GDP per capita, 9 countries (average annual growth in 5-year period ending in year indicated). Data sources: International Monetary Fund, world economic outlook database, October 2012 edition (2012), and World Bank, world development indicators (2012). Figure covers BTW, Botswana; LSO, Lesotho; MWI, Malawi; MOZ, Mozambique; NMB, Namibia; RSA, South Africa; SSA-9, sub-Saharan Africa-9; SWZ, Swaziland; ZMB, Zambia; and ZWE, Zimbabwe.

expectancy (or similar variables), this empirical approach returns a large negative impact of HIV/AIDS on growth but rests on two untested hypotheses: (1) The correlation of growth and life expectancy reflects a causal link between health and growth and (2) HIV/AIDS affects economic outcomes in a similar way as changes in the state of population health reflected in changes in life expectancy across countries. Both assumptions are doubtful. Some observers point to common factors like institutions affecting the functioning of health systems, governance, and growth. Also, the health impact of HIV/AIDS has a specific profile that does not simply reverse health gains achieved over the past decades, and it has occurred much more quickly than the gradual improvements in health outcomes achieved over the past decades.

If the links between health and economic outcomes are of a longer term nature, this could mean that the impacts of HIV/AIDS on economic growth have not fully materialized yet. For example, economic theory suggests that higher mortality risks reduce the returns to education. HIV/AIDS could therefore slowdown the accumulation of human capital and economic growth. As this effect would take several decades to materialize (as cohorts grow from school benches through the working-age population), it would barely show up in economic growth data at present, and there would not be a clear contemporary correlation between HIV prevalence and economic growth. Some microempirical evidence points to lower school attendance in areas highly affected by HIV/AIDS, consistent with such a hypothesis about the long-term economic consequences of HIV/AIDS.

Another possible reason why the impacts of HIV/AIDS on growth have been small so far is the fact that economic activity, within countries, is distributed unevenly. It has been observed that HIV is associated with certain economic activities like mining, and that migrant workers also play a large role in disseminating HIV. However, as value added per worker in mining is high, companies can afford to take actions

to prevent any disruptions to production from increased mortality or morbidity, at a low cost relative to turnover or value added.

The discourse regarding the macroeconomic effects of HIV/AIDS has focused on the growth impacts of the epidemic. It is important to take note of the fact that HIV/AIDS also results in a shift in the composition of spending. As governments and households shift expenditures to respond to the epidemic and address its consequences, these funds are no longer available for other purposes, i.e., private or public consumption and investment. Compared with a no-AIDS situation, HIV/AIDS-related spending therefore adds to the economic costs of the epidemic. The discussion in the Section Macroeconomic Aspects of the Response to HIV/AIDS, suggests that public HIV/AIDS spending accounts for several percent of GDP in a number of countries. Private HIV/AIDS spending and shifts in the allocation of time within households add to these economic costs.

The steep declines in life expectancy that can arise because of HIV/AIDS can also be interpreted as an economic cost. Such interpretations of the health impact of HIV/AIDS draw on estimates of the value of statistical life, which typically suggest that a loss in life expectancy of one percent is equivalent to an income loss of 3–4%. A loss in life expectancy of 23% (as in Botswana, 2005–10, compare **Figure 2**) would then translate into an economic cost exceeding one-half of GDP. Even in countries like the US, with an HIV prevalence of 0.6% and a loss in life expectancy of half an year, the costs of increased mortality, by this count, exceed 2% of GDP.

Small aggregate impacts of HIV/AIDS may mask shifts below the surface of national averages, which are relevant from a welfare perspective. For example, it is plausible that high HIV prevalence increases the risk to material living standards and – for parts of the population – of falling into poverty (even though other households may benefit, taking advantage of employment opportunities vacated by people affected by HIV/AIDS). Also, even though HIV prevalence tends to be somewhat higher among wealthier population groups, differences in access to treatment across population groups, in a country facing an HIV epidemic, can exacerbate inequalities in health prospects. Although demographic and health surveys consistently return higher rates of access to health services for wealthier population groups, little data are available regarding the benefit incidence of HIV/AIDS-related health services and the consequences of increased demand for HIV/AIDS-related health services (and a corresponding scaling-up in the supply of such services) for access to health services more generally.

Macroeconomic Aspects of the Response to HIV/AIDS

The global response to HIV/AIDS has altered the course of the epidemic. The macroeconomic impact of HIV/AIDS therefore partly reflects the consequences of policy interventions, in several dimensions: (1) HIV incidence, (2) the microeconomic consequences, (3) the growth impacts of HIV/AIDS, and (4) the costs of the response to the epidemic.

In many countries, HIV incidence has declined very considerably from its peak. In South Africa, for example, HIV incidence among the population of ages 15–49 years declined from a peak of 2.8% in 1998 to 1.3% in 2011. As a consequence, the health outlook in countries experiencing such declines is improving, and the economic consequences of HIV/AIDS become less forceful.

More immediately, the adverse economic consequences of HIV/AIDS are modified by increased access to treatment. This intuition is supported by empirical analysis on the microeconomic level, illustrating a reversal in worker's productivity following initiation of treatment. These estimates, however, are available only in settings where labor input and output are directly observable (e.g., tea pluckers) and may not translate one-to-one to other sectors and contexts, such as capital-intensive mining or services, which account for a large share in GDP.

The studies of the macroeconomic effects of HIV/AIDS also provide some pointers regarding the consequences of treatment (and the later studies frequently offer explicit estimates). In addition to mitigating productivity losses, antiretroviral treatment reduces the decline in population growth and reduces the private and public costs of care. Looking ahead, the prospect of access to treatment changes the risks associated with an HIV infection. Along with declining risk of becoming infected, it therefore increases the incentives to invest in education, therefore mitigating one of the most forceful effects through which HIV/AIDS could affect long-term growth.

Macroeconomic studies, which explicitly account for the impact of antiretroviral treatment, illustrate the extent to which increased access to treatment mitigates the economic impacts of HIV/AIDS, frequently suggesting a reversal in the growth impact of approximately one-third to one-half of the unfettered ('no treatment') impact of HIV/AIDS. This reversal is less than complete even where the rate of access to treatment is very high because treatment only mitigates and delays the adverse health consequences of HIV/AIDS, and because the costs of treatment crowd out other investments. Some observers argue that access to treatment could be financed from this 'growth dividend' (and reduced costs of other HIV/AIDS-related health services). This, however, is not necessarily the case, as the 'growth dividend' is not directly available for higher health spending (people surviving longer because of treatment need to eat).

The policy response does not merely reverse the adverse macroeconomic impacts of the epidemic. The costs of the response in many countries have attained a level that is significant from a fiscal perspective, and HIV/AIDS-related external aid may account for a substantial proportion of aid received. Globally, HIV/AIDS accounted for US$ 8.0 billion out of total disbursements of official development assistance of US$ 150 billion in 2011, and out of US$ 19.4 billion in the areas of health and population policies, according to the "creditor reporting system" database maintained by the Organisation for Economic Co-operation and Development.

The high costs of the response to HIV/AIDS in numerous countries are illustrated in **Figure 4**. The burden of funding the HIV/AIDS program, relative to GDP, is not necessarily the largest in the countries facing the highest HIV prevalence (Botswana, Lesotho, Namibia, South Africa, and Swaziland) but in a number of low-income countries facing HIV prevalence between 3% and 15%. In particular, some least-developed

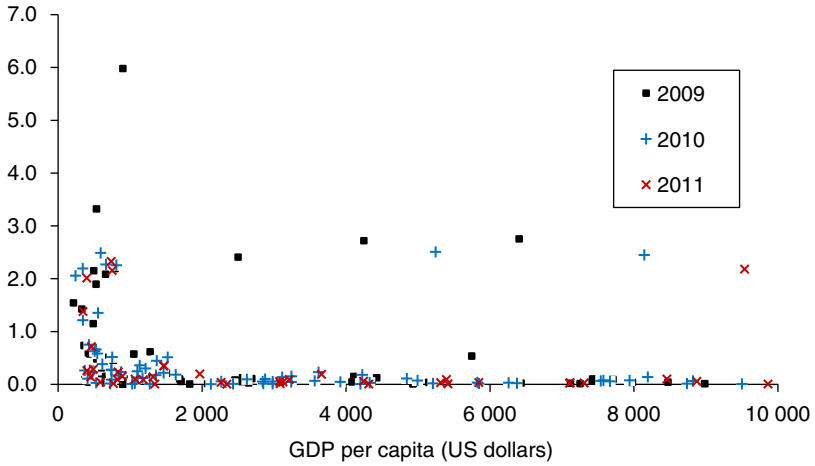


Figure 4 Total HIV/AIDS spending (percent of GDP). Data sources: International Monetary Fund, world economic outlook database, October 2012 edition (2012) and UNAIDS, AIDS spending data (2012).

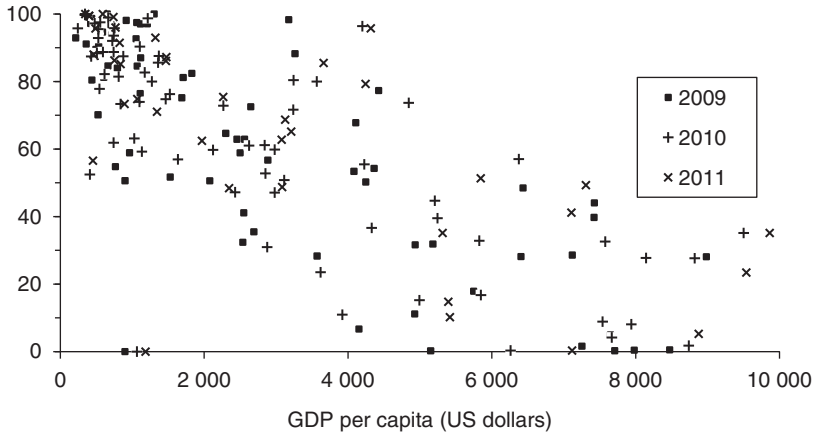


Figure 5 Externally financed HIV/AIDS spending (percent of total spending). Data sources: International Monetary Fund, world economic outlook database, October 2012 edition (2012) and UNAIDS, AIDS spending data (2012).

countries face a very large financing challenge, even though HIV prevalence is moderate. This is the case because the unit costs of HIV/AIDS interventions differ across countries much less than the level of GDP per capita.

The spending figures summarized in Figure 4 confirm that HIV/AIDS spending is significant from a fiscal perspective in many countries. In a typical low-income country (the figures are based on the median for this country group), public spending accounts for approximately 25% of GDP, of which 8% (equal to 2% of GDP) go toward health. According to Figure 4, the costs of the national response to HIV/AIDS (whether delivered through the public sector or non-governmental organisations) thus exceed total public health spending in a number of countries. These high levels of spending would be hard to envisage without high levels of external assistance.

Health is an area in which external assistance is playing a large role across developing countries in general. Owing to the uneven distribution of HIV/AIDS across countries, and the high costs of HIV/AIDS in a number of countries, the role of external assistance is even more pronounced in the area of HIV/AIDS spending, as illustrated by Figure 5. For low-income countries (broadly, those with GDP per capita of less than US$ 1000 in Figure 5), external financing usually accounts for more than 80% of the total costs of the HIV/AIDS program and in some cases close to 100%. In contrast, external assistance for public health spending rarely exceeds two-thirds of total spending. The differences in external funding between HIV/AIDS and health are even more pronounced for middle-income countries including countries which are not facing very high HIV prevalence rates.

Looking ahead, two aspects of the fiscal dimension of HIV/AIDS are worth noting. First, the costs of HIV/AIDS programs are going to remain high for a long time, even where HIV incidence is declining. This is the case because the number of people receiving treatment are still rising, and an increasing number of people who have contracted HIV in the past will require treatment. Second, there is a perception (and some early evidence) that external funding for HIV/AIDS is stagnating or even declining. This will place the funding of

HIV/AIDS programs under pressure, especially in low-income countries where HIV/AIDS spending is high relative to the government's fiscal resources.

Concluding Remarks

The impact of HIV/AIDS on economic growth has been small so far. This finding raises some questions regarding the empirical literature on health and growth (which would predict a large impact), but it could also be the case that the link from increased mortality to growth occurs so slowly and has not fully materialized yet. In many countries, HIV/AIDS programs have attained a scale that is significant from a fiscal perspective. The response to HIV/AIDS has been enabled by high rates of external assistance in the past, but the availability of funding is perceived to decline. Under these circumstances, sustaining the funding of HIV/AIDS programs will present a challenge especially for a number of low-income countries.

See also: HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. What Is the Impact of Health on Economic Growth – and of Growth on Health?

Further Reading

Bachmann, M. O. and Booysen, F. L. R. (2006). Economic causes and effects of AIDS in South African households. *AIDS* **20**(14), 1861–1867.

Botswana Institute for Development Policy Analysis (2000). *Macroeconomic impacts of the HIV/AIDS epidemic in Botswana*. Gaborone, Botswana: BIDPA.

Case, A. and Ardington, C. (2006). The impact of parental death on school outcomes: Longitudinal evidence from South Africa. *Demography* **43**(3), 401–420.

Case, A. and Paxson, C. (2011). The impact of the AIDS pandemic on health services in Africa: Evidence from demographic and health surveys. *Demography* **48**(2), 675–697.

Deaton, A. (2006). Global patterns of income and health: Facts, interpretations, and policies. *NBER Working Paper No. 12269*. Cambridge, MA: NBER.

Ellis, L., Laubscher, P. and Smit, B. (2006). *The macroeconomic impact of HIV/AIDS under alternative intervention scenarios (with specific reference to ART) on the South African economy*. Stellenbosch, South Africa: Bureau for Economic Research, University of Stellenbosch.

Haacker, M. (ed.) (2004). *The macroeconomics of HIV/AIDS*. Washington, DC: International Monetary Fund.

Nattrass, N. (2003). *The moral economy of AIDS in South Africa*. Cambridge: Cambridge University Press.

Parkhurst, J. O. (2010). Understanding the correlations between wealth, poverty and human immunodeficiency virus infection in African countries. *Bulletin of the World Health Organization* **88**(7), 519–526.

Sahn, D. E. (2010). *The socioeconomic dimensions of HIV/AIDS in Africa*. Ithaca, NY and London: Cornell University Press.

UNAIDS (2012). *Report on the global AIDS epidemic 2012*. Geneva: UNAIDS.

Whiteside, A. (2008). *HIV/AIDS: A very short introduction*. Oxford and New York: Oxford University Press.

HIV/AIDS: Transmission, Treatment, and Prevention, Economics of

D de Walque, The World Bank, Washington, DC, USA

Abbreviation

UNAIDS Joint United Nations Program on HIV/AIDS.

Glossary

Concurrency When an act of sex with one partner occurs between two acts of sex with another partner.

Disinhibition behaviors People may increase the riskiness of their behavior in response to perceived decreases in risk of acquiring or transmitting the virus.

Serodiscordance When one member in a sexual partnership is HIV positive and the other is not.

Introduction

At the end of 2011, according to Joint United Nations Program on Human immunodeficiency virus (HIV)/Acquired immunodeficiency syndrome (AIDS) (UNAIDS), an estimated 34 million people were living with HIV worldwide. The number of people dying of AIDS-related causes fell to 1.7 million in 2011, down from a peak of 2.2 million in the mid-2000s. There were 2.5 million new HIV infections in 2011, including an estimated 390 000 among children. This was 15% less than in 2001, and 21% below the number of new infections at the peak of the epidemic in 1997. Sub-Saharan Africa remains the region most heavily affected by HIV. In 2011, approximately 69% of all people living with HIV resided in sub-Saharan Africa, a region with only 12% of the global population. Sub-Saharan Africa also accounted for 70% of new HIV infections in 2010, although there was a notable decline in the regional rate of new infections. As Africa shoulders the heaviest burden, this article emphasizes evidence from this continent.

The article focuses on the economics of HIV/AIDS and therefore does not emphasize the biomedical determinants of the epidemic. However, it briefly summarizes some of the recent biomedical prevention interventions. The focus of the discussion is on behaviors, economic behaviors, and incentives in particular. For that reason, the article does not address the HIV epidemic among children, even though it constitutes a heavy burden and an important challenge. This article will mainly review the microeconomic aspects of HIV/AIDS.

The article articulates the discussion around the three themes of HIV transmission, prevention, and AIDS treatment. It starts by exploring the determinants of HIV transmission, focusing on behavioral (gender and marriage, serodiscordant couples and multiple partners, and concurrency) and socioeconomic (poverty, education, and occupation) determinants. A short Section '(Micro-) Economic Consequences of HIV/AIDS' follows. The Section 'HIV Prevention' reviews the recent advances in biomedical prevention interventions (male circumcision, treatment for prevention, and preexposure chemoprophylaxis) before discussing behavioral interventions: information and education campaigns (IECs), HIV testing and counseling (HTC), school-based interventions, and conditional cash transfers. The Section 'AIDS Treatment' reviews briefly the literature on adherence to treatment before presenting the evidence about the socioeconomic benefits of antiretroviral treatment. Before the Section 'Conclusion', the article addresses, at the intersection between AIDS treatment and HIV prevention, the issue of disinhibition behaviors.

Determinants of HIV Transmission

This article does not focus on the biological determinants of HIV transmission but rather on the behavioral and socioeconomic determinants of HIV infection.

Behavioral Determinants

Gender and marriage

An alarming demographic trend in developing countries has been the steadily increasing percentage of adolescents and women who are HIV positive. If globally, 50% of all people living with HIV are women, in sub-Saharan Africa, that proportion rises to 61% and young women (15–24 years) are 3–6 times more likely to be infected than men in the same age group. These patterns have been identified as reflecting marriage patterns and risk: women are marrying younger than men and are often initiating sexual activity earlier, but women are also biologically more vulnerable to HIV infection. Several researchers argue that early marriage by females presents an important risk factor for HIV infection that is generally not being addressed and that could be contributing to the increase in HIV among this relatively large segment of the population (almost a third of girls between the ages of 10 and 19 in developing countries marry before their 18th birthday). Using data from 22 Demographic and Health Surveys (DHS) conducted in Africa, Latin America, and the Caribbean, these researchers conclude that two main factors increase the vulnerability of young brides to HIV infection: (1) marriage dramatically increases the frequency of unprotected sex for

 doi:10.1016/B978-0-12-375678-7.00113-9

most young brides and (2) many young brides marry older men, who are more likely to be HIV positive, because of their longer sexual activity.

Another study documents the increased risk of HIV infection for young married females by comparing prevalence data among the partners of young married females and the boyfriends of unmarried females the same age who are seropositive. It reports that in Kenya 30% of male partners of young wives are HIV positive, whereas only 11.5% of partners of unmarried females the same age are seropositive.

Yet another draws the opposite conclusion. The analysis done in this study, based on DHS in Ghana, Kenya, and on cross-country comparisons, suggests that late marriage and a long interval between first sex and first marriage are risk factors for HIV infection. Other researchers use data from five DHS that include HIV testing for a nationally representative sample (Burkina Faso (2003), Cameroon (2004), Ghana (2003), Kenya (2003), and Tanzania (2003–04)) to assess the question empirically. Overall, except in Cameroon, their results do not support the hypothesis that early marriage increases the HIV risk for women. Getting married at an early age does not seem to put young married women at any greater risk of contracting HIV than women their age who do not get married. However, except in Burkina Faso, marriage does not seem to protect women against HIV either.

One study focuses on the risk associated with remarriage. Using DHS nationally representative data from 13 sub-Saharan African countries, it concludes that, in almost all of the countries examined, there are high rates of remarriage and these remarried individuals have significantly higher rates of HIV prevalence than the adult population in general and that of other married individuals. It stresses that this relationship is not necessarily causal, but that remarried individuals constitute a large segment of the population that is highly vulnerable to HIV/AIDS and has not been clearly identified as such by the existing prevention efforts. Using the same data sources, another study also investigates how reported condom use varies within and outside marriage. It reinforces and expands on previous findings that men report using condoms more frequently than women do and that unmarried respondents report that they use condoms with casual partners more frequently than married individuals report using them with their spouses. The study documents that married men from most countries report using condoms with extramarital partners about as frequently as unmarried men report using them with casual partners. Married women from most of the countries included in the study reported using condoms with extramarital partners less frequently than unmarried women reported using them with casual partners. This result is especially troubling because marriage usually ensures regular sexual intercourse, thereby providing more opportunities for a person to pass HIV infection from an extramarital partner to his or her spouse.

Serodiscordant couples

Recent research on discordant couples (couples in which only one partner is HIV positive) also shed new lights on the dynamics of HIV infection within marriage. In five countries – Burkina Faso, Cameroon, Ghana, Kenya, and Tanzania – an analysis of HIV status among discordant couples yields two findings that challenge conventional notions about HIV transmission. First, in at least two-thirds of HIV-positive couples (couples with at least one HIV-positive partner), only one partner is HIV positive. Second, in close to half of those serodiscordant couples only the woman is positive. These findings have important implications for HIV prevention policies and have been confirmed in a meta-analysis for a larger set of African countries.

A pervasive, if unstated, belief is that males are by and large responsible for spreading the infection among married and cohabiting couples. The results from the analysis of discordant couples suggest, however, that HIV prevention policies should take into account the fact that women are almost as likely to be the infected partner.

Multiple partners and concurrency

In terms of behaviors, strong emphasis has been put on the hypothesis that concurrent sexual partnerships have been and remain an important driver of the HIV epidemic, especially in southern and eastern Africa. Concurrency is defined when an act of sex with one partner occurs between two acts of sex with another partner. In a network where people engage in concurrent sexual partnerships, if one person is living with HIV, the virus can spread much more rapidly among the other partners, as at any point in time a larger number of individuals is connected through the sexual network and is susceptible of becoming infected and then transmitting the infection. This network effect is further reinforced by the fact that immediately after becoming infected with HIV, HIV-positive individuals are more infectious and at higher risk of transmitting HIV within their network. Although concurrent sexual partnerships occur everywhere in the world, they might be more prevalent or last longer in southern or eastern Africa, which might be one of the key factors explaining the higher HIV prevalence in those regions. However, the hypothesis that concurrency is one of the main drivers of the HIV epidemic is difficult to establish empirically and there is a debate as to whether the evidence is strong enough to support it. The debate focuses on the measurement of concurrency (recent surveys using improved questionnaire design show reported concurrency to be between 0.8% and 7.6% in sub-Saharan Africa), the assumptions used in mathematical models of concurrency, and on whether a correlation between HIV and concurrency can be established.

Socioeconomic Determinants

Poverty

To what extent is poverty to be blamed for the AIDS epidemic? Globally, the countries hardest hit by the AIDS epidemic are poor; within sub-Saharan Africa, however, the hardest hit countries are relatively richer. The macroeconomic evidence is discussed in more detail in another article in this Encyclopedia.

Despite the lack of evidence, poverty is still believed to be a driver of the epidemic. A number of compelling arguments have been made that would support the notion that poverty causes AIDS. A naive reason underpinning this view is that health and disease exposure are usually positively correlated with poverty: richer people live longer, are in better health, and are less exposed to the deadliest diseases in low-income

countries (diarrheal diseases, malaria, and so forth). This argument does not work in the case of HIV/AIDS, because the HIV virus is contracted very differently from other contagious diseases. Indeed, it is associated with behaviors and characteristics that are often associated with higher income, such as more concurrent partners, geographic mobility, and urbanization. One study characterizes these traits as those that are a direct function of wealth (e.g., increased demand for partners) and those that are correlated with wealth (such as residence and population density).

Another study examines empirically if higher household incomes are associated with less risky behaviors for individuals (particularly females) in Cape Town, South Africa. Females in poorer households are more likely to be sexually active and experience earlier sexual debut. They are more likely to reduce condom use when they experience economic shocks, but are less likely to have multiple partners. Males are more likely to have multiple partners when confronted with a negative economic shock. However, overall, the study does not find systematic difference in condom use at last sex by income level or the experience of economic shocks.

Education

There have been different conclusions reached about the association between HIV infection and education. There are various reasons why the association may be different, including the specific context and ways of analyzing the data but the factor that seems to have the biggest influence is the time the data was recorded relative to the stage of the HIV/AIDS epidemic in the country.

Several researchers completed two systematic reviews of studies relating to the association between educational attainment and risk of HIV infection in sub-Saharan Africa. The first review concluded that there was either no association between educational attainment and HIV infection (16 studies) or that there was a positive association between education and HIV infection (15 studies), with the exception of one case of negative association in Uganda where the response to the epidemic was the most developed.

An updated version of the review combined additional data published between 2001 and 2006 with the previous data. Overall, 44 studies did not show any statistically significant association between HIV infection and education, 20 studies showed a positive association, and in only 8 studies was there a negative association. In this updated version, there is evidence that the HIV epidemic is changing as shown by the fact that a larger proportion of studies conducted from 1996 onwards identified a lower risk of infection associated with the most educated than studies from before 1996; 7 studies showed a negative association with post-1996 data compared with only one study showing a negative association with the pre-1996 data. In addition, studies from after 1996 (5/40 studies) were less likely to show a positive association between HIV infection and the highest level of education than studies from before 1996 (15/32 studies). In studies from 1996 onwards that showed changes over time, there seemed to be a shift from strong positive associations toward weaker or negative associations between the highest levels of educational attainment and HIV infection. Additionally, HIV prevalence seemed to fall more consistently among the higher educated groups. Another study also noted a shift toward a more negative association between HIV and education between 1995 and 2003 based on the analysis, controlling for wealth, of data from serial population-based surveys in both urban and rural Zambia.

Referring to the two systematic reviews above, some researchers highlight the theory that the nature of the relationship between education and HIV infection is changing over time, whereby the early positive association between education and HIV is weakening as the epidemic matures in a particular country, though they also say that there is no hard evidence that these shifting associations can be attributed to a causal effect of education on HIV infection rates.

It was also found that there is a negative association between HIV and education among young women in an analysis of an individual-level longitudinal dataset in rural Uganda. It explores the evolution of this association over a period of 12 years and finds it changes over time. The study found no robust association between HIV/AIDS and education in 1990 but then found a negative association for young females in 2000.

Occupation

Occupation can also contribute to the risk of HIV infection and transmission. Commercial sex workers have been identified as a particularly vulnerable group. One study uses a panel set of 192 self-reported daily diaries compiled by commercial sex workers in Kenya to analyze decisions to engage in unprotected sex with clients. It finds that women who engage in transactional sex substantially increase their supply of risky, better compensated sex to cope with unexpected health shocks, particularly the illness of another household member. Women are 3.1% more likely to see a client, 21.2% more likely to have anal sex, and 19.1% more likely to engage in unprotected sex on days in which another household member (typically a child) falls ill. Similar responses are observed on days just after a woman recovers from the symptoms of a sexually transmitted infection (STI), which arguably might be seen as an exogenous shock to her ability to supply sex, or from other health problems. Women do this in order to capture the roughly 42 Kenyan shilling (US$0.60) premium for unprotected sex and the 77 shilling (US$1.10) premium for anal sex. Other studies, in very different settings, Calcutta and Mexico respectively, confirms the existence of a compensating differential and that female sex workers not using condoms obtain higher prices.

Truck drivers, migrants, and miners are also often perceived as occupations at risk. Two researchers investigate the role that mines and migration played in southern Africa. They start from the observation that Swaziland and Lesotho are the countries with the highest HIV prevalence in the world. They have in common another distinguishing feature: during the past century they sent massive numbers of migrant workers into South African mines. A job in the mines implies spending a long period away from the household of origin surrounded by an active sex industry. This creates potential incentives for multiple concurrent partnerships. Using DHS, their analysis shows that migrant miners aged 30–44 years are 15% points more likely to be HIV positive and having a migrant miner as a partner increases the probability of infection for women by

8% points. The study also shows that miners are less likely to abstain and to use condoms and that female partners of miners are more likely to engage in extramarital sex. The fact that mobility might be one of the key factors of HIV transmission is also highlighted by another study that shows a positive relationship between HIV prevalence and the volume of exportations. However, a recent study examining the effects of the early twenty-first century copper boom on risky sexual behavior in Zambian copper mining cities found that the copper boom substantially reduced rates of transactional sex and multiple partnerships in copper mining cities. Copper boom induced in-migration to mining cities appears to have contributed to these reductions.

(Micro-) Economic Consequences of HIV/AIDS

From a microeconomic point of view, the costs of the epidemic are numerous. The negative impact on labor markets has been documented. For example, using firm-level data from South Africa and Botswana, one study calculates that the value of an incident HIV infection was between 0.5 and 3.6 times the annual salary of the worker. It estimated that costs varied widely between firms and among job levels within the firm. Another studied the productivity and attendance of 54 tea workers who died or were medically retired because of AIDS between 1997 and 2002 compared with other workers. After adjusting for age and environmental factors, cases were absent from work 31 days or more often (an increase of 87%); spend 22 days more on light duty (an increase of 66%); produce an average of 7.1 kg less tea leaf per plucking day (a decrease of 17%), when compared with the control group.

One of the most devastating consequences of the HIV/AIDS epidemic is the large increase in the number of orphans. In 2008, more than 14.1 million children in sub-Saharan Africa were estimated to have lost one or both parents to AIDS. There is a large literature on the consequences of orphanhood. Summarizing it has been done elsewhere and would be beyond the scope of this article. In brief, though the results from cross-sectional studies point to a large heterogeneity in the orphan/nonorphan differential across countries, longitudinal studies who can contrast the situation of the child before and after the death of the adult generally conclude that orphans are disadvantaged in terms of schooling outcomes, even if it is not always in terms of enrollment.

Beyond orphanhood, the HIV epidemic could reduce the incentives to invest in education and affect fertility behaviors. By looking at the DHS data from 15 countries in sub-Saharan Africa, one study examines the relationship between HIV prevalence and changes in human capital investment over time and finds that areas with higher HIV prevalence experienced relatively larger declines in schooling. One of the suggested mechanisms is that a lower life expectancy reduces the incentives to invest in human capital. Another also finds that short life-spans might be one of the reasons why, even when confronted with high HIV prevalence numbers, the extent of behavior change has been limited in most African settings.

Yet another study shows evidence for the fact that HIV has had little impact on fertility, both overall and in a sample of HIV-negative women; however, it was estimated that the presence of HIV reduces the average number of births a woman gives during her lifecycle by 0.15.

HIV Prevention

Although this article focuses on the economic and behavioral aspects of the HIV/AIDS epidemic, it is worth noting that currently the field of HIV prevention is dominated by recent advances in biomedical interventions for HIV prevention. This section starts by reviewing some of these advances, with some emphasis on the behavioral responses to these advances. The discussion moves next to behavioral interventions for HIV prevention.

Biomedical Interventions

The first biomedical approach to be rigorously tested for HIV was the treatment of other STIs. As summarized in one particular study, the earliest study of the efficacy of treating other STIs on HIV incidence conducted in Tanzania suggested that when STIs are treated, HIV infection declined by almost 40% over a 2 year period. Following this result, STI treatment was included in the catalog of HIV prevention measures endorsed by the World Health Organization (WHO) and UNAIDS. However, another randomized control trial in Uganda showed contradictory results and other studies have not replicated the level of efficacy found in the initial study.

However, male circumcision has been shown to be protective and more recently, new biomedical approaches have been more successful. In particular, 'treatment for prevention' or 'test-and-treat,' and preexposure chemoprophylaxis for HIV prevention have shown promising results.

Male circumcision

The evidence showing the protective effect of male circumcision from three randomized control trials is strong. Unlike other HIV prevention strategies, male circumcision is a one-time procedure with lifelong benefits and thus potentially highly cost effective. However, till date, there is no rigorous impact evaluation of male circumcision at scale. Those would be important studies to carry not only to confirm the external validity of the randomized control trials but also to learn what are the most effective delivery mechanisms for scaling up male circumcision or to assess whether behavioral responses such as disinhibition might differ in an environment where the benefits of male circumcision have been largely publicized and where a large number of men have been recently circumcised.

Treatment for prevention

The 'treatment for prevention' approach proposes to test regularly a large fraction of the population and treat immediately those who have tested positive with antiretroviral therapies, without waiting for the AIDS symptoms to develop. By treating HIV positives immediately after they have tested, the objective is to reduce the viral load of HIV positives and therefore their infectiousness. While earlier studies advocating this approach were based on modeling, recent results from the HPTN 052 study indicate that treatment for prevention is efficacious.

Preexposure chemoprophylaxis for HIV prevention

One study also reports on recent trials evaluating preexposure chemoprophylaxis for HIV prevention. In the Center for the AIDS Programme of Research in South Africa (CAPRISA) study in South Africa, high-risk women used an applicator that delivered 1% tenofovir gel into the vaginal vault up to 12 h before, and within 12 h after intercourse. Investigators reported a 39% reduction in overall acquisition of HIV, and the maximum reduction was 54% among the most adherent women. In the Iniciativa Profilaxis Pre Exposicion or Prexposure Prophylaxis Initiative (iPrEx) study in 2010, HIV-negative men who have sex with men were given daily an antiretroviral combination, emtricitabine and tenofovir disoproxil fumarate (TDF plus FTC) for up to 2.8 years. The study recorded a 44% reduction in HIV acquisition and, as with the CAPRISA study, efficacy was strongly associated with concentrations of antiretroviral drugs, a direct marker of adherence. By contrast, the Preexposure Prophylaxis Trial for HIV Prevention among African Women (FEM-PrEP) trial of TDF plus FTC offered to high-risk women was discontinued because an equal number of infections occurred in both the placebo and treatment groups.

As with treatment for prevention, the efficacy and efficiency of preexposure chemoprophylaxis for HIV patients needs to be further established and confirmed, but if they are confirmed it would open very promising perspectives for the prevention of sexual transmission. Compared with treatment as prevention, preexposure chemoprophylaxis offers two advantages. First, there is no need for frequent and widespread testing in order to identify HIV-positive individuals. This is logistically challenging in most settings in sub-Saharan Africa, especially if one of the objectives is to detect individuals with recent HIV infections that are more infectious, but more difficult to detect with accuracy. Second, preexposure chemoprophylaxis for HIV prevention can be self-targeted by individuals who feel they are most at risk. However, both approaches require a high level of adherence in the absence of symptoms and are operationally challenging to implement considering that it has proved difficult so far to fully scale up HIV testing in the general population and access to antiretroviral treatment for all AIDS patients.

Behavioral Interventions

One study reviews 37 randomized controlled trials of HIV prevention interventions and finds only six demonstrating effects in reducing HIV incidence. Those six were all evaluating biomedical interventions (male circumcision trials, STI treatment, and care). None of the behavioral interventions reviewed demonstrated impact in reducing HIV incidence. The review suggests that lack of statistical power, poor adherence, and diluted versions of the intervention in comparison groups may have been important issues in some of the trials that did not show any results.

Information and education campaigns

IECs have been among the first behavioral interventions for HIV prevention. One researcher reviews the much touted abstain, be faithful, use condoms (ABC) campaigns in Uganda. The study concludes that the effects of such a national mass media campaign on behavior are difficult to estimate as a control group is not available. The ABC initiative in Uganda, combined with a high level of political commitment to HIV prevention, seemed to have been successful in significantly reducing the prevalence of HIV. However in mass efforts such as this, it is difficult to ascribe success to individual components (there is a debate about the relative importance of condoms in the ABC strategy), but they do provide suggestive evidence that broad-based and well supported efforts at behavior change can be effective prevention strategies. Overall, IECs by itself have not been shown to have more than a minor impact on patterns of HIV transmission and the trajectory of the epidemic. Numerous studies have shown that information alone is typically insufficient to change risk behavior. The impact of mass media campaigns tends to be short in the absence of an ongoing effort, and these campaigns can be aided by condom distribution and by more targeted education programs aimed at youth in and out of school.

HIV testing and counseling

HTC is recognized as the necessary gateway for HIV/AIDS treatment. However, the prevention benefits of individual HTC remain under discussion. One study estimates the behavioral responses by individuals to a public HIV testing program. It posits that only individuals who are surprised by the test results, i.e., low-risk individuals testing HIV positive or high-risk individuals testing negative, will change their behaviors. For those individuals HTC can lead to unexpected behaviors that might not reinforce prevention. It finds that although the aggregate effect of the testing program is quite small, the effects disaggregated by private beliefs about own risks are consistent with information elastic behavior for the average individual. It concludes that the subgroups of the population affected by HTC may have roughly offsetting behavioral responses, which may lead to little effect or possibly even perverse outcomes with regard to an objective of lowering disease transmission.

Another study finds that beliefs are an important determinant of risky behavior, with downward revisions in the belief of being HIV positive increasing risky behavior and upward revisions decreasing it. Yet another tests the hypothesis that only individuals who are surprised by the test results will change their behaviors, using STIs as objectively measured proxies for unsafe sexual behavior. On the one hand, individuals who believed they were at low risk for HIV before testing, are nine times more likely to contract an STI following an HIV-positive test, indicating riskier sexual behavior. On the other hand, individuals who believed they were at high risk for HIV have an 84% decrease in their likelihood of contracting an STI following an HIV-negative test, indicating safer sexual behavior. When HIV tests agree with a person's belief of HIV infection, there is no statistically significant change in contracting an STI. Using the randomly assigned incentives and distance from results centers as instruments for the knowledge of HIV status, one researcher finds that sexually active HIV-positive individuals who learned their results are 3 times more likely to purchase condoms 2 months later than sexually active HIV-positive individuals who did not learn

their results. However, there is no significant effect of learning HIV-negative status on the purchase of condoms.

Meta-analyses of the prevention benefits of HTC conclude that HIV counseling and testing appears to provide an effective means of secondary prevention for HIV-positive individuals but is not an effective primary prevention strategy for uninfected participants and that, overall, there is only moderate evidence in support of HTC as an effective prevention strategy.

Joint couple or partner testing appears to have stronger prevention benefits, especially in the case of serodiscordant couples. However, despite the importance of couple testing for treatment and prevention, there are few successful experiences of HTC programs reaching couples. Recent evidence on the effectiveness of ART for the prevention of HIV transmission among couples makes this a key intervention of prevention programs in generalized epidemic countries. Recent evidence from Rwanda suggests that pay-for-performance schemes at the health facility level can be an effective intervention to target discordant couples for HTC.

School-based interventions

The school environment offers a useful platform to deliver HIV information and prevention messages to individuals just before or as they start their sexual life.

Several researchers analyzed results from a randomized evaluation comparing two different HIV prevention interventions and one economic intervention, and their impact on the students in certain behaviors considered to be risk factors for HIV infection. They tested three different types of school-based interventions in rural Kenya. One intervention involved training teachers in the national HIV/AIDS curriculum for them to present to their students. The second intervention consisted of students being encouraged to debate the benefits of using condoms and write essays on ways to protect themselves against HIV. The third intervention involved lowering the cost of schooling by providing school uniforms to students attending school as a way to get students to stay in school longer. To measure effectiveness, the researchers primarily evaluated teenage childbearing as a proxy for unprotected sex, the main risk factor for HIV/AIDS in Africa. They also collected information on knowledge, attitudes, and behavior regarding HIV/AIDS. The teacher training was found to have little impact on teen childbearing, students' knowledge, and self-reported sexual activity and condom use. The debate and essay intervention increased self-reported condom use, but not self-reported sexual activity. Paying for uniforms reduced dropout rates by 15%, resulted in an almost 10% decrease in teen childbearing, girls were 12% less likely to be married, and boys were 40% less likely to be married.

The UK Department for International Development (DFID) trial (2004) in rural Tanzania evaluated the impact of an intervention aimed at changing the knowledge and sexual behavior of adolescents on HIV rates, other STIs, unintended pregnancy and adolescents' knowledge, and reported attitudes and behaviors. The intervention included an in-school teacher-led, peer-assisted sexual and reproductive health education component, training for health workers to make reproductive health services at the clinics more youth-friendly, community-based condom promotion, and periodic community activities promoting sexual health. Comparing the communities that received the interventions with the control communities showed that the intervention communities had statistically significant improvement in knowledge and reported sexual attitudes for both males and females. Males also reported delayed sexual debut, fewer sexual partners, and more condom use at last sex. However, there was no evidence of a consistent impact of the intervention on biological outcomes including HIV incidence, other STIs, and unintended pregnancies.

A review of 11 quasiexperimental designs that measured the impact of a variety of school-based HIV prevention interventions in sub-Saharan Africa reinforce the finding from the DFID trial that behavior is more difficult to change than knowledge.

Although general HIV knowledge may not often result in behavior change, another study shows that specific information that distinguishes the levels of HIV risk may be more useful in changing behavior. The study rigorously tests an information campaign telling teenagers about the relative risks of different types of partners, based on their HIV infection rates. The objective of the campaign was to make teenagers aware of the relative risks of partners of different ages in the hope that they will take these different levels of risk into account when choosing a partner. As a result of the campaign, the incidence of cross-generational pregnancies among the treatment group decreased by 61% while intragenerational pregnancies remained stable. This information on the relative risks of different partners resulted in a sizable decrease in unprotected sex between older men and teenage girls but without an increase in unprotected sex between teenage boys and girls. In contrast, another program that only gave general information about HIV risk had no impact on the incidence of unprotected sex as measured by pregnancy rates.

Conditional cash transfers

Conditional cash transfer programs have become an increasingly popular approach for incentivizing socially desirable behavioral change. The principle of conditionality – making payments contingent, for example, on a minimal level of schooling attendance or preventative care use – distinguishes conditional cash transfer programs from more traditional means-tested social programs. The evaluation of conditional cash transfer programs have shown that they can be effective at raising consumption, education, and preventative health care, as well as actual health outcomes. Similarly, 'contingency management' approaches have shown important substance abuse reductions by conditioning rewards on negative tests for drug or alcohol. The evidence on the efficacy of conditional cash transfers for STI or HIV prevention is still unfolding and remains limited. In Malawi, small financial incentives have been shown to increase the uptake of HTC. Another study in Malawi, conducted a conditional cash transfer program for adolescents in which the cash transfer was conditional on school attendance but which, in addition to increased enrollment and attendance also caused a reduction in HIV and herpes simplex virus type 2 (HSV-2) incidence. HIV prevalence among program beneficiaries was 60% lower than the control group (1.2% vs. 3%). Similarly, the prevalence of HSV-2 (which is the common cause of genital herpes) was more than 75% lower in the combined treatment group (0.7% vs. 3%). No significant differences were detected between those offered

conditional and unconditional payments. In addition, cash payments offered to the girls who had already dropped out of school at the beginning of the trial made no difference on their risk of HIV or HSV-2 infection. The same program also led to a modification of self-reported sexual behaviors with adolescent girls having younger partners.

Till date, two studies evaluated conditional cash transfers in which the conditionality is attached to negative test results for STIs. In Malawi, one study tested an intervention promising a single cash reward in 1 year's time for individuals who remained HIV negative. This design had no measurable effect on HIV status, but the number of seroconversions in the sample was very small and statistical power was therefore low. The Rewarding STI Prevention and Control in Tanzania (RESPECT) study evaluated a randomized intervention that used economic incentives to reduce risky sexual behavior among young people aged 18–30 years and their spouses in rural Tanzania. The goal was to prevent HIV and other STIs by linking cash rewards to negative STI test results assessed every 4 months. The study tested the hypothesis that a system of rapid feedback and positive reinforcement using cash as a primary incentive to reduce risky sexual behavior could be used to promote safer sexual activity among young people who are at high risk of HIV infection. Results of the randomized controlled trial after 1 year showed a significant reduction in STI incidence in the group that was eligible for the US$20 quarterly payments, but no such reduction was found for the group receiving the US$10 quarterly payments. Further, though the impact of the Conditional Cash Transfers (CCTs) did not differ between males and females, the impact was larger among poorer households and in rural areas. Although the results from those studies are important in showing that the idea of using financial incentives can be a useful tool for preventing HIV/STI transmission, this approach would need to be replicated elsewhere and implemented on a larger scale before it could be concluded that such conditional cash transfer programs, for which administrative and laboratory capacity requirements are significant, offer an efficient, scalable, and sustainable HIV prevention strategy.

AIDS Treatment

Antiretroviral therapy (ART) has dramatically reduced morbidity and mortality for people living with HIV/AIDS. By the end of 2010, an estimated 6.6 million people in low- and middle-income countries received ART. In sub-Saharan Africa, approximately 47% of the 14.2 million eligible people living with HIV were on ART. This is an extraordinary achievement, considering that as recently as 2003, relatively few people living with HIV/AIDS had access to ART in Africa. A total of 2.5 million deaths have been averted in low- and middle-income countries since 1995 due to the ART being introduced, according to new calculations by UNAIDS.

Adherence to Treatment

Medical research has established that a minimum level of adherence to antiretroviral drug (ARV) treatment of 95% is necessary to achieve significantly better health outcomes as assessed by the viral load, immune system, and occurrence of opportunistic infections. Nonadherence predicts disease progressions and survival rates, and increases the risk of transmission of drug-resistant viruses. Failure to achieve proper adherence to treatment is thus both an individual and collective risk.

Determinants of adherence depends on several factors such as the treatment regimen (which can be quite complex and include food restrictions, specific schedules, etc.), disease characteristics, the quality of the patient–provider relationship, or the clinical setting. Sociodemographic factors do not consistently predict adherence behavior. The meta-analysis on socioeconomic status as a determinant of adherence finds that while the relationship is weak, there is generally a positive association between income, education, or employment status and adherence. It is worth noting that adherence is not found to be consistently lower in developing countries, and largely depends on access to treatment and financial barriers. When therapy is fully subsidized in developing countries, it can be at least as good as in developed countries.

Even when treatment per se is free, transportation costs to the health facility to get a prescription refilled are found to be a powerful barrier to adherence. Moreover, patients have to make 'impossible choices' between competing claims: transport costs and good nutrition of the patients compete with schooling fees or medical costs for children, food for the rest of the family, etc. As already mentioned, malnutrition can be an obstacle to adherence.

Several interventions aiming to improve adherence have been evaluated. For example, weekly Short Message Service (SMS) reminders have been shown to increase the percentage of participants achieving 90% adherence to ART by approximately 13–16% compared with no reminder and were also effective at reducing the frequency of treatment interruptions.

The Economic Benefits of Antiretroviral Treatment

The most immediate benefit of the scaling up of antiretroviral treatment is a reduction in mortality and morbidity. A second-order set of benefits is related to the increase of labor supply and productivity of AIDS patients and their family members as well as related changes in income, time allocation, and school participation of children.

A study from Botswana provides evidence on the link between a worker's health status (measured by his/her cluster of differentiation 4 (CD4) count) and absenteeism in a given month, using measurements of the CD4 count at 0, 6, and 12 months after treatment initiation. The estimates provide robust evidence of an inverse V-shaped pattern in worker absenteeism around the time of ARV treatment inception. In the 1–5 years before the start of treatment, there is no difference in the rate of worker absenteeism before the start of treatment. At 12–15 months before the start of treatment, there is a sharp increase in absenteeism to approximately 20 days in the year before the start of treatment and a peak of 5 days in the month of treatment initiation (absence rate of 22%). Recovery is quick within the first year. At 1–4 years after treatment starts, treated workers have low rates of absenteeism similar to

nontreated workers. In Tamil Nadu, India, at 6 months after initiation of ART, AIDS patients were 10% points more likely to be economically active and worked 5.5 additional hours per week.

On the basis of data from rural Kenya, several researchers compare the change in the extensive and intensive margins of labor supply of patients on ARV and their household members. They document a 20% increase in the likelihood of patient participating in labor force and a 35% increase (7.9 h) in weekly hours worked within 6 months of treatment. Young boys in treated patients' households work significantly less after treatment initiation, whereas girls and adult household members do not change their labor supply. In the same setting in Kenya, with ARV treatment, females increase time for water and firewood collection, but decrease time on medical care translating into a lower burden on children with less time spent on housework and chores. Finally, based on the same longitudinal survey data from Kenya, weekly hours of school attendance of children, particularly for girls, in the patient's household increased by more than 20% within 6 months after ARV treatment was initiated for the adult patient. In Kenya, there is weaker evidence that the short-term nutritional status of young children also improves. However, in a recent study in Zambia, the researcher finds that adult access to ART resulted in increased weight-for-age and decreased incidence of stunting among children younger than 60 months of age.

At the Intersection of Prevention and Treatment: Disinhibition Behaviors

Part of the economics literature on HIV/AIDS has investigated disinhibition – or risk compensation – behaviors. The main proposition of this literature is that people may alter their behavior in response to perceived changes in risk. In the specific case of HIV/AIDS, the focus has been mainly related to the increased access to antiretroviral treatment. The concern is that increased access to ART may lead to a decrease in the perceived risk and costs of contracting HIV and, as a consequence, may lead to an increase of risky sexual behaviors. Such disinhibition behaviors, if large enough, may (at least partially) offset the benefits of scaling up access to ART. This conjecture is supported by several studies in the US and Europe, which have identified an upward trend in risky sexual behaviors since the introduction of ART in 1996. More specifically, an association has been identified between decreased concern about HIV due to ART availability and unprotected sex, and in particular among men who have sex with men.

Investigations of disinhibition behaviors in sub-Saharan Africa are limited. Studies exploring directly the behaviors of ART patients have generally concluded that there was no evidence of increase in risky behaviors after the ART initiation, even if sexual activity increased. One of the earliest studies looked at change in the use of condom by sex workers in Nairobi, Kenya. This analysis provided at least some suggestive evidence that condom use by sex workers decreased when 'fake' cures of AIDS were announced. Such a pattern is consistent with disinhibition behaviors, although the result may not be generalizable to the general population as it uses a much selected segment of the population. Another study used population-based surveys to test risk compensation behavior in the general population in a sub-Saharan African context. The researchers observed that in Kisumu (Kenya), ART-related risk compensation and the belief that ART cures HIV were associated with an increased HIV seroprevalence in men but not women. Others study the effect of increased access to ART on self-reported risky sexual behavior, using the data collected in Mozambique in 2007 and 2008. Controlling for unobserved individual characteristics, the findings support the hypothesis of disinhibition behaviors. In particular, risky behaviors are more positively associated with efficacious ART for family members of HIV-positive persons and for individuals from neighboring households, whereas disinhibition behaviors are not found among AIDS patients themselves.

Although disinhibition might more directly be a consequence of the availability of ART, disinhibition behaviors could also be present as a consequence of HIV prevention interventions. For example, one study advances that HTC might be effective in persuading HIV-positive individuals to reduce their risky behaviors and the risk of transmission of HIV to their partners, but potentially leads to disinhibition among those who receive an HIV-negative test result. Disinhibition should be considered and investigated in the case of male circumcision, treatment for prevention, and preexposure prophylaxis. In the case of male circumcision, it is possible that as a consequence of male circumcision – which is protective, but only to a certain extent – male individuals and their partners opt for less safe sexual practices and, for example, become less likely to use condoms or more likely to engage in concurrent partnerships. Another study discusses compensating behaviors related to male circumcision. The assessment is that the current empirical evidence does suggest that disinhibition is unlikely to substantially reduce the effectiveness of medical male circumcision. This assessment is based on the evidence from self-reported sexual behaviors of study participants in the randomized control trials that have established the efficacy of medical male circumcision. It would be important to assess the possibility of disinhibition from male circumcision interventions at scale.

Overall, it is fair to conclude that the evidence on disinhibition behaviors is limited and inconclusive. Several studies have provided a comprehensive review, with studies finding evidence of disinhibition and others not. The evidence is even more limited in sub-Saharan Africa but the potential risks associated with disinhibition on a large scale are important enough to be taken into consideration in further studies.

Conclusion

After reviewing the behavioral and socioeconomic determinants of HIV transmission, this article has focused on HIV prevention intervention and AIDS treatment. There is a tendency to present prevention and treatment as alternatives competing for scarce (donor) resources. However, HIV prevention remains crucial. Only by sustaining recent reductions in mortality and bringing down the number of new infections will the total number of people with HIV finally decline and will an AIDS transition be attainable.

It has been stressed that behavioral responses are very important mediators of HIV transmission and of the efficacy of HIV prevention and AIDS treatment. Currently, the field of HIV prevention is dominated by recent advances in biomedical interventions for HIV prevention such as male circumcision, treatment for prevention, and preexposure chemoprophylaxis. Though these interventions represent important breakthroughs, it is important to keep in mind potential behavioral responses, such as disinhibition to these interventions as well as the role that incentives can play. Further, it will be important to evaluate those interventions at scale. Such impact evaluations would not only confirm the external validity of the randomized control trials but also would allow learning what are the most effective delivery mechanisms for scaling up those interventions.

See also: Health Status in the Developing World, Determinants of. HIV/AIDS, Macroeconomic Effect of. Infectious Disease Externalities. Sex Work and Risky Sex in Developing Countries

Further Reading

Baird, S., Chirwa, E., McIntosh, C. and Özler, B. (2010). The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health Economics* **19**(S1), 55–68, doi:10.1002/hec.1569.

Baird, S., Garfein, R., McIntosh, C. and Özler, B. (2012). Impact of a cash transfer program for schooling on prevalence of HIV and HSV-2 in Malawi: A cluster randomized trial. *Lancet,* doi:10.1016/S0140-6736(11)61709-1.

Duflo, E., Dupas, P., Kremer, M. and Sinei, S. (2006). Education and HIV/AIDS prevention: Evidence from a randomized evaluation in western Kenya. *World Bank Research Policy Working Paper No. 4024*. Washington, DC: The World Bank.

Dupas, P. (2011). Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics* **3**, 1–34.

Evans, D. K. and Miguel, E. (2007). Orphans and schooling in Africa: A longitudinal analysis. *Demography* **44**, 35–57.

Eyawo, O., de Walque, D., Ford, N., et al. (2010). HIV status in discordant couples in sub-Saharan Africa: A systematic review and meta-analysis. *Lancet Infectious Diseases* **10**, 770–777.

Fortson, J. G. (2008). The gradient in sub-Saharan Africa: Socioeconomic status and HIV/AIDS. *Demography* **45**(2), 303–322.

Fortson, J. G. (2009). HIV/AIDS and fertility. *American Economic Journal: Applied Economics* **1**(3), 170–194.

Fortson, J. G. (2011). Mortality risk and human capital investment: The impact of HIV/AIDS in sub-Saharan Africa. *Review of Economics and Statistics* **93**(1), 1–15.

Gertler, P. J., Shah, M. and Bertozzi, S. M. (2005). Risky business: The market for unprotected commercial sex. *Journal of Political Economy* **113**, 518–550.

Graff Zivin, J., Thirumurthy, H. and Goldstein, M. (2009). AIDS treatment and intrahousehold resource allocation: Children's nutrition and schooling in Kenya. *Journal of Public Economics* **93**(7–8), 1008–1015.

Granich, R. M., Gilks, C. F., Dye, C., De Cock, K. M. and Williams, B. G. (2009). Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: A mathematical model. *Lancet* **373**(9657), 48–57, doi:10.1016/S0140-6736(08)61697-9.

Lakdawalla, D., Sood, N. and Goldman, D. (2006). HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics* **121**(3), 1063–1102.

Oster, E. (2012a). HIV and sexual behavior change: Why not Africa? *Journal of Health Economics* **31**(1), 35–49.

Oster, E. (2012b). Routes of infection: Exports and HIV incidence in sub-Saharan Africa. *Journal of the European Economic Association* **10**(5), 1025–1058.

Over, M. (2011). *Achieving an AIDS transition: Preventing infections to sustain treatment*. Washington, DC: Center for Global Development.

Pop-Eleches, C., Thirumurthy, H., Habyarimana, J. P., et al. (2011). Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: A randomized controlled trial of text message reminders. *AIDS* **25**(6), 825–834.

Robinson, J. and Yeh, E. (2011). Transactional sex as a response to risk in western Kenya. *American Economic Journal: Applied Economics* **3**(1), 35–64.

Thirumurthy, H., Graff Zivin, J. and Goldstein, M. (2007). The economic impact of AIDS treatment labor supply in western Kenya. *Journal of Human Resources* **43**(3), 511–552.

Thornton, R. (2008). The demand for and impact of learning HIV status. *American Economic Review* **98**, 1829–1863.

de Walque, D. (2007). How does the impact of an HIV/AIDS information campaign vary with educational attainment? Evidence from rural Uganda. *Journal of Development Economics* **84**, 686–714.

de Walque, D., Dow, W. H., Nathan, R., et al. (2012). Incentivizing safe sex: A randomized trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania. *BMJ Open* **2**, e000747, doi:10.1136/bmjopen-2011-000747.

de Walque, D., Kazianga, H. and Over, M. (2012). Antiretroviral therapy perceived efficacy and risky sexual behaviors: Evidence from Mozambique. *Economic Development and Cultural Change* **61**(1), 97–126.

Wilson, N. (2012). Economic booms and risky sexual behavior: Evidence from Zambian copper mining cities. *Journal of Health Economics* **31**(5), 797–812.

Relevant Websites

http://www.iaen.org/
International AIDS Economics Network.

http://www.iasociety.org/
International AIDS Society.

http://www.unaids.org/en/
Joint United Nations Program on HIV/AIDS.

Home Health Services, Economics of

G David and D Polsky, University of Pennsylvania, Philadelphia, PA, USA

Introduction

Throughout the nineteenth century in the western world, home health care (HHC) existed to care for new mothers and those with infectious diseases. In the mid-twentieth century, HHC began to transform, as the proportion of older people in the general population steadily increased and with it the need for care for chronic degenerative diseases. The emergence of new medical innovation allowed the shift from facilities to the patient's residence and demographic trends such as a decrease in the size of families and a decline in families' colocation changed the social attitudes toward formal care. Finally, rising hospital costs led government to favor lower cost settings.

Although the trends described above are shared by most developed countries, the size of the home health sector as well as the way in which it is delivered, financed, and regulated varies across countries. Spending on home care accounts for a large proportion of resources spent on long-term care. According to 2009 data published by the Organization for Economic Co-operation and Development (OECD), spending on long-term care as a percent of gross domestic product (GDP) was as high as 2.72% in Denmark and as low as 0.84% in Spain. The US spends 0.98% of GDP on long-term nursing care, and approximately 40% of that on HHC.

Home health services are provided by agencies that are primarily engaged in providing skilled nursing or medical care in the home, under the supervision of a physician. The services provided can range from assisting with basic 'activities of daily living' (bathing, dressing, getting out of bed, and feeding oneself) to providing complex care. Skilled care can include audiology and speech pathology, dietary and nutritional services, drug services, home health aide, laboratory, medical social services, nursing, occupational therapy, and physical therapy.

Unlike the US, where a mix of public and private home health agencies (HHAs) provides both skilled nursing as well as home aide services, the organization of home health services is different in Europe. In some countries there is a divide, where skilled services are provided by the health care sector, whereas home aide services are provided by social services (e.g., Norway, Finland, and Sweden). In other countries, both skilled and nonskilled care are provided by either municipalities (e.g., UK, France, Italy, and Spain) or covered under social insurance and provided by a mix of governmental and private agencies (e.g., Germany and the Netherlands).

This article will both discuss the salient features of the home health industry, with a focus on the institutional structure in the US. The authors emphasize how these features pose challenges for economic analysis of competition, regulation, and integration. The typical way economists analyze hospital or nursing home markets to not always apply for HHC markets. In particular, the location in which services are rendered – the patient residence – changes the nature of competition, the ability to engage in effective monitoring, and the benefits of organizing services along the health care continuum.

Home Health Care Industry

Freestanding home care services of all types accounted for US$68.3 billion in annual expenditures in 2009, approximately 3% of all personal health care spending. The largest payer of HHC services is Medicare, accounting for 41%. The total coverage from all government sources is 80% because Medicaid covers 24% and other government sources cover 15%. Private insurance accounts for only 8% and most of the remainder is paid from out-of-pocket expenditure.

The Bureau of Labor Statistics estimated that 1 071 960 persons were employed in home health service sector in 2012. The central figure within home care agencies is the registered nurse. The RNs comprise approximately 15% of total home care employment and receive an annual median salary of US$63 850. Approximately 59% of jobs in this segment are in low-income service occupations, mostly home health aides and personal and home care aides. Home health aides, comprising the largest fraction of employees at 35%, receive an annual median salary of $20 560. Nursing and therapist jobs also account for substantial shares of employment in this segment. It should be noted that formal home health is just a fraction of home caregiving; more than one in three US households (an estimated 48.9 million caregivers more than age 18) are informal caregivers for a person older than age 18, with an additional 16.8 million caring for children or both children and adults, for a total of 65.7 million individual caregivers.

From an organizational perspective, there are 10 422 Medicare-certified HHAs. Approximately 85% of them are freestanding; the remainder are predominantly affiliated with hospitals. Approximately 70% of the freestanding HHAs were classified as proprietary or for-profit and the remaining freestanding HHAs were nonprofit agencies, including Visiting Nursing Associations, government or voluntary agencies, public agencies (typically run by the state or local government) and private nonprofits. There are HHAs that do not certify with Medicare but data on these facilities are sparse. HHC agencies are distinct from other home care organizations such as hospices where the focus is on care of terminally ill patients and their families, home care aide agencies where the focus is on assistance with activities of daily living, and home care equipment providers. Home-hospice, home infusion therapy, and home dialysis are outside the scope of this article. HHC agencies are also a distinct from other organized settings for postacute care. These other settings include skilled nursing facilities (SNFs), long-term care hospitals, and inpatient rehabilitation facilities.

The service lines of these HHAs are separated into personal care services (care provided by home health aids or personal

care for the elderly such as bathing, dressing when there is no concurrent need for skilled care, and homemaking), which are more likely to be covered by Medicaid and services to treat an illness or injury to regain independence which are covered by Medicare. Medicare home health services consist of skilled nursing care by a registered nurse or licensed practical nurse with supporting services by home health aides; therapy services including physical therapy, occupational therapy, and speech–language therapy; medical social services; and medical supplies. Home health visits typically last approximately 45 min. A typical clinical episode of care may be approximately a month, but payment is fixed as long as the clinical episode does not exceed 60 days. If the clinical episode needs to be extended beyond 60 days, there can be sequential 60-day payment episodes through recertification. Although there is great variation, there are approximately 12 visits on average during a typical clinical episode. Most of this article focuses on the Medicare service line, which is the largest segment of HHC services. However, Medicaid's role in home health has been growing rapidly as long-term nursing care is moving away from institutional settings and into community-based settings.

The Value Proposition for Home Health Services

Given that care in the home is less expensive to Medicare than care in a hospital or a SNF – in 2009, the average Medicare charges on a per day basis for hospital came to US$6 200, SNF was US$622, and home health Medicare charges averaged US$135/day – there are great opportunities for value in home health. Value is derived when home health can cost-effectively substitute for these more intensive locations of care or when home health services can play an important role in avoiding rehospitalizations during postacute care or hospitalizations for chronically ill patients. However, because standards for what constitutes appropriate or necessary care do not exist, the value of what gets delivered in home health on the margin is often questioned.

Empirically, value in home health is typically shown for select conditions where the evidence for home health is strongest (i.e., diabetes, chronic obstructive pulmonary disease, and congestive heart failure patients). Measured in 1995, savings accrued when home health was successfully substituted for more intensive sites of care in cases of pediatric AIDS (US$2263 per hospital per day vs. US$531 at home per day), respiratory care (US$188 909 per year at hospital vs. US$109 836 per year at home), and hip-fracture (savings of US$2300 per incident if home health used in conjunction with hospital care). For the majority of conditions, however, there are few studies that even attempt to demonstrate value. As a result, great geographic variations exist in home health.

Holding HHAs accountable for outcomes may be an avenue to improve both the quality of home health services and patient outcomes in general, but the measurement and assessment of outcomes in home health is a challenge. Although outcomes can be measured from the Outcome and Assessment Information Set (OASIS), there is no consensus regarding the outcomes that capture the effectiveness of home health. And more importantly, because outcomes are typically measured within HHC, home health outcomes are not compared to the alternative of reduced access to home care. This makes it difficult to assess whether improvements in staffing or increasing the number or coverage of agencies would, in fact, spillover to other services, for example, through reduced hospitalization rates.

Reimbursement under Medicare

Reimbursement Mechanism

To be eligible for Medicare's home health benefit, beneficiaries must need part-time (fewer than 8 h per day) or intermittent (temporary but not indefinite) skilled care to treat their illnesses or injuries and must be unable to leave their homes without considerable effort. Medicare does not require beneficiaries to pay copayments or a deductible for home health services. In the Balanced Budget Amendment of 1997, Medicare changed from paying a fee per home health visit to a Prospective Payment System (PPS). Under the PPS system, which began in 2000 after a 3-year interim system, Medicare pays a fixed amount for HHC in 60-day episodes. These Medicare payment episodes begin when patients are admitted to HHC. Patients who complete their course of care before 60 days have passed are discharged. If they do not complete their care within 60 days, another episode starts and Medicare makes another episode payment. As long as they meet the eligibility standards for the benefit, beneficiaries may receive an unlimited number of consecutive home health episodes. Medicare adjusts the payment based on several factors including measures of patients' clinical and functional severity and the use of therapy during the home health episode. This case-mix adjusted payment rate is similar to the Medicare SNF and inpatient hospital PPS's. However, a major difference among the systems is the unit of payment. SNFs are paid by the day, whereas the home health PPS pays by the 60-day episode. In 2009, the Medicare payment per user of home health was US$5748. This was up from US$3803 in 2002.

Yet the system will continue to be changed as savings are sought within Medicare. One reason for this is that HHAs continued to be paid by Medicare significantly above cost, with margins of 16.6% in 2007 though there have been recent changes that include a payment-rate update that represents a 5% decrease and caps on outlier payments. Several changes were part of health care reform that have expanded the role of the physician so that a physician face-to-face encounter is now a requirement for certification of eligibility for home health services, the final rule provided that the encounter must occur with the 90 days before start of care, or within the 30 days after. This is a means of increasing physician accountability and providing an additional check on beneficiaries' eligibility for home health benefits.

Incentives Created by Reimbursement Mechanism

The shift from per-visit payment to prospective payment shifted incentives from rewarding the number of visits, which can lead to a more intensive pattern of visits, to rewarding a limited number of visits within an episode, but encouraging expansion through the number of episodes. Care patterns

appear to be very sensitive to the payment system. For example, during the interim period (Interim Payment System (IPS)) between the end of per-visit payment and the beginning of PPS, there was an annual reduction of 1.3 million HHC episodes with a 30% decline in the number of Medicare-certified HHAs. However, PPS did not have the same disincentive for visits as the IPS as the PPS scheme includes lower payments if 5 visits are not achieved and enhanced payments when therapy visits exceed 10 visits. As a result the transition from IPS to PPS has resulted in an increase in both episodes and agencies.

Various changes to reimbursement design illustrate the influence of incentives in determining where Medicare beneficiaries receive postacute care. The results of switching from fee-for-service (FFS) to PPS were profound, suggesting highly elastic patterns based on reimbursement design. When the Balanced Budget Act (BBA) was passed and the IPS was implemented after years of FFS, the industry changed rapidly. In addition to heavily cutting reimbursement rates, the IPS ended a period in which providers had no little incentive to control the amount of service per user. After the IPS's enactment, a trend emerged in which patients were shifted from HHAs and SNFs to having no formal care. Also, because reimbursement was not case-mix adjusted, the IPS created backward incentives for HHAs to cut service to high-cost patients. HHAs that did not use strategic admission of low-cost patients suffered the risk of insolvency. Furthermore, the scale of the industry responded quickly and intensely: The number of active agencies decreased by 20% after IPS. Between 1996 and 1999, the number of new agencies declined by a drastic 86%, and the number of terminated agencies increased by 523%. In 1996, the ratio of terminated HHAs to new HHAs was less than 1, but 1997 after the IPS, terminated HHA's outnumbered new HHA's 9 to 1. The industry is highly reactive to reimbursement changes, and the roughly 30% of the decline in HHAs between 1997 and 2001 has been attributed to changes in Medicare home health coverage and reimbursement enacted as part of the BBA.

The PPS, introduced in October 2000, continued prospective payment but adjusted for case-mix when determining reimbursement payments. By replacing the IPS with the PPS's risk-adjusted episode system, Medicare alleviated HHAs' financial risk of treating patients. The PPS reversed some of the IPS's impacts: From 1999 to 2002, the number of new HHAs increased 78% and the number of HHA termination fell by 88%. By 2002, the PPS had stopped the contraction of HHAs providers, and more agencies were added than terminated. However, throughout both the IPS and PPS, proprietary and freestanding HHAs experienced greater volatility. Not until 2009, with 10 581 agencies, did the number of HHAs surpass that of 1997. With respect to quality, the Office of Inspector General found that the change in the reimbursement system did not lead to increased use of hospital and ER services. Recently MedPAC has responded to the HHC industry's high margins (16.6% in 2007), which it feels undermine the efficiency goals of a PPS. Consequently, it has recommended cuts in reimbursement rates.

Even before the BBA there was strong evidence of drastic industry responses to incentive changes. A 1987 court case, Duggen versus Bowen, resulted in changes in reimbursement and incentive changes. Before 1986, Medicare suffered from excessive administrative complexity and unreliable reimbursements. The lawsuit's success contributed to increased annual Medicare home health outlays and a doubling of the number of Medicare-Certified HHAs between 1989 and 1996. Additionally, growth of the HHC services industry was 18%, whereas it was 7.2% for the total US health care services.

Managed Care in Home Health

After the passage of the Medicare Modernization Act, Medicare Advantage enrollment has increased rapidly. As of February 2010, 25.2% of Medicare beneficiaries were enrolled in Medicare Advantage.

Incentivized by the increasingly competitive nature of the health care industry, HHAs have entered into managed care provider networks. However, the extent to which HHAs participate in managed care is largely unstudied. An early study by Center for Medicare and Medicaid Services (CMS's) predecessor, Health Care Financing Administration, found that managed care patients used less home health resources but also had worse outcomes when compared with FFS patients. Further research is needed on the effect of managed care plans on outcomes in HHC.

The Nature of Competition

The most salient distinctive feature of HHC is the site of care. With services delivered in the home rather than in a centralized facility, the nature of competition is different. For hospitals and physician offices, location provides a degree of market power that does not exist for HHAs because the consumers do not face travel costs when receiving home health services. Travel costs, in both emergencies and nonemergencies, lead most consumers to prefer a closer provider and similarly for admitting and referring physicians. Without location as a natural barrier to competition, home health markets are expected to be highly competitive.

Quality of care in home health may be more important for agency choice because consumers do not need to tradeoff quality off against distance, as is the case for hospitals, nursing homes, ambulatory surgery centers, and other facilities. Studies of hospitals and other health care facilities have shown distance to be an important factor in the choice of health care provider. For example, the effect of distance to provider for mental health institutions was found to overshadow other incentives to initiate treatment. Similarly, patients often prefer to receive care at a near hospital, even if it has higher mortality rates and less experience with certain procedures. Distance to nearest hospital was shown to significantly impact utilization of preventative care, psychiatric, geriatric, and elective surgery and had a much stronger effect on the probability of hospital choice than waiting time. Moreover, physicians typically mention in surveys that the hospital's location strongly influences their decision on where to admit patients. Geographic proximity was found to be a strong predictor of whether or not a physician utilizes a hospital.

In classic spatial models of competition each firm chooses a location such that it attracts the profit maximizing amount of consumers. In markets for services such as HHC or home repair the site of exchange is the consumer's home and although proximity to consumers remains the source of market power, it is the firms who engages in travel. Under fixed prospective payments, the firm bears the costs of travel. When firms choose a price schedule, discriminatory pricing occurs if the firm bears the transportation cost. More importantly, the notion of a marginal consumer (a consumer that is indifferent between traveling to the closest firm to her right and the closest firm to her left), is different than the one in Salop (1979). Here the marginal consumer is the one that makes the firm indifferent between serving her or not, and as such does not directly defines the boundaries of the demand for the firm (unless the firm is a local monopoly). Therefore, multiple firms may compete for the same consumers in equilibrium.

Because provision of care takes place in patients' homes, service delivery in this industry is both labor-intensive and decentralized. These two features have a potentially important effect on the nature of competition in HHC markets. The fact that there are few capital requirements lowers the barriers to entry. In the next section the authors discuss the fact that states have imposed an artificial barrier to competition by restricting the creation of new HHAs through Certificate of Need (CON) regulation. The decentralized nature of service delivery has two important effects: First, because patients are 'matched' to a home aide, nurse, or therapist by agencies, switching costs within and across agencies may be similar. Secondly, monitoring quality of care is difficult for both agencies and regulatory bodies.

Nature, Roles, and Impacts of Regulation

Entry Regulation through Certificate-of-Need Laws

Although states universally adopted CON for hospitals in the 1970s, 38 states also applied CON regulation to the HHC sector. When the federal mandate was repealed in 1987, only 18 states continued active CON regulations for HHC. Interestingly, the lessons from hospitals will not necessarily apply to home health. Unlike hospitals, SNFs, or physician offices, where location provides a degree of market power, HHAs deliver services at the patient residence. Without location as a natural barrier to competition, one might expect home health markets to be a highly competitive. Similarly, unlike hospitals and other facilities that require major capital investments in order to become operational, HHC is labor intensive and is expected to be highly competitive absent of entry regulation.

CON for hospitals, nursing homes, and rehabilitation centers were designed to give state governments the authority to restrict the construction of new and expansions to existing facilities, as well as the purchase of expensive technology. These restrictions were designed to prevent overutilization and duplication of services and ensure quality by centralizing medical services to high-volume facilities. Although acquisition or expansion of hospitals requires large capital investments, home health is a labor intensive industry with little capital investment and no evidence of a volume-outcome relationship. Therefore, there is no reason to expect an effect of CON on expenditures, costs, procedure volume, or mortality. Moreover, CON for home health, operates as a mechanism for restricting entry of new agencies. Most states with CON regulations follow specific policies and guidelines for the approval of additional HHAs in a given market, but in practice new agencies are rarely approved, leaving markets in CON-regulated states uncontested by potential entrants. CON laws serve as an artificial barrier on the number of competitors in a given market. Unlike in the case of hospitals, it is nearly impossible for a potential entrant to demonstrate 'need,' as incumbent agencies are not constrained by capacity and have no hurdles when it comes to expansion of services. Not surprisingly, CON states have almost half the number of Medicare-certified agencies compared to non-CON states although Medicare expenditures are similar in CON and non-CON states.

An alternative rationale for CON programs in home health is that they can improve quality of care through enhanced ability to monitor agencies. With fewer agencies, state regulators may be more effective at having a positive influence on the quality of care delivered by the HHAs in their state. However, although HHC in CON states was found to be less intensive (lower frequency of visits and lower skill mix), to date there is no evidence to suggest CON in HHC is quality enhancing. This may not be surprising, as the number of evidence-based standards of care in home health on which effective quality regulation can be based is limited.

Price Regulation

As discussed in Section Reimbursement under Medicare, the price of a 60-day home health episode is fixed and set at admission according to the severity of the patient's condition. Because prices are regulated, providers can no longer compete for patients based on price of services and instead compete for patients on the quality of their services. Economic theory suggests that market competition in the presence of regulated prices can drive up quality. Indeed, most empirical studies of the relationship between competition and quality under regulated prices in the case of dialysis centers and hospitals found more competition to result in higher quality (as measured by lower mortality).

Although the effect of market concentration on quality has been studied extensively in the hospital sector, this relationship has received little attention in the HHC industry. Some studies focused on the effect that Medicare PPS for home health services had on market concentration. One study has found that reimbursement cuts under IPS and PPS led to massive closure of HHAs, which found it difficult to remain fiscally viable. Moreover, states with higher barriers to entry through CON laws showed relatively lower rates of agency termination.

The Role of Integration

Vertical integration of acute care sites (i.e., hospitals) into postacute care (e.g., SNFs, rehabilitation centers, and HHAs) is

common and has the potential to influence the nature of health interventions. Vertical integration increased dramatically during the 1990s, with three-quarters of hospitals integrated with postacute care in 2001. Although patient care is produced along a care continuum, which includes both acute and postacute care entities, reimbursement for entities along the same continuum does not incorporate the fact that patient outcomes depend on the entire patient experience, including the transition between facilities. Vertical integration has the potential to correct such distortions, and is a key feature of the Accountable Care Organization concept.

Environmental changes in health care in the form of PPS's, managed care, and aging of the population have resulted in greater interdependence among acute and postacute providers. Although postacute care has been described as highly fragmented and with much redundancy, the increase in the level of interdependence among contracting parties increases the costs of external market exchange and favors integration. From an efficiency perspective, vertical integration in the health care sector can reduce transaction costs, and raise quality of care due to greater coordination and continuity of care.

Another study looked at vertical integration of hospitals and SNFs before and after the introduction of PPS for hospitals. PPS produced strong incentives to reduce costs per admission by shortening the average length of patient stays, which in turn created a new dependency of hospitals on nursing homes. The price paid to the nursing home to accept a hospital patient is established unilaterally by Medicare and therefore cannot be negotiated between the hospital and the nursing home. Hence, vertical integration becomes the only feasible route to affect the implicit transfer prices governing patient flows between the hospital and its own nursing home division. Hospitals with larger fractions of their patients covered by Medicare were significantly more likely to integrate vertically into nursing home services than were hospitals with proportionately fewer Medicare patients. A similar argument was put forth in another study, which concludes that financial pressure was the key driver leading to vertical integration of hospitals and HHAs in the mid-80s. As environmental pressures increase, hospitals benefited from tighter linkages with home health providers. Furthermore, an even earlier study compared the medical process at two hospitals, one with and one without a home nursing department. Regression analysis showed that home nursing care significantly reduced both the length of hospital stays and the number of follow-up visits to outpatient clinics. After accounting for the cost of the home nursing program, however, the program did not significantly reduce overall hospital expenditures.

Consistent with these findings, in a recent paper the authors introduce a theoretical framework, in which vertical integration allows hospitals to shift patient recovery tasks downstream to lower cost delivery entities (e.g., SNFs or HHAs) by discharging patients earlier. Because integrated hospitals fully control the postacute tier, they can ensure that patients discharged earlier and in poorer health receive greater posthospitalization service intensity. Although integration facilitates a change in the timing of hospital discharge, health outcomes are no worse when patients receive care from an integrated provider. It is shown that vertically integrated hospitals tend to discharge patients to their own HHAs sooner, with poorer health at the time of transition out of the hospital, yet with similar overall health outcomes. The authors used rehospitalization rate within 60 days of hospital discharge as the outcome variable. According to a recent report to Congress, "Hospital readmissions are sometimes indicators of poor care or missed opportunities to better coordinate care. Research shows that specific hospital-based initiatives to improve communication with beneficiaries and their other caregivers, coordinate care after discharge, and improve the quality of care during the initial admission can avert many readmissions." The Hospital Readmissions Reduction Program is a new Medicare program that establishes a financial incentive for hospitals to lower readmission rates. Under the program, Medicare's base operating diagnosis-related group payment amounts will be reduced for hospitals with excess readmissions.

The Use of Technology in Home Health Care

Telemedicine is a term used to cover a broad category of services, defined by the Institute of Medicine as "the use of electronic information and communications technologies to provide and support health care when distance separates the participants." The term is also applied more narrowly to medical care that uses interactive video, generally for consultations with specialists. However, telemedicine (or more generally, telehealth) is also comprised of the transmission of still images, e-health including patient portals, remote monitoring, medical education, and nursing call centers.

In the 1960s, the first uses of electronic telemedicine were to support neurologic and psychiatric services in Nebraska. With the exception of teleradiology, its adoption by physicians since then has been slow. Some of the main difficulties are licensing providers across state lines, liability concerns, reimbursement concerns, and physician awareness. From the 1960s through the 1990s, telemedicine consisted mostly of specialty consultations though videoconference technology. The millennium, however, saw more attention focused on noninteractive data storage and transmission. The thawing of Medicare's and other insurers' collective reluctance to cover telemedicine helped contribute to the 2000s' expansion. Both interactive and noninteractive technologies are increasingly used for remote monitoring of health status in homes.

Remote patient monitoring (RPM), or 'home telehealth,' is a subset of telemedicine that includes technology in a patient's home that records biometric data and transmits it to a central monitoring facility for interpretation. Consequently, patients can receive monitoring that might otherwise require physical nurse visits or trips to outpatient or inpatient facilities. Currently, Medicare spending on telemedicine is tracked as a whole, but not by class. Teleradiology has the largest expenditures, but the total amount is not documented, nor is it for RPM. Medicare reimburses for remote cardiac monitoring technologies and remote screening. Videoconference technology for rural patients has seen rapid growth, but it is still underutilized with less than US$1 million in expected reimbursements for 2011. Home telemedicine (and delivery for it) is paid for under the prospective payment reimbursement system.

An early and successful application of RPM was in heart monitoring, which culminated in greater safety for at-risk, rural-dwelling patients. RPM has rendered home health more likely to be substitutable for medical treatment in a more intensive location. By lifting the burden of face-to-face contact between providers and patients, telemedicine in theory should be access expanding, cost-effective, and quality improving. There is evidence that access has improved as technology enabled rural patients now receive care that was once too costly and impractical to provide, but there are no well-controlled studies that demonstrate cost-effectiveness or quality improvement with these technologies.

RPM is characterized by large, up-front costs to acquire the capital, the need for highly trained labor to operate it, and the integration of care with response teams and specialists. An important potential limitation of delivering RPM in a cost-effective system is related to the way care is reimbursed in Medicare. If home health providers cannot recover the added capital expense of RPM, they may underinvest. But if home health providers are reimbursed for RPM at a higher rate there may not be sufficient controls to only use this technology in those patients who would gain the most from it. The challenge is for hospitals to work more effectively with providers and technology developers. When determining reimbursement for RPM, it is important to consider the true costs of the alternative form of care and to align incentives such that those making decisions about the course of treatment are not penalized for selecting treatment patterns that may save the system money. The ideas behind Accountable Care Organizations where savings to Medicare are shared among providers may create the environment for a more cost-effective use of telehealth.

The Nature, Role, and Impact of Quality Initiatives in Home Health Care

Quality in health care is significant because it greatly impacts an individual's well-being and is more influential on well-being than quality of most other goods and services. Prompted by consumers, providers, and the growing body of evidence about the poor quality of health care, policymakers developed a strong interest in designing and implementing system-wide, market-based reforms to promote quality in health care. CMS implemented quality reporting in HHC in 2003 and has a demonstration project testing pay for performance in home health.

Public Reporting in Home Health: Home Health Compare

The public reporting initiative in home health started in October of 2003 when CMS launched a website called Home Health Compare. This website posts quality performance information for HHAs that serve a particular zip code. The quality measures generally measure how well the patients of an HHA regain or maintain their ability to function. There are 10 quality measures posted on HHC which come from a subset of larger set of 41 OASIS outcome measures that are well known to the HHAs, including improvements in ambulation, bathing, transferring, management of oral medication, pain interfering with activity, dyspnea (shortness of breath), and urinary incontinence, as well as measures of acute care hospitalization, emergent care, and discharge to community. The emphasis of this initiative was to give consumers information regarding the quality of care provided by HHAs. Other similar initiatives, such as Hospital and Nursing Home Quality Initiatives, suggest that measured quality improves in response with these two initiatives.

In HHC, there are two pathways for which quality to be improved. The first is 'selection,' which is that knowledge about performance leads patients, their payers, and agents engaged in referrals to be more likely to select higher quality providers. This will raise average quality in a market because a greater share of patients receives care from high performers. The second pathway is 'change' which is that more information in the hands of stakeholders creates motivation for organizations and their providers to improve quality and that more feedback about performance within an organization can also lead to positive change.

There is limited research on the environments in which HHC will be most effective. Although competition's effect on quality has been studied extensively in the hospital sector, it has not yet been in home health. HHC should be studied separately because with services delivered in the home rather than in the facility of the provider – the nature of competition is different. Theoretically, patients in more competitive markets will have higher quality based on conduct measures (visits per admission) and performance measures (improved functional outcomes and fewer adverse events). Furthermore, HHC should result in quality improvement, and competitive markets should have greater quality improvement in outcomes. The only evidence currently available comes from the initial demonstration project for HHC which did show some improvement in quality.

Pay-for-Performance

In 2003, MedPAC recommended that Medicare reward providers who provide 'high-quality care or improve the quality of care for their patients.' Pay-for-performance ties a direct financial payment to performance on selected quality measures and creates incentives for individual providers to improve the quality of care. The program aims to reward quality where it is possible to measure. In home health, the measures based on currently mandatory patient evaluations met the proposed criteria. MedPAC seeks to make sure that measure sets are not fixed and that they progress to integrate new measures and to eliminate any obsolete or ineffective measures.

Readmissions reduction payments have been considered as well. With respect to lowering readmissions, hospitals are the most obvious focus, but in a 2007 report, MedPAC focused on aligning incentives across all with influence on outcomes. However, there is disagreement over the best way to reward reductions in hospital readmissions. It can be done by directly penalizing or rewarding hospitals or secondary means of reduction, such as RPM and HHC improvements.

In 2007, a P4P pilot was implemented in seven states been 2008 and 2009. The 'incentive pool' used to fund the program

was generated from savings due to less utilization of costly Medicare services. The payout structure was setup such that 75% of the pool went to agencies in the top 20% of the highest level of patient care and 25% of the pool went to the top 20% of those making the biggest improvements in patient care. If there were no savings, there would be no compensation. Results: for 2008, aggregate Medicare savings were US$15.4 million for three of four regions, with the Midwest region not achieving any savings. The demonstration is still under evaluation.

Acknowledgements

The authors are grateful to Richard Chesney, Bruce Kinosian, and Rachel Werner for their feedback during the development of the ideas in this article. Special Thanks to Robert Sanders who provided exceptional skilled research assistance to the writing of this article. This work is supported by NIH/NHLBI grant #R01 HL088586-01.

See also: Market for Professional Nurses in the US

Further Reading

Anderson, K. B. and Kass D. I. (1986). Certiciate of need regulation of entry into home health care, a multi-product cost function analysis, an economic policy analysis. *Bureau of Economics Staff Report to the Federal Trade Commission*. Washington, DC: The Federal Trade Commission.

Avalere Health, LLC (2009). *Medicare spending and rehospitalization for chronically ill medicare beneficiaries: Home health use compared to other post-acute care settings*. Washington, DC: Avalere Health LLC.

Banks, D., Parker, E. and Wendel, J. (2001). Strategic interaction among hospitals and nursing facilities: The efficiency effects of payment systems and vertical integration. *Health Economics* **10**(2), 119–134, Article first published online: March 2001. doi:10.1002/hec.585.

Choi, S. and Joan, D. (2009). Changes in the medicare home health care market, the impact of reimbursement policy. *Medical Care* **47**(3), 302–309.

Dansky, K. H., Milliron, M. and Gramm, L. (1996). Understanding hospital referrals to home health agencies. *Hospital & Health Administration* **41**(3), 331–342.

David, G., Rawley, E. and Polsky, D. (2013). Integration and task allocation: Evidence from patient care. *Journal of Economics and Management Strategy* **22**(3), 617–639.

Dranove, D. (1985). An empirical study of a hospital-based home nursing care program. *Inquiry* **22**(1), 59–66.

Field, M. J. and Grigsby, J. (2002). Telemedicine and remote patient monitoring. *Journal of the American Medical Association* **288**(4), 423–425. doi: 10.1001/jama.288.4.423.

Goldsmith, J. (2004). Technology and the boundaries of the hospital: Three emerging technologies. *Health Affairs* **23**(6), 149–156.

Kenney, G. and Dubay, L. (1992). Explaining area variation in the use of medicare home health services. *Medical Care* **30**(1), 43–57.

Martin, A., Lassman, D., Whittle, L., Catlin, A. and National Health Expenditure Accounts Team (2011). Recession contributes to slowest annual rate of increase in health spending in five decades. *Health Affairs (Millwood)* **30**(l), 11–22.

Medicare Payment Advisory Commission (MedPAC) (2006). Adding quality measures in home health. *Report to the Congress: Medicare Payment Policy, Home Health Services, Section 4b 103–113*. June 2006.

Medicare Payment Advisory Commission (MedPAC) (2009). Home health services: Section 2E. *Report to the Congress: Medicare Payment Policy, Home Health Services, Section E 193–203*. Washington, DC: MedPac.

Polsky, D., David G., Yang, J., Kinosian, B. and Werner, R. (in press). The effect of entry regulation: The case of home health. *Journal of Public Economics*.

The National Association for Home Care & Hospice (NAHC) (2010). *Basic statistics about home care*. Available at: http://www.nahc.org/facts/10HC_Stats.pdf (accessed 30.05.11).

ENCYCLOPEDIA OF
HEALTH ECONOMICS

ENCYCLOPEDIA OF
HEALTH ECONOMICS

EDITOR-IN-CHIEF
Anthony J Culyer
University of Toronto, Toronto, Canada
University of York, Heslington, York, UK

ELSEVIER

Amsterdam • Boston • Heildelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

First edition 2014

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the Library of Congress.

ISBN 978-0-12-375678-7

Printed and bound in the United States of America

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

Project Manager: Gemma Taft
Associate Project Manager: Joanne Williams

EDITORIAL BOARD

CONTRIBUTORS TO VOLUME 2

J Abraham
University of Minnesota, Minneapolis, MN, USA

M Asaria
University of York, York, UK

DI Auerbach
RAND, Boston, MA, USA

MC Auld
University of Victoria, Victoria, BC, Canada

KS Babiarz
Stanford University, Stanford, CA, USA

M Baiocchi
Stanford University, Stanford, CA, USA

BH Baltagi
Syracuse University, Syracuse, NY, USA

E Bariola
Monash University, Clayton, VIC, Australia

H Bergquist
University of Pennsylvania, Philadelphia, PA, USA

A Bhattacharjee
Indian Institute of Management Bangalore, Karnataka, India

M Bitler
University of California Irvine, Irvine, CA, USA

C Blouin
Institut national de santé publique du Québec, Québec, Canada

J Brazier
School of Health and Related Research, University of Sheffield, Sheffield, UK

PI Buerhaus
Vanderbilt University Medical Center, Nashville, TN, USA

SH Busch
Yale School of Public Health, New Haven, CT, USA

AC Cameron
University of California – Davis, Davis, CA, USA

R Chanda
Indian Institute of Management Bangalore, Karnataka, India

JB Christianson
University of Minnesota School of Public Health, Minneapolis, MN, USA

P Clarke
The University of Melbourne, VIC, Australia

K Claxton
University of York, York, North Yorkshire, UK

J Connell
University of Sydney, NSW, Australia

R Cookson
University of York, York, UK

G David
University of Pennsylvania, Philadelphia, PA, USA

B Dowd
University of Minnesota, Minneapolis, MN, USA

A Edwards
University of Toronto, Toronto, ON, Canada

RS Eisenberg
University of Michigan Law School, Ann Arbor, MI, USA

G Erreygers
University of Antwerp, Antwerpen, Belgium

S Felder
Universität Basel, Switzerland

JM Fletcher
Yale School of Public Health, New Haven, CT, USA

LP Garrison
University of Washington, Seattle, WA, USA

U-G Gerdtham
Lund University, Lund, Sweden

M Gersovitz
Johns Hopkins University, Baltimore, MD, USA

TE Getzen
International Health Economics Association, Philadelphia, PA, USA

E Golberstein
University of Minnesota School of Public Health, Minneapolis, MN, USA

C Goulão
Toulouse School of Economics (GREMAQ, INRA), Toulouse, France

DC Grabowski
Harvard Medical School, Boston, MA, USA

H Grabowski
Duke University, Durham, NC, USA

WH Greene
New York University, New York, NY, USA

BA Griffin
RAND Corporation, Arlington, VA, USA

S Griffin
University of York, York, UK

PV Grootendorst
University of Toronto, Toronto, ON, Canada

V Ho
Rice University, Houston, TX, USA

A Hollis
University of Calgary, Calgary, AB, Canada

D Horsfall
University of York, Heslington, York, UK

V Iemmi
London School of Economics and Political Science, London, UK

T Iizuka
University of Tokyo, Tokyo, Japan

P Karaca-Mandic
University of Minnesota, Minneapolis, MN, USA

MR Keogh-Brown
London School of Hygiene and Tropical Medicine, London, UK

DP Kessler
Stanford University, Stanford, CA, USA

M Kifmann
Universität Hamburg, Hamburg, Germany

G Kjellsson
Lund University, Lund, Sweden

B van der Klaauw
VU University, Amsterdam, The Netherlands

MM Kleiner
University of Minnesota and NBER, Minneapolis, MN, USA

SA Kleiner
Cornell University, Ithaca, NY, USA

M Knapp
London School of Economics and Political Science, London, UK, and King's College London, Institute of Psychiatry, London, UK

RT Konetzka
University of Chicago, Chicago, IL, USA

PFM Krabbe
University of Groningen, Groningen, The Netherlands

M Kyle
Toulouse School of Economics, Toulouse, France, and Center for Economic Policy Research, Toulouse, France

SF Lehrer
Queen's University, Kingston, ON, Canada

M Lindeboom
VU University, Amsterdam, The Netherlands

N Lunt
University of York, Heslington, York, UK

WG Manning
University of Chicago, Chicago, IL, USA

M Martínez Álvarez
London School of Hygiene and Tropical Medicine, London, UK

JD Matsudaira
Cornell University, Ithaca, NY, USA

M Mazzocchi
Università di Bologna, Bologna, Italy

DF McCaffrey
ETS, Princeton, NJ, USA

J McKie
Monash University, Clayton, VIC, Australia

G Miller
Stanford University, Stanford, CA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

S Morris
University College London, London, UK

BH Neelon
Duke University, Durham, NC, USA

S Nicholson
Cornell University, Ithaca, NY, USA

E Nord
Norwegian Institute of Public Health and the University of Oslo, Norway

AJ O'Malley
Harvard Medical School, Boston, MA, USA

P Olivella
Universitat Autonoma de Barcelona and Barcelona GSE, Cerdanyola del Valles (Barcelona), Spain

A Oliver
London School of Economics and Political Science, London, UK

JC van Ours
Tilburg University, Tilburg, The Netherlands, and University of Melbourne, Melbourne, VIC, Australia

J Perelman
Universidade Nova de Lisboa (UNL), Lisbon, Portugal

P Pita Barros
Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal

RJ Pitman
Oxford Outcomes Ltd, Oxford, UK

D Polsky
University of Pennsylvania, Philadelphia, PA, USA

JS Preisser
University of North Carolina, Chapel Hill, NC, USA

PJ Rathouz
University of Wisconsin School of Medicine & Public Health, Madison, WI, USA

JB Rebitzer
Boston University, Boston, MA, USA; National Bureau of Economic Research, Cambridge, MA, USA; Case Western Reserve School of Medicine, Cleveland, OH, USA; Center for the Institute of the Study of Labor (IZA), Bonn, Germany, and The Levy Institute, Hudson, NY, USA

T Rice
University of California, Los Angeles, Los Angeles, CA, USA

J Richardson
Monash University, Clayton, VIC, Australia

JN Rosenquist
Harvard Medical School, Boston, MA, USA

D Rowen
School of Health and Related Research, University of Sheffield, Sheffield, UK

H Royer
University of California-Santa Barbara, Santa Barbara, CA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

CJ Ruhm
University of Virginia, Charlottesville, VA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

DE Sahn
Cornell University, Ithaca, NY, USA

A Schmid
Universität Bayreuth, Germany

P Serneels
University of East Anglia, Norwich, Norfolk, UK

B Shankar
Leverhulme Centre for Integrative Research on Agriculture and Health, London, UK, and University of London, London, UK

J Shen
Newcastle University, Newcastle Upon Tyne, UK

JV Terza
Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

JR Thomas
Georgetown University Law Center, Washington, DC, USA

A Towse
Office of Health Economics, London, UK

WB Traill
University of Reading, Reading, UK

PK Trivedi
Indiana University, Bloomington, IN, USA

V Ulrich
Universität Bayreuth, Germany

L Vallejo-Torres
University College London, London, UK

T Van Ourti
Erasmus University Rotterdam, Rotterdam, The Netherlands, and Tinbergen Institute Rotterdam, Rotterdam, The Netherlands

ME Votruba
Case Western Reserve School of Medicine, Cleveland, OH, USA

P Wilde
Tufts University, Boston, MA, USA

J Wildman
Newcastle University, Newcastle Upon Tyne, UK

J Williams
University of Melbourne, Melbourne, VIC, Australia

A Witman
University of California-Santa Barbara, Santa Barbara, CA, USA

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you're interested in: an alphabetical contents list, cross-references to other relevant articles within each article, and a full subject index.

1 Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

2 Cross-References

Most of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an entry as a See also list, serve four different functions:

i. To draw the reader's attention to related material in other entries.
ii. To indicate material that broadens and extends the scope of the article.
iii. To indicate material that covers a topic in more depth.
iv. To direct readers to other articles by the same author(s).

Example

The following list of cross-references appears at the end of the entry Abortion.

> *See also*: Education and Health in Developing Economies. Fertility and Population in Developing Countries. Global Public Goods and Health. Infectious Disease Externalities. Nutrition, Health, and Economic Performance. Water Supply and Sanitation

3 Index

The index includes page numbers for quick reference to the information you're looking for. The index entries differentiate between references to a whole entry, a part of an entry, and a table or figure.

4 Contributors

At the start of each volume there is list of the authors who contributed to that volume.

SUBJECT CLASSIFICATION

Demand for Health and Health Care

Collective Purchasing of Health Care
Demand Cross Elasticities and 'Offset Effects'
Demand for Insurance That Nudges Demand
Education and Health: Disentangling Causal Relationships from Associations
Health Care Demand, Empirical Determinants of
Medical Decision Making and Demand
Peer Effects, Social Networks, and Healthcare Demand
Physician-Induced Demand
Physician Management of Demand at the Point of Care
Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment
Quality Reporting and Demand
Rationing of Demand

Determinants of Health and Ill-Health

Abortion
Addiction
Advertising as a Determinant of Health in the USA
Aging: Health at Advanced Ages
Alcohol
Education and Health
Illegal Drug Use, Health Effects of
Intergenerational Effects on Health – *In Utero* and Early Life
Macroeconomy and Health
Mental Health, Determinants of
Nutrition, Economics of
Peer Effects in Health Behaviors
Pollution and Health
Sex Work and Risky Sex in Developing Countries
Smoking, Economics of

Economic Evaluation

Adoption of New Technologies, Using Economic Evaluation
Analysing Heterogeneity to Support Decision Making
Budget-Impact Analysis
Cost-Effectiveness Modeling Using Health State Utility Values
Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties
Economic Evaluation, Uncertainty in
Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis
Infectious Disease Modeling
Information Analysis, Value of
Observational Studies in Economic Evaluation
Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes
Problem Structuring for Health Economic Model Development
Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation
Searching and Reviewing Nonclinical Evidence for Economic Evaluation
Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies
Statistical Issues in Economic Evaluations
Synthesizing Clinical Evidence for Economic Evaluation
Value of Information Methods to Prioritize Research
Valuing Informal Care for Economic Evaluation

Efficiency and Equity

Efficiency and Equity in Health: Philosophical Considerations
Efficiency in Health Care, Concepts of
Equality of Opportunity in Health
Evaluating Efficiency of a Health Care System in the Developed World
Health and Health Care, Need for
Impact of Income Inequality on Health
Measuring Equality and Equity in Health and Health Care
Measuring Health Inequalities Using the Concentration Index Approach
Measuring Vertical Inequity in the Delivery of Healthcare
Resource Allocation Funding Formulae, Efficiency of
Theory of System Level Efficiency in Health Care
Welfarism and Extra-Welfarism

Global Health

Education and Health in Developing Economies
Fertility and Population in Developing Countries

Health and Its Value

Health and the Macroeconomy

Health Econometrics

Health Insurance

Human Resources

Markets in Health Care

Pharmaceutical and Medical Equipment Industries

Public Health

Supply of Health Services

Ambulance and Patient Transport Services
Cost Function Estimates
Healthcare Safety Net in the US
Home Health Services, Economics of
Long-Term Care
Production Functions for Medical Services
Understanding Medical Tourism

PREFACE

What Do Health Economists Do?

This encyclopedia gives the reader ample opportunity to read about what it is that health economists do and the ways in which they set about doing it. One may suppose that health economics consist of no more than the application of the discipline of economics (that is, economic theory and economic ways of doing empirical work) to the two topics of health and healthcare. However, although that would usefully uncouple 'economics' from an exclusive association with 'the (monetized) economy,' markets, and prices, it would miss out a great deal of what it is that health economists actually do, irrespective of whether they are being descriptive, theoretical, or applied. One distinctive characteristic of health economics is the way in which there has been a process of absorption into it (and, undoubtedly, from it too); in particular, the absorption of ideas and ways of working from biostatistics, clinical subjects, cognitive psychology, decision theory, demography, epidemiology, ethics, political science, public administration, and other disciplines already associated with 'health services research' (HSR) and, although more narrowly, 'health technology assessment' (HTA). But to identify health economics with HSR or HTA would also miss much else that health economists do.

… And How Do They Do It?

As for the ways in which they do it, in practice, the overwhelming majority of health economists use the familiar theoretical tools of neoclassical economics, although by no means all (possibly not even a majority) are committed to the welfarist (specifically the Paretian) approach usually adopted by mainstream economists when addressing normative issues, which actually turns out to have been a territory in which some of the most innovative ideas of health economics have been generated. Health economists are also more guarded than most other economists in their use of the postulates of soi-disant 'rationality' and in their beliefs about what unregulated markets can achieve. To study healthcare markets is emphatically not, of course, necessarily to advocate their use.

A Schematic of Health Economics

To think of health economics merely in these various restricted ways would be indeed to miss a great deal. The broader span of subject matter may be seen from the plumbing diagram, in which I have attempted to illustrate the entire range of topics in health economics. A version of the current schematic first appeared in Williams (1997, p. 46). The content of the encyclopedia follows, broadly, this same structure. The arrows in the diagram indicate a natural logical and empirical order, beginning with **Box A** (Health and its value) (**Figure 1**).

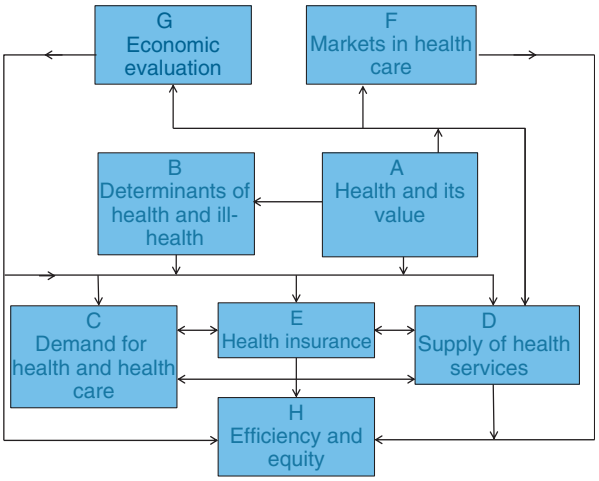


Figure 1 A schematic of health economics.

Box A, in the center-right of the schematic, contains fundamental concepts and measures of population health and health outcomes, along with the normative methods of welfarism and extra-welfarism; measures of utility and health outcomes, including their uses and limitations; and methods of health outcome valuation, such as willingness to pay and experimental methods for revealing such values, and their uses and limitations. It includes macro health economic topics like the global burden of disease, international trade, public and private healthcare expenditures, Gross Domestic Product (GDP) and healthcare expenditure, technological change, and economic growth. Some of the material here is common to epidemiology and bioethics.

Box A Health and its value

Concepts and measures of population health and health outcomes.
Ethical approaches (e.g., welfarism and extrawelfarism).
Measures of utility and the principal health outcome measures, their uses, and limitations.
Health outcome valuation methods, willingness to pay, their uses, and limitations.
Macro health economics: global burdens of disease, international trade, healthcare expenditures, GDP, technological change, and economic growth.

Box B (Determinants of health and ill health) builds on these basics in various 'big-picture' topics, such as the population health perspective for analysis and the determinants of lifetime health, such as genetics, early parenting, and schooling; it embraces occupational health and safety, addiction (especially tobacco, alcohol, and drugs), inequality as a determinant of ill health, poverty and the global burden of disease in low- and middle-income countries, epidemics, prevention, and public health technologies. Here too, much is

Box B Determinants of health and ill health

The population health perspective.
Early determinants of lifetime health (e.g., genetics, parenting, and schooling).
Occupational health and safety.
Addiction: tobacco, alcohol, and drugs.
Inequality as a determinant of ill health.
Poverty and global health (in LMICs).
Epidemics.
Prevention.
Public health technologies.

shared, both empirically and conceptually, with other disciplines.

From this it is a relatively short step into **Box C** (Demand for health and healthcare): here we are concerned with the difference between demand and need; the demand for health as 'human capital'; the demand for healthcare (as compared with health) and its mediation by 'agents' like doctors on behalf of 'principals'; income and price elasticities; information asymmetries (as in the different types of knowledge and understandings by patients and healthcare professionals, respectively) and agency relationships (when one, such as a health professional, acts on behalf of another, such as a patient); externalities or spillovers (when one person's health or behavior directly affects that of another) and publicness (the quality which means that goods or services provided for one are also necessarily provided for others, like proximity to a hospital); and supplier-induced demand (as when a professional recommends and supplies care driven by other interests than the patient's).

Box C Demand for health and healthcare

Demand and need.
The demand for health as human capital.
The demand for healthcare.
Agency relationships in healthcare.
Income and price elasticities.
Information asymmetries and agency relationships.
Externalities and publicness.
Supplier-induced demand.

Then comes **Box D** (Supply of healthcare) covering human resources; the remuneration and behavior of professionals; investment and training of professionals in healthcare; monopoly and competition in healthcare supply; for-profit and nonprofit models of healthcare institutions like hospitals and clinics; health production functions; healthcare cost and production functions that explore the links between 'what goes in' and 'what comes out;' economies of scale and scope; quality of care and service; and the safety of interventions and modes of delivery. It includes the estimation of cost functions and the economics of the pharmaceutical and medical equipment industries. A distinctive difference in this territory from many other areas of application is the need to drop the assumption

Box D Supply of health services

Human resources, remuneration, and the behavior of professionals.
Investment and training of professionals in healthcare.
Monopoly and competition in healthcare supply.
Models of healthcare institutions (for-profit and nonprofit).
Health production functions.
Healthcare cost and production functions.
Economies of scale and scope.
Quality and safety.
The pharmaceutical and medical equipment industries.

of profit-maximizing as a common approach to institutional behavior and to incorporate the idea of 'professionalism' when explaining or predicting the responses of healthcare professionals to changes in their environment.

Supply and demand are mediated (at least in the high-income world) by insurance: the major topic of **Box E** and a large part of health economics as practiced in the US. This covers the demand for insurance; the supply of insurance services and the motivations and regulations of insurance as an industry; moral hazard (the effect of insurance on utilization); adverse selection (the effect of insurance on who is insured); equity and health insurance; private and public systems of insurance; the welfare effects of soi-disant 'excess' insurance; effects of insurance on healthcare providers; and various specific issues in coverage, such as services to be covered in an insured bundle and individual eligibility to receive care. Although the health insurance industry occupies a smaller place in most countries outside the US, the issues invariably crop up in a different guise and require different regulatory and other responses.

Box E Health insurance

The demand for insurance.
The supply of insurance services.
Moral hazard.
Adverse selection.
Equity and health insurance.
Private and public systems.
Welfare effects of 'excess' insurance.
Effects of insurance on healthcare providers.
Issues in coverage: services covered and individual eligibility.
Coverage in LMICs.

Then, in **Box F**, comes a major area of applied health economics: markets in healthcare and the balance between private and public provision, the roles of regulation and subsidy, and the mostly highly politicized topics in health policy. This box includes information and how its absence or distortion corrupts markets; other forms of market failure due to externalities; monopolies and a catalog of practical difficulties both for the market and for more centrally planned systems; labor markets in healthcare (physicians, nurses, managers, and allied professions), internal markets (as when the public sector of healthcare is divided into agencies that commission care on behalf of populations and those that

Box F Markets in healthcare

Information and markets and market failure.
Labor markets in healthcare: physicians, nurses, managers, and allied professions.
Internal markets in the healthcare sector.
Rationing and prioritization.
Welfare economics and system evaluation.
Comparative systems.
Waiting times and lists.
Discrimination.
Public goods and externalities.
Regulation and subsidy.

provide it); rationing and the various forms it can take; welfare economics and system evaluation; waiting times and lists; and discrimination. It is here that many of the features that make healthcare 'different' from other goods and services become prominent.

Box G is about evaluation and healthcare investment, a field in which the applied literature is huge. It includes cost-benefit analysis, cost-utility analysis, cost-effectiveness analysis, and cost-consequences analysis; their application in rich and poor countries; the use of economics in medical decision making (such as the creation of clinical guidelines); discounting and interest rates; sensitivity analysis as a means of testing how dependent one's results are on assumptions; the use of evidence, efficacy, and effectiveness; HTA, study design, and decision process design in agencies with formulary-type decisions to make; the treatment of risk and uncertainty; modeling made necessary by the absence of data generated in trials; and systematic reviews and meta-analyses of existing literature. This territory has burgeoned especially, thanks to the rise of 'evidence-based' decision making and the demand from regulators for decision rules in determining the composition of insured bundles and the setting of pharmaceutical prices.

Box G Economic evaluation

Decision rules in healthcare investment.
Techniques of cost-benefit analysis in health and healthcare.
Techniques of cost-utility analysis and cost-effectiveness analysis in health and healthcare in rich and poor countries.
Techniques of cost-consequences analysis.
Decision theoretical approaches.
Outcome measures and their interpretation.
Discounting.
Sensitivity analysis.
Evidence, efficacy, and effectiveness.
Economics and health technology assessment.
Study design.
Risk and uncertainty.
Modeling.
Systematic reviews and meta-analyses.

The final Box, H, draws on all the preceding theoretical and empirical work: concepts of efficiency, equity, and possible conflicts between them; inequality and the socioeconomic 'gradient;' techniques for measuring equity and inequity; evaluating efficiency at the system level; evaluating equity at system level: financing arrangements; evaluating equity at system level: service access and delivery; institutional arrangements for efficiency and equity; policies against global poverty and for health; universality and comprehensiveness as global objectives of healthcare; and healthcare financing and delivery systems in low- and middle-income countries (LMICs). This is the most overtly 'political' and policy-oriented territory.

Box H Efficiency and equity

Concepts of efficiency, equity, and possible conflicts.
Inequality and the socioeconomic 'gradient.'
Evaluating efficiency: international comparisons.
Techniques for measuring equity and inequity.
Evaluating equity at system level: financing arrangements.
Evaluating equity at system level: service access and delivery.
Institutional arrangements for efficiency and equity.
Global poverty and health.
Universality and comprehensiveness.
Healthcare financing and delivery systems in LMICs.

A Word on Textbooks

The scope of a subject is often revealed by the contents of its textbooks. There are now many textbooks in health economics, having various degrees of sophistication, breadth of coverage, balance of description, theory and application, and political sympathies. They are not reviewed here but I have tried to make the (English language) list in the Further Reading as complete as possible. Because the assumptions that textbook writers make about the preexisting experience of readers and about their professional backgrounds vary, not every text listed here will suit every potential reader. Moreover, a few have the breadth of coverage indicated in the schematic here. Those interested in learning more about the subject to supplement what is to be gleaned from the pages of this encyclopedia are, therefore, urged to sample what is on offer before purchase.

Acknowledgments

My debts of gratitude are owed to many people. I must particularly thank Richard Berryman (Senior Project Manager), at Elsevier, who oversaw the inception of the project, and Gemma Taft (Project Manager) and Joanne Williams (Associate Project Manager), who gave me the most marvelous advice and support throughout. The editorial heavy lifting was done by Billy Jack and Karen Grépin (Global Health); Aki Tsuchiya and John Wildman (Efficiency and Equity); John Cawley and Kosali Simon (Determinants of Health and Ill health); Richard Cookson and Mark Suhrcke (Public Health); Erik Nord (Health and its Value); Richard Smith (Health and the

Macroeconomy); John Mullahy and Anirban Basu (Health Econometrics); Tom McGuire (Demand for Health and Healthcare); John Nyman (Health Insurance); Jim Burgess (Supply of Health Services); Martin Gaynor and Sean Nicholson (Human Resources); Patricia Danzon (Pharmaceutical and Medical Equipment Industries); Pau Olivella and Pedro Pita Barros (Markets in Healthcare); and John Brazier, Mark Sculpher, and Anirban Basu (Economic Evaluation). Finally, my thanks to the Advisory Board: Ron Akehurst, Andy Briggs, Martin Buxton, May Cheng, Mike Drummond, Tom Getzen, Jane Hall, Andrew Jones, Bengt Jonsson, Di McIntyre, David Madden, Jo Mauskopf, Alan Maynard, Anne Mills, the late Gavin Mooney, Jo Newhouse, Carol Propper, Ravindra Rannan-Eliya, Jeff Richardson, Lise Rochaix, Louise Russell, Peter Smith, Adrian Towse, Wynand Van de Ven, Bobbi Wolfe, and Peter Zweifel. Although the Board was not called on for frequent help, their strategic advice and willingness to be available when I needed them was a great comfort.

Anthony J Culyer
Universities of Toronto (Canada) and York (England)

Further Reading

Cullis, J. G. and West, P. A. (1979). *The economics of health: An introduction*. Oxford: Martin Robertson.

Donaldson, C., Gerard, K., Mitton, C., Jan, S. and Wiseman, V. (2005). *Economics of health care financing: The visible hand*. London: Palgrave Macmillan.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. and Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*, 3rd ed. oxford: Oxford University Press.

Evans, R. G. (1984). *Strained mercy: The economics of Canadian health care*. Markham, ON: Butterworths.

Feldstein, P. J. (2005). *Health care economics*, 6th ed. Florence, KY: Delmar Learning.

Folland, S., Goodman, A. C. and Stano, M. (2010). *The economics of health and health care*, 6th ed. Upper Saddle River: Prentice Hall.

Getzen, T. E. (2006). *Health economics: Fundamentals and flow of funds*, 3rd ed. Hoboken, NJ: Wiley.

Getzen, T. E. and Allen, B. H. (2007). *Health care economics*. Chichester: Wiley.

Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (eds.) (1996). *Cost-effectiveness in health and medicine*. New York and Oxford: Oxford University Press.

Henderson, J. W. (2004). *Health economics and policy with economic applications*, 3rd ed. Cincinnati: South-Western Publishers.

Hurley, J. E. (2010). *Health economics*. Toronto: McGraw-Hill Ryerson.

Jack, W. (1999). *Principles of health economics for developing countries*. Washington, DC: World Bank.

Jacobs, P. and Rapoport, J. (2004). *The economics of health and medical care*, 5th ed. Sudbury, MA: Jones & Bartlett.

Johnson-Lans, S. (2006). *A health economics primer*. Boston: Addison Wesley/ Pearson.

McGuire, A., Henderson, J. and Mooney, G. (1992). *The economics of health care*. Abingdon: Routledge.

McPake, B., Normand, C. and Smith, S. (2013). *Health economics: An international perspective*, 3rd ed. Abingdon: Routledge.

Mooney, G. H. (2003). *Economics, medicine, and health care*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

Morris, S., Devlin, N. and Parkin, D. (2007). *Economic analysis in health care*. Chichester: Wiley.

Palmer, G. and Ho, M. T. (2008). *Health economics: A critical and global analysis*. Basingstoke: Palgrave Macmillan.

Phelps, C. E. (2012). *Health economics*, 5th (international) ed. Boston: Pearson Education.

Phillips, C. J. (2005). *Health economics: An introduction for health professionals*. Chichester: Wiley (BMJ Books).

Rice, T. H. and Unruh, L. (2009). *The economics of health reconsidered*, 3rd ed. Chicago: Health Administration Press.

Santerre, R. and Neun, S. P. (2007). *Health economics: Theories, insights and industry*, 4th ed. Cincinnati: South-Western Publishing Company.

Sorkin, A. L. (1992). *Health economics – An introduction*. New York: Lexington Books.

Walley, T., Haycox, A. and Boland, A. (2004). *Pharmacoeconomics*. London: Elsevier.

Williams, A. (1997). Being reasonable about the economics of health: Selected essays by Alan Williams (edited by Culyer, A. J. and Maynard, A.). Cheltenham: Edward Elgar.

Witter, S. and Ensor, T. (eds.) (1997). *An introduction to health economics for eastern Europe and the Former Soviet Union*. Chichester: Wiley.

Witter, S., Ensor, T., Jowett, M. and Thompson, R. (2000). *Health economics for developing countries. A practical guide*. London: Macmillan Education.

Wonderling, D., Gruen, R. and Black, N. (2005). *Introduction to health economics*. Maidenhead: Open University Press.

Zweifel, P., Breyer, F. H. J. and Kifmann, M. (2009). *Health economics*, 2nd ed. Oxford: Oxford University Press.

CONTENTS OF ALL VOLUMES

VOLUME 1

VOLUME 2

VOLUME 3

Illegal Drug Use, Health Effects of

JC van Ours, Tilburg University, Tilburg, The Netherlands and University of Melbourne, Melbourne, VIC, Australia
J Williams, University of Melbourne, Melbourne, VIC, Australia

Introduction

The potential health risks associated with using illicit drugs remain the key argument for maintaining their criminal status. And although many studies find that drug users are in worse health than nonusers, the proper interpretation of this evidence is contentious. This is because, in order to conclude that it is in fact their drug use that causes them poor health, two alternative explanations for the association must be eliminated. This issue is not new. Determining the true nature of the relationship between drug use and health has a long history. An early example of a discussion of the issues can be found in the 1894 Indian Hemp Commission Report (Kendell, 2003). The first alternative explanation is referred to as reverse causality. Under reverse causality, the observed relationship between drug use and poor health runs in the reverse direction – from poor health to drug use. This may occur if, for example, people use illegal drugs to treat symptoms of their illness. The second alternative explanation is referred to as spurious correlation. This is an issue if there exists an unobserved factor, for example, childhood abuse, which causes both drug use and poor health. If this is the case, then the resulting correlation between drug use and poor health is spurious because drug use is simply capturing the unmeasured effect of the confounding factor, childhood abuse, on health. Untangling these competing and more than likely coexisting mechanisms generating the observed relationship between drug use and health is not merely of academic importance. The economic cost of maintaining criminal sanctions for illicit drug use is large. This cost is typically justified on the grounds that criminalizing drug use prevents health-related harms associated with drug use. For this reason, it is important to know whether and to what extent drug use causes ill health. This article reviews the evidence on this issue.

To begin, section The Extent of Illegal Drug Use introduces facts and figures regarding the extent of illicit drug use. To do so, the authors present data on the prevalence and intensity of use for the major illicit drugs: heroin, cocaine, amphetamines, ecstasy, and cannabis. These data illustrate the dominance of cannabis among illegal drugs. Although the prevalence of drug use provides an overview of the extent of drug use in a population, it is not necessarily informative about the type of drug use that may give rise to health-related problems. For example, the prevalence of use is unable to distinguish between those who have experimented once or twice (in the given time frame) and the more policy-relevant group who become long-term heavy users. Second, there is mounting evidence that uptake of drugs in the teenage years carries significantly more risks than uptake at later ages. Therefore, it is not simply the prevalence of use, but the age of first use and the duration of use that is informative in terms of risk of potential health-related harms. To provide information on these dimensions, the authors describe the dynamics of drug use. They do so for cannabis as this is by far the most popular illegal drug.

In Section Health Effects of Illegal Drug Use, the authors present and discuss a number of recent studies on the direct and indirect health effects of cannabis use. They distinguish between epidemiological and econometric studies. Section Discussion and Conclusion concludes that although consumers of illegal drugs are assumed to face substantial health risks, the evidence base regarding the nature and extent of these risks is, by and large, yet to be well established. For the most popular illegal drug, cannabis, there do not seem to be serious harmful effects with moderate use. There may be negative harmful effects for heavy users who are susceptible to mental health problems.

The Extent of Illegal Drug Use

Annual Prevalence of Illegal Drug Use

Table 1 provides information on the annual prevalence of use for the most important illegal drugs: amphetamines, ecstasy, cannabis, cocaine, and heroin. The annual prevalence refers to the percentage of the population aged 15–64 years who report any use of the substance in the year before being surveyed. The age range varies slightly for some countries. This information is reported for 10 developed countries and the authors refer the interested reader to United Nations (2011) for information on additional countries.

As shown in Table 1, with the exception of cannabis, the annual prevalence rate of use for any of these illegal drugs is not more than a few percentages of the population. The annual prevalence of amphetamine use ranges from a low of 0.2% of the population in France to a high of 2.7% of the population in Australia, and the annual prevalence of ecstasy use ranges from 0.1% of the population in Sweden to 4.2% in Australia. The range for the annual prevalence of cocaine use is similar, with a low of 0.5% of the population in Sweden to a high of 2.6% of the population in Spain. The annual prevalence rate of heroin use is low in all countries, ranging from 0.1% of the population in Spain to 0.8% of the population in England and Wales. For cannabis, the annual prevalence rate of use is substantially higher, ranging from 1.2% of the population in Sweden to a 14.6% of the population in Italy. The information in Table 1 makes it clear that cannabis is the most popular illegal drug by a wide margin. This is not an artifact of the countries that has been reported on. Globally, cannabis is the most commonly used illegal drug. In 2009, between 2.8 and 4.5% of the world's population aged 15–64 years, corresponding to between 125 and 203 million people, had used cannabis at least once in the past year (United Nations, 2011).

Table 1 Annual prevalence of illegal drugs; various countries (percentages)

Country	*Year*	*Age*	*Amph.*	*Ecstasy*	*Cannabis*	*Cocaine*	*Heroin*
Australia	2007	15–64	2.7	4.2	10.6	1.9	0.4
Denmark	2008	16–64	1.2	0.4	5.5	1.4	0.6
England	2010	16–59	1.0	1.6	6.6	2.5	0.8
France	2005	15–64	0.2	0.5	8.6	0.6	0.5
Germany	2009	18–64	0.7	0.4	4.8	0.9	0.2
Italy	2008	15–64	0.6	0.7	14.6	2.2	0.6
The Netherlands	2005	15–64	0.3	1.2	5.4	0.6	0.3
Spain	2010	15–64	0.6	0.8	10.6	2.6	0.1
Sweden	2008	15–64	0.8	0.1	1.2	0.5	0.2
United States	2009	15–64	1.5	1.4	13.7	2.4	0.6

Note: England includes Wales; Amph., amphetamines; heroin includes opium and except for the United States, it also includes other opioids such as morphine, methadone, etc. The information for heroin always refers to the population aged 15–64 years; the information is for the following years: Denmark and the Netherlands (2005), Germany and Italy (2008), United States (2009), all other countries (2007).

Source: Reproduced from United Nations (2011). *World Drugs Report 2011*. Vienna, Austria: United Nations Office on Drugs and Crime.

Table 2 Cannabis use; various countries (percentages)

Country	*Year*	*Population (age)*	*Ever use*	*Last year use*	*Last month use*
Australia	2007	≥14	34	9	5
Denmark	2008	16–64	39	6	2
England and Wales	2008–09	16–59	31	8	5
France	2005	15–64	31	9	5
Germany	2006	18–64	23	5	2
Italy	2008	15–64	32	14	7
The Netherlands	2009	15–64	26	7	4
Spain	2007–08	15–64	27	10	7
Sweden	2008	15–64	21	1	1
United States	2009	≥12	42	11	7

Source: Reproduced from van Laar M. (2011). Nationale Drug Monitor. Utrecht: Trimbos Instituut.

Intensity of Cannabis Use

Table 2 reports more detailed information on cannabis use for the same set of countries contained in Table 1. Specifically, Table 2 distinguishes between lifetime use, use in the last year, and use in the last month. There is substantial variation in these measures of use both across countries and within countries. The variation across countries is demonstrated by comparing Sweden, where just 21% of the population aged 15–64 years has used cannabis in their lifetime, with the US where 42% of those aged 12 years or older have used cannabis at some point in their lifetime. Similarly, just 1% of those aged 15–64 years in Sweden has used cannabis in the past year compared with 14% of those in Italy. Equally striking is the variation between lifetime and past year use within each country. In the Netherlands, for example, 26% of the population aged 15–64 years has used cannabis in their lifetime but only 7% have done so in the last year. Apparently, cannabis use is not very addictive for a substantial proportion of users (see van Ours, 2005 for details).

The proportion of the population who has used cannabis in the past month gives an indication of the extent of current use. However, as shown in Table 3, there remain substantial differences across countries in the frequency with which past month users consume cannabis. In Denmark, for example, almost 60% of past month users consumed cannabis no more than 1–3 days in the past 30 days whereas just 16% used at least 20 days out of the past 30. Even in Spain, where almost 9% of the population aged 15–64 years has used cannabis in the last 30 days, less than 3% of the population (or one-third of current users) has used cannabis on 20 or more days out of the last 30. In Germany, Italy, and the Netherlands, less than 1% of the population aged 15–64 years used cannabis on at least 20 days out of the past 30, and in France just 1.5% has done so. This demonstrates that, although cannabis is by far the most widely used among the illegal drugs, the prevalence of heavy use in the population is still low among the countries reported in Table 3.

Dynamics in Cannabis Use

Although a significant proportion of the population will have tried cannabis at some point in their life, many will simply experiment once or twice without suffering harmful consequences. To assess the degree of risk of harmful consequences, one needs to understand the profile of the duration of cannabis use. In addition, a growing literature provides evidence that early onset of cannabis use has especially harmful effects on health and life outcomes. Therefore, in this section, information on the dynamics of cannabis use, including age at first use and the duration of use, is provided.

Figure 1 shows typical patterns in the dynamics of cannabis use derived from a sample of Amsterdam residents (van Ours, 2005). Figure 1(a) provides information on the uptake

Table 3 Frequency of cannabis use in the past 30 days

Country	Year	Days in the past 30 days (%)				Total (%)	Last month prevalence (%)	
		1–3	4–9	10–19	20+		Total	20+ days
Denmark	2005	58	19	7	16	100	2.6	0.4
France	2005	36	17	15	32	100	4.8	1.5
Germany	2003	47	16	14	23	100	3.4	0.8
Italy	2005	47	25	10	18	100	5.8	1.0
The Netherlands	2005	38	12	27	23	100	3.3	0.8
Spain	2005/06	32	23	15	31	100	8.7	2.7

Note: Population aged 15–64 years.
Source: Reproduced from European Monitoring Center for Drugs and Drug Addiction.

of cannabis and Figure 1(b) provides information on quitting behavior as a function of the duration of cannabis use. The first graph in Figure 1(a) shows the hazard rate for starting cannabis use, defined as the probability of starting cannabis use at each age conditional on having not used up until that age. As can be seen from the graph, uptake typically occurs between the ages of 15 and 25 years, with clear spikes in the rate of uptake at ages 16, 18, and 20 years. The starting rate for ages greater than 25 years is small. This means that, if a person has not started cannabis use by the age of 25 years, they are unlikely to do so at a later age. The second graph in Figure 1(a) shows the cumulative starting probability. This is defined as the proportion of individuals at each age who have started cannabis use. The cumulative starting probability shows that 10% of 15-year olds have ever used cannabis. This proportion rises to 50% by the age of 25 years. The slowing in the rate of uptake after the age of 25 years is reflected in the flattening of the cumulative starting probability, which increases from 52% to 55% over the ages of 25–30 years.

Figure 1(b) shows the quit rate, defined as the probability of quitting cannabis use at each duration of use (measured in years) conditional on not previously quitting, and the cumulative quit probability, defined as the proportion of those who have ever used cannabis quitting at each duration of use. The graph of the quit rate shows that approximately 20% of cannabis users stop using within a year of starting use. The graph of the cumulative quit probability shows that although many cannabis users quit use after a couple of years, a significant proportion do not. For example, 20 years after first using cannabis, between 30% and 40% are still using. Based on these dynamics three groups of individuals can be distinguished; those who never use (abstainers), those who use but only for a short time (experimenters), and persistent users some of whom are recreational users whereas others are addicts. It is important to note that although these graphs were constructed using data on residents of Amsterdam, the patterns in Figure 1 are typical of the dynamics found in other countries.

In addition, the characteristics found in the dynamics of cannabis use are similar to those found for other illegal drugs, although the magnitude of use and the timing over the lifecycle may differ slightly from drug to drug. For example, for the sample of Amsterdam inhabitants on which Figure 1 is based, van Ours (2005) reports that the mean age of first use is 20 years for cannabis, 23 for amphetamines, 24 for heroin, 25 for cocaine, and 26 for ecstasy. In comparison, the mean age of first use for alcohol and tobacco is 17.5 years. He also finds drug-specific critical ages, such that if individuals have not started using by the critical age, then they are not very likely to do so at a later age. As seen above, the critical age is 25 years for cannabis. For cocaine it is 30 years, whereas for tobacco the critical age is approximately 20 years.

It is often found that the age of onset influences user quit rates. The earlier the individuals start using a particular drug, the less likely they are to stop using that drug. Although the general pattern in user dynamics is very much the same across the various drugs, there are also differences between drugs. Cannabis and cocaine use are characterized by relatively low starting rates that begin in the mid-teen years and by high quit rates especially in the first year after starting use. Tobacco use is characterized by high starting rates at a young age and by low quit rates. Once individuals start using cigarettes, they are very unlikely to stop using. Apparently, among the users of cannabis and cocaine there are many experimenters, that is, individuals who use the drug for a very short time but then decide very quickly to stop using.

From the dynamics in illegal drug use it is clear that there are differences between drugs in terms of the duration of use. These differences are related to the variation in the degree of psychic dependence of illegal drugs. As shown in column (1) of Table 4 the degree of psychic dependence is strongest for heroin, tobacco and alcohol, and weakest for cannabis. Nutt *et al.* (2010) present an attempt to score drugs according to 16 criteria of harm ranging from the intrinsic harms of the drug to social and health-care costs. Based on the criteria they distinguish between harm to users and harm to others. Drugs are scored on a scale of 0–100, with 100 assigned to the most harmful drug and 0 indicating no harm. The points were assigned in consultation with expert groups. The outcome is replicated in the second to fourth column of Table 4. In the second column harm to individual users is represented. The most harmful drugs to users are heroin and alcohol, whereas cannabis and ecstasy are least harmful to the users. The third column presents harm to others and here alcohol is the most harmful followed by heroin; ecstasy is the least harmful. The overall harm score is presented in the fourth column. Overall, alcohol is the most harmful drug and ecstasy the least harmful.

Of course such rankings of harm are not uncontroversial. Caulkins *et al.* (2011), for example, argue that the harmfulness of a drug cannot be indicated using one number as the harm is more than the harm to the user and spillover effects in terms of harm to others. Furthermore, harms related to drug-related

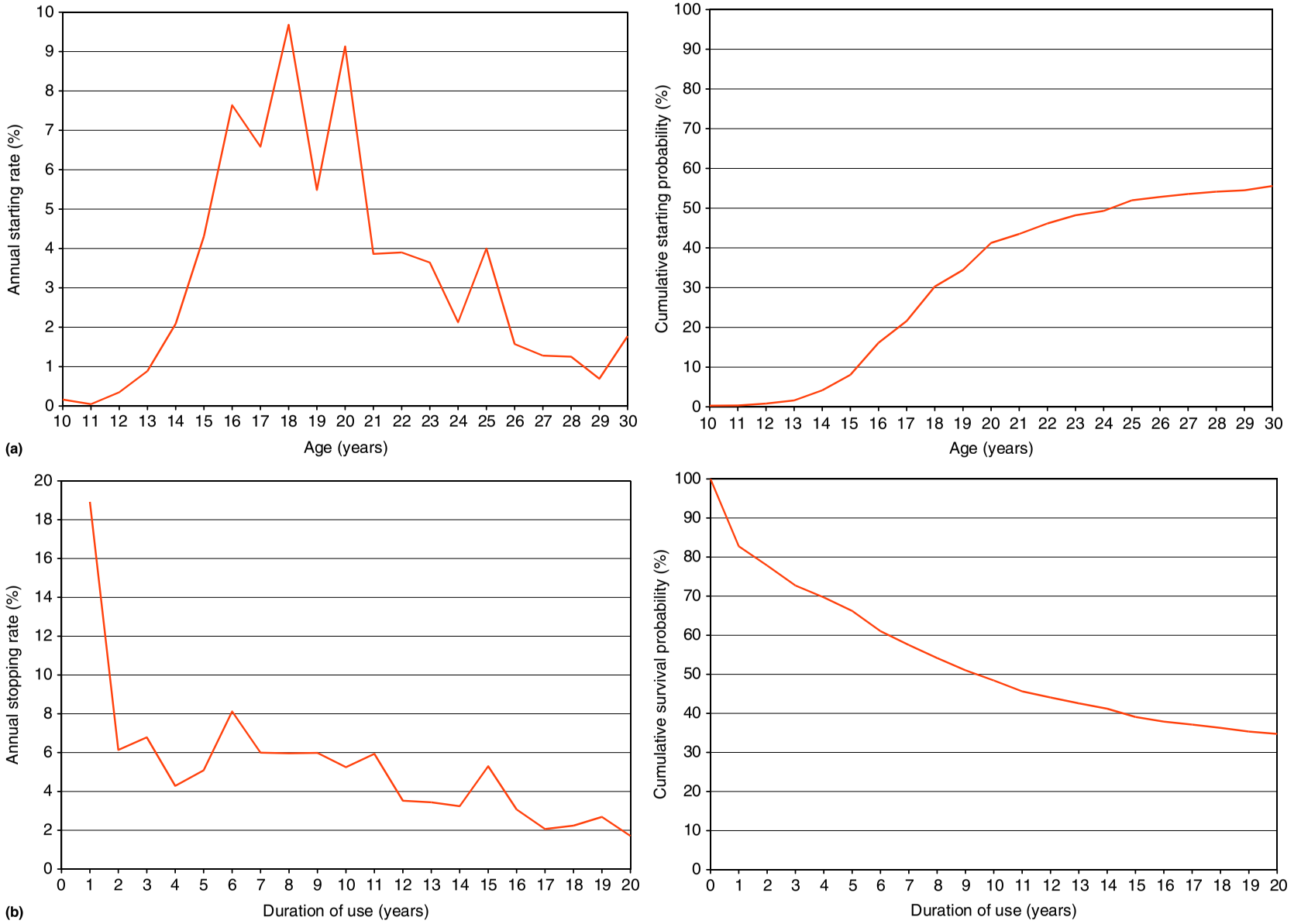


Figure 1 Dynamics in cannabis use in Amsterdam. (a) Starting rates (left) and cumulative starting probability (right) by age. (b) Quit rates (left) and cumulative quit probabilities (right) by duration of use in years. Reproduced from van Ours, J. C. (2005). Dynamics in the use of drugs, *Health Economics* **15**(12), 1283–1294 – Based on surveys in 1994, 1997, and 2001.

Table 4 Illegal drugs and legal drugs; addictiveness, degree of psychic dependence, and danger ratio

	Degree of psychic dependence	*Harm score*			*Danger ratio (%)*
	(1)	*Users (2)*	*Others (3)*	*Total (4)*	*(5)*
Illegal drugs					
Amphetamines	Middling	19	4	23	10.0
Ecstasy	–	8	1	9	6.3
Cannabis	Weak	12	8	20	<0.1
Cocaine	Strong, intermittent	19	8	27	6.7
Heroin	Very strong	34	21	55	16.7
Legal drugs					
Alcohol	Very strong	26	46	72	10.0
Tobacco	Very strong	19	9	26	<0.1

Note: Danger ratio = normal dose as a percentage of lethal dose.

Source: Column (1): Room, R., Fischer, B., Hall, W., Lenton, S. and Reuter P. (2010) *Cannabis policy: Moving beyond stalemate.* Oxford, England: Oxford University Press; columns (2–4): Nutt, D., King, L. and Phillips, L. (2010). Drug harms in the UK: A multi-criteria decision analysis. *Lancet* **376**, 1558–1565, column (5): Gable, R. (2004). Comparison of acute lethal toxicity of commonly abused psychoactive substances. *Addiction*, **99**, 686–696; except tobacco which is assumed to be similar to cannabis.

crime, environmental damage, and the cost of police and prisons depend on the legal status of the drug. Finally, the fifth column of Table 4 provides information about the acute lethal toxicity of illegal drugs, i.e., the 'danger ratio,' defined as the usual effective dose as a percentage of usual lethal dose (see Gable, 2004). Heroin is the most dangerous drug as the usual effective dose is almost 17% of the usual lethal dose. Heroin is followed by amphetamines and alcohol with a danger ratio of 10%. Cannabis (and tobacco) do not present an immediate danger of a lethal dose.

Health Effects of Illegal Drug Use

The literature that seeks to determine the impact of drug use on health is reviewed. As cannabis is the most widely used illegal drug, the focus is on the relationship between cannabis use and health. Much of the research in this area is contributed from epidemiology and is focused on the mental health effects of drug use. There is a smaller and more recent literature contributed by economics. A distinguishing feature of this literature is the utilization of methodologies designed to identify causal effects of drug use. In addition to studying the direct effects of drug use on health, the economics literature also considers the impact of drug use on labor market outcomes. Because a significant cost of poor health resulting from drug use is considered to be reduced labor market success, understanding the evidence regarding the indirect health effects of drug use is of significant interest. The indirect health effects of illegal drug use which originate, from effects on crime, violence, traffic accidents, etc., are not discussed. The relative contribution of illegal drug use to the economics costs of risky behavior more generally is discussed by Cawley and Ruhm (2011). They find that illegal drug use makes a modest contribution to these costs.

Medical and Epidemiological Literature

The earliest attempt to identify the causal impact of cannabis use on mental illness is by Andreasson *et al.* (1988) who study a cohort of more than 50 000 18- to 20-year-old Swedish conscripts. The authors find that the postconscription risk of developing schizophrenia is increasing in the number of times cannabis is used before conscription. This was a controversial finding and prompted a raft of epidemiological studies on the relationship between cannabis use and mental health more generally. This literature is so large that there is now a large number of studies dedicated to reviewing it.

In their 2003 review, Degenhardt *et al.* (2003) conclude that there is a modest but significant association between early-onset regular or problematic cannabis use and depression later in life, although there is little evidence of an association between depression and infrequent cannabis use. The authors go on to conclude that even if the association between cannabis use and depression is assumed to be causal, regular cannabis use can only explain a small proportion of depression in the population. Macleod *et al.* (2004) review more than 200 studies based on longitudinal data that seek to determine the psychosocial impact of cannabis use. They conclude that although there is evidence of associations between cannabis use and various measures of psychosocial harm, the extent of the associations and the strength of the evidence is not always large. Furthermore, the authors conclude that the causal nature of the associations is far from clear.

Many of the overview studies have focused on the relationship between cannabis use and psychosis. Arseneault *et al.* (2004) conclude on the basis of their review of previous research that cannabis use is likely to have a causal role in the development of psychosis but the magnitude of its impact is unclear. Kalant (2004) concludes from his review of previous studies that there is more evidence for a causal relationship running from cannabis use to psychiatric problems than there is for reverse causality, i.e., psychiatric problems leading to cannabis use. Henquet *et al.* (2005) review seven studies and conclude that cannabis use has a causal effect on later psychosis. They note, however, that the effect is not very large and the mechanism underlying the causality is unclear. Semple *et al.* (2005) provide an overview of 17 case–control studies that examined the association between cannabis use and schizophrenia or schizophrenia-like psychosis. They also

conclude that cannabis is a risk factor for psychosis but indicate that it is not clear whether cannabis is a precipitating or a causative factor in the development of schizophrenia. Hall (2006) argues on the basis of his review of studies that there is a strong association between cannabis use and psychosis, but it remains controversial whether the association is causal. Moore *et al.* (2007) present an overview of 11 studies on psychosis based on data from seven cohort studies. Although they find that there is an association between cannabis use and psychosis, they are unable to rule out spurious correlation resulting from unobserved confounding factors as the underlying explanation for this association.

In their recent review, Hall and Degenhardt (2009) argue that previous research on the relationship between mental health and illegal substance use has produced mixed findings, with some papers reporting a positive association between cannabis use and mental health problems and others reporting no association. McLaren *et al.* (2010) review the methodological strengths and limitations of major cohort studies that have sought to determine the causal nature of the relationship between cannabis use and psychosis. The authors conclude that, on the basis of the current studies, no inference can be made about a potential causal relationship from cannabis use to psychosis. Discussing a variety of papers Werb *et al.* (2010) conclude that the research to date is insufficient to conclusively claim that the association between cannabis use and psychosis is causal in nature. The fact that population-level rates of psychotic disorders do not appear to correlate with population-level rates of cannabis use suggests that these two phenomena may not be causally related.

Econometric Studies – Direct Health Measures

In examining the relationship between mental health and cannabis use, the literature from epidemiology cited above has attempted to identify the causal effect of cannabis use by controlling for observed factors that may be a source of confounding. However, as noted by Pudney (2010), the potential for unobserved common confounding factors makes inference regarding the causal impact of cannabis use difficult. In contrast, economic research routinely makes use of statistical techniques designed to account for unobserved confounding factors in studying the impact of one outcome on another. Despite the potential to provide strategies for addressing the issue of unobserved confounders, and thus better assess the health risks faced by drug users, there are very few contributions from the economics literature on this issue. As detailed below, the economic studies that do attempt to tease out causal effects suggest that there may be risks to both mental and physical health from using cannabis.

Williams and Skeels (2006) and van Ours and Williams (2011) use Australian data to study the impact of cannabis use on physical and mental health, respectively. Williams and Skeels (2006) find the probability of reporting very good or excellent self-assessed health to be 8% lower among those who consumed cannabis in the past year compared with those who had not, and 18% lower for those who reported weekly use. Along similar lines, van Ours and Williams (2011) find that cannabis use increases the likelihood of mental health problems, with the probability of experiencing mental distress increasing with the frequency of past year use. Although each of these studies considers a single dimension of health, there is significant evidence that poor mental health is correlated with poor physical health. van Ours and Williams (2012) investigate the impact of cannabis use on health in a framework that accounts for the potential for shared frailties in the domains of physical and psychological well-being, as well as selection into cannabis use. Their analysis of Amsterdam data suggests that cannabis use reduces the mental well-being of men and women and the physical well-being of men. Although statistically significant, the magnitude of the effect of using cannabis on mental and physical health is found to be small.

van Ours *et al.* (2013) is the only study to address both the potential for common unobserved confounders and reverse causality in studying the health impact of cannabis use. Their analysis of the relationship between suicidal ideation and cannabis use is based on a 30-year longitudinal study of a birth cohort. They find that intensive cannabis use – at least several times per week – leads to a higher transition rate into suicidal ideation for susceptible males. There is no evidence that suicidal ideation leads to regular cannabis use for either males or females.

Econometric Studies – Indirect Health Measures

In addition to their stock of human capital, a person's labor market productivity is determined by their health capital stock (Grossman and Benham, 1974). Drug use is conjectured to reduce labor market productivity through its deleterious effects on an individual's stock of health. Although intuitively appealing, empirically assessing the validity of this conjecture is complicated by the fact that individuals choose, or self-select into, drug use. Specifically, there may be important unobserved determinants of wages or employment that also influence the decision to use drugs. An example of an omitted variable particularly relevant in this context is an individual's discount rate. Individuals who discount the future heavily are more likely to use drugs because they place little weight on the future negative health consequences of their drug use (Becker and Murphy, 1988). They are also more likely to choose jobs with little investment in on-the-job training, and that consequently pay relatively high current wages but relatively low future wages. This may give rise to a positive correlation between drug use and wages even if drug use is negatively causally related to wages. Similarly, individuals with strong preferences for leisure may also be more likely to use drugs if drug use and leisure are complements in the production of euphoria. Such a relationship would produce a negative correlation between drug use and labor supply even in the absence of a causal effect of drug use on labor supply.

The empirical strategy pursued by the first-wave studies for estimating the causal impact of drug use on wages and employment is instrumental variables. Three of these studies draw on data on 18- to 27-years old from the 1984 cross section of the National Longitudinal Survey of Youth (NLSY) and all three studies found evidence that, rather than reduce wages, drug use increases wages. Kaestner (1991) finds that for

males, drug use measured as past 30-day use of cannabis, lifetime use of cannabis, past 30-day use of cocaine, or lifetime use of cocaine, raises hourly wages. Similarly, male wages are found to be increasing in the frequency of cannabis use in the past 30 days by Register and Williams (1992). Gill and Michaels (1992) report that the use of any drugs in the past year or any hard drugs (cocaine, heroin, inhalants, psychedelics, other drugs, other narcotics) in the past year increases the hourly wage rate received in a combined sample of males and females. The estimated magnitudes of the wage effects are quite large. For example, Kaestner (1991) estimates that males who have tried cannabis earn 18% more than otherwise similar males who have not tried cannabis, Register and Williams (1992) estimate that using cannabis on one more occasion per month increases hourly wages by 5%, and Gill and Michaels (1992) find that drug users earn approximately 4% more per hour than nonusers, and that hard drug users earn approximately 10% more per hour than nonhard drug users. Moreover, both Kaestner (1991) and Gill and Michaels (1992) report that the premiums for drug use are attributable to unobserved differences between the users and nonusers and not differences in returns to human capital and other characteristics.

Kaestner (1994a,b) uses the 1984 and 1988 waves of the NLSY to compare cross-sectional and longitudinal estimates of the impact of cocaine and cannabis use on labor supply and wages, respectively. He finds that the results based on the 1984 data, which show that cannabis and cocaine use increases wages and cannabis use decreases hours spent working in the sample of males, cannot be replicated using the 1988 data. Moreover, when unobserved differences that affect drug use and labor market outcomes are controlled for through a fixed-effect estimator, drug use is found to have a negative but insignificant impact on wages for males (Kaestner, 1994b), and mixed, although generally insignificant, effects on hours worked (Kaestner, 1994a). The overall conclusion reached by Kaestner is that drug use does not have a systematic impact on labor supply or wages.

The counterintuitive and inconsistent findings of the above studies motivated a second wave of economic research into the impact of drug use on wages and labor supply. Taken at face value, most of the second-wave studies tend to find evidence that nonproblematic use of drugs (light to moderate use, or the use of soft drugs) has no impact on labor supply, measured by employment or hours worked, but that problematic use (heavy use, or the use of hard drugs) does, although Burgess and Propper (1998); DeSimone (2002); Zarkin *et al.* (1998) and van Ours (2006) provide counterexamples. Similarly, most of the second-wave studies find that infrequent or nonproblematic drug use has no impact on wages, whereas problematic use does have negative wage effects. Once again, there are also exceptions to this generalization, such as MacDonald and Pudney (2000). It is noteworthy that many of these studies (especially those based on US data) tend to treat drug use as exogenous to labor market outcomes.

Focusing on the studies that are more rigorous in their efforts to address the potential endogeneity of drug use, the results are mixed. For example, although van Ours (2007) finds that using cannabis at least 25 times in one's lifetime reduces the wage of prime-age males, the use of cocaine is found to have no effect, and MacDonald and Pudney (2000) are unable to detect any impact of either hard or soft drug use on their proxy for wages, that is, occupational attainment. Similarly, with respect to the employment of males, DeSimone (2002) finds that both past year cannabis and cocaine use reduces the probability of employment, whereas, MacDonald and Pudney (2000) find no employment impact of soft drug use (which includes cannabis) and van Ours (2006) finds no impact of cannabis or cocaine use on employment. Finally Conti (2010) introduces cognitive ability as additional variable in a wage equation with cannabis use as explanatory variable, showing that this causes the effect of cannabis use to become insignificantly different from zero.

Given the conflicting nature of the empirical findings, it is simply uncertain as to whether there are negative labor market consequences of drug use in general, and cannabis use in particular. Furthermore, it is unclear as to whether this literature should be interpreted as reflecting a lack of robust evidence of a negative health effect of drug use, or as reflecting the presence of a productivity improving effect of drug use that is confounding the negative health effects.

Discussion and Conclusion

The use of illegal drugs is limited to a small part of the population. Not many people consume amphetamines, ecstasy, cocaine, or heroin. The most popular illegal drug by far is cannabis. However, even for the most popular illegal drug, heavy use is quite rare. And whereas a substantial proportion of the population has used cannabis in their lifetime, for many their use was a short-lived experiment. Even among individuals who persist in cannabis use, many do so on a recreational basis. Despite a large number of epidemiological studies and a handful of econometric studies little is known with any degree of certainty about the health effects of illegal drug use. Researchers agree that drug use is associated with worse health. The issue is whether this association is causal, with drug use causing poor health, or whether spurious correlation or reverse causality underlies this association. The main impediment to determining the nature of the relationship between illegal drug use and health is that the optimal setup for addressing this issue is a randomized control trial in which individuals are randomly allocated to the treatment group (who are administered illegal drugs) or to the control group (who receive a placebo). However, this type of experiment is not possible for at least two reasons. First and foremost, individuals will always know whether they are in the treatment group receiving illegal drugs, or in the control group receiving the placebo. Second, long-term exposure to illegal drugs would be necessary in order to determine the health effects, and this would be rather unethical should the outcome be that there are serious health problems related to illegal drug use. The so-called 'natural experiments,' in which a policy change that affects drugs use is exploited as if it were an experiment, are rare simply because drug policies have the tendency not to change.

The lack of econometric research that seeks to identify causal effects of drug use on health is surprising but likely to

be related to lack of good data as a basis for the research. Drug use is not a static phenomenon. On the contrary, dynamics in use are very important. Within the population some individuals may start using a drug but others will abstain. Among those who have started using a drug there are individuals who will stop using and other individuals who will persist in drug use. By and large, in the population there are never-users, experimental users, and persistent users. Even within the group of persistent users there may be transitions from high intensity of use to low intensity of use and vice versa. To understand the dynamics of illegal drug use, information is needed from the time when individuals are first confronted with the choice of whether to use a particular drug. Ideally, this information should capture how relevant circumstances change over time. Information that could be relevant includes: family situation, experiences at school, changing drug supply conditions, and drug prices. Unfortunately, this type of information is not typically available. Another issue which makes it hard to research in this area is the fact that it is hard to quantify drug use. Whereas standard quantity measures are available for tobacco (cigarettes per day) and for alcohol (standard units of alcohol per day), there are no obvious standard quantity measures for the use of illegal drugs.

Despite the absence of experimental research it is still possible to draw some conclusions from previous research on the direct and indirect health effects of illegal drug use. Intensive use of illegal drugs over a long period of time generates negative health effects for its users whereby the magnitude depends on the nature of the drug involved. Whether short-term use or long-term, recreational use is harmful is not clear. For cannabis, the evidence finds that use is neither necessary nor sufficient for mental health problems to occur. It could be that individuals who are susceptible to mental health problems are vulnerable for cannabis use, but as yet this is unclear. Most likely, experimenters will not suffer serious health effects, whereas the same holds for persistent but recreational users. The group of persistent heavy users is at risk of negative health effects. However, the size of this group is limited to 1% or 2% of the adult population. In this sense, from an aggregate point of view, the magnitude of the health effects of illegal drug use is limited. Nevertheless, for individuals the negative health effects may be severe. How severe it may be is yet to be established.

Given the limited circumstances for which cannabis use may pose a threat of harm, there is growing interest in possible medical applications of cannabis, the so-called 'medical marijuana' most notably as a treatment for the symptoms of muscle spasm and tremors in multiple sclerosis patients and the symptoms of vomiting and nausea in cancer patients undergoing chemotherapy (Hall *et al.*, 2001). Cannabinoids may allay pain, improve sleep, and possibly inhibit degenerative processes (McCarberg, 2007). Caulkins *et al.* (2012) refer to a summary of 12 double-blind clinical trials where 57% report positive outcome of cannabis use, 33% found no effect and 10% found adverse outcomes. Research on the therapeutic use of cannabis and cannabinoid drugs is hampered by 'Catch 22' situation that as long as cannabis is illegal the medical benefits cannot be established in a way that it would be accepted as a treatment and cannabis remains illegal if the medical benefits of cannabis use cannot be established. Nevertheless, 18 US states and the District of Columbia allow patients who have a recommendation from a doctor to use cannabis for medical purposes without the risk of being prosecuted.

When assessing the health effects of illegal drug use some caveats are important to keep in mind. First, all health effects are established under one type of policy regime, prohibition. Although there is variation in the way prohibition is implemented, there is no country or jurisdiction that has legalized selling, buying, or using any illegal drug. In the USA, Colorado and Washington states have recently passed referendums to legalize cannabis but at the time of writing, the framework for implementing legalization was yet to be established. However, the legal status of a drug may affect the relationship between drug use and health. Furthermore, because it is an illegal activity, it is not easy to collect reliable data on drug use. A second caveat is that the health consequences of using an illegal drug are likely to depend on the manner in which it is consumed. Smoking heroin, for example, is less dangerous than injecting heroin and inhaling cannabis that has been vaporized is less dangerous than smoking cannabis. A third caveat to keep in mind is that the health risks posed by specific illegal drugs may have changed over time. For example, in recent years, the proportion of Δ^9-tetrahydrocannabinol present in cannabis is thought to have risen, whereas the proportion of cannabidiol is thought to have decreased. Δ^9-tetrahydrocannabinol is believed to exaggerate the psychotic effects of cannabis, whereas cannabidiol is thought to moderate the psychotic effects. However, due to paucity of information on the composition of cannabis, the health effects of any changes are unknown.

It is concluded that adverse health effects of cannabis use are clearly present but their magnitude seems rather limited. Nevertheless, using illicit drugs is not good for one's health. Even cannabis, which is considered to be a soft drug in some countries because of its limited health effects, has a negative health effect. Whether one should worry about this is another matter. In the grand scheme of things cannabis use – and even hard drug use – has a limited health effect compared with other risky behavior. Heavy cannabis use and early onset of cannabis use, which often but not always coincide, have the largest negative health effects. Preventing youngsters from starting to use cannabis or least preventing them from doing this early on in life could be sufficient to prevent serious health effects.

As to the health effects of other illegal drugs the weight of evidence supports the finding that the harms associated with cannabis use are much less serious than those associated with 'hard' drugs such as cocaine or heroin and may even be smaller than those associated with alcohol and cigarettes. And although it is generally acknowledged that there are risks associated with long-term heavy use of cannabis such as respiratory diseases, cancer, and perhaps psychotic disorders, only a small fraction of those who ever use cannabis actually become long-term heavy users.

See also: Addiction. Alcohol. Mental Health, Determinants of. Peer Effects in Health Behaviors. Smoking, Economics of

References

Andreasson, S., Engstrom, A., Allebeck, P. and Rydberg, U. (1988). Cannabis and schizophrenia: A longitudinal study of Swedish conscripts. *Lancet* **2**(8574), 1483–1486.

Arseneault, L., Cannon, M., Witton, J. and Murray, R. (2004). Causal association between cannabis and psychosis: Examination of the evidence. *British Journal of Psychiatry* **184**, 110–117.

Becker, G. and Murphy, K. (1988). A theory of rational addiction. *Journal of Political Economy* **96**(4), 675–700.

Burgess, S. M. and Propper, C. (1998). Early health related behaviours and their impact on later life chances: Evidence from the US. *Health Economics* **7**(5), 381–399.

Caulkins, J., Hawken, A., Kilmer, B. and Kleiman, M. (2012). *Marijuana legalization: What everyone needs to know*. Oxford, UK: Oxford University Press.

Caulkins, J., Reuter, P. and Coulson, C. (2011). Basing drug scheduling decisions on scientific ranking of harmfulness: false promise from false premises. *Addiction* **106**, 1886–1890.

Cawley, J. and Ruhm, C. J. (2011). The economics of risky health behaviors. *Handbook of health economics*, vol. 2, pp 95–199. Amsterdam: North-Holland.

Conti, G. (2010). Cognition, cannabis and wages. Mimeo.

Degenhardt, L., Hall, W. and Lynskey, M. (2003). Exploring the association between cannabis use and depression. *Addiction* **98**, 1493–1504.

DeSimone, J. (2002). Illegal drug use and employment. *Journal of Labor Economics* **20**(4), 952–977.

Gable, R. (2004). Comparison of acute lethal toxicity of commonly abused psychoactive substances. *Addiction* **99**, 686–696.

Gill, A. and Michaels, R. J. (1992). Does drug use lower wages? *Industrial and Labor Relations Review* **45**(3), 419–434.

Grossman, M. and Benham, L. (1974). Health, hours and wages. In Perlman, M. (ed.) *The economics of health and medical care: Proceedings of a conference held by the International Economic Association at Tokyo*, pp 205–233. London: Macmillan.

Hall, W. (2006). Cannabis use and the mental health of young people. *Australian and New Zealand Journal of Psychiatry* **40**, 105–113.

Hall, W. and Degenhardt, L. (2009). The adverse health effects of non-medical cannabis use. *Lancet* **374**, 1383–1391.

Hall, W., Degenhardt, L. and Currow, D. (2001). Allowing the medical use of cannabis. *Medical Journal of Australia* **175**, 39–40.

Henquet, C., Murray, R., Linszen, D. and Van Os, J. (2005). The environment and schizophrenia: The role of cannabis use. *Schizophrenia Bulletin* **31**(3), 608–612.

Kaestner, R. (1991). The effect of illicit drug use on the wages of young adults. *Journal of Labor Economics* **9**(4), 381–412.

Kaestner, R. (1994a). The effect of illicit drug use on the labor supply of young adults. *Journal of Human Resources* **29**(1), 126–155.

Kaestner, R. (1994b). New estimates of the effect of marijuana and cocaine use on wages. *Industrial and Labor Relations Review* **47**(3), 454–470.

Kalant, H. (2004). Adverse effects of cannabis on health: An update of the literature since 1996. *Progress in Neuro-Psychopharmacology* **28**, 849–863.

Kendell, R. (2003). Cannabis condemned: the proscription of Indian hemp. *Addiction* **98**, 143–151.

MacDonald, Z. and Pudney, S. (2000). Illicit drug use, unemployment and occupational attainment. *Journal of Health Economics* **19**(6), 1089–1115.

Macleod, J., Oakes, R., Copello, A., et al. (2004). Psychological and social sequelae of cannabis and other illicit drug use by young people: A systematic review of longitudinal, general population studies. *Lancet* **363**, 1579–1588.

McCarberg, B. (2007). Cannabinoids: Their role in pain and palliation. *Journal of Pain & Palliative Care Pharmacotherapy* **21**(3), 19–28.

McLaren, J., Silins, E., Hutchinson, D., Mattick, R. and Hall, W. (2010). Assessing evidence for a causal link between cannabis and psychosis: A review of cohort studies. *International Journal of Drugs Policy* **21**, 10–19.

Moore, T., Zammit, S., Lingford-Huges, A., et al. (2007). Cannabis use and risk of psychotic or affective mental health outcomes: A systematic review. *Lancet* **370**, 319–328.

Nutt, D., King, L. and Phillips, L. (2010). Drug harms in the UK: A multi-criteria decision analysis. *Lancet* **376**, 1558–1565.

van Ours, J. C. (2005). Dynamics in the use of drugs. *Health Economics* **15**(12), 1283–1294.

van Ours, J. C. (2006). Cannabis, cocaine and jobs. *Journal of Applied Econometrics* **21**, 897–917.

van Ours, J. C. (2007). The effect of cannabis use on wages of prime age males. *Oxford Bulletin of Economics and Statistics* **69**, 619–634.

van Ours, J. C. and Williams, J. (2011). Cannabis use and mental health problems. *Journal of Applied Econometrics* **26**, 1137–1156.

van Ours, J. C. and Williams, J. (2012). The effects of cannabis use on physical and mental health. *Journal of Health Economics* **31**, 564–577.

van Ours, J. C., Williams, J., Fergusson, D. and Horwood, L. (2013). Cannabis use and suicidal ideation. *Journal of Health Economics* **32**(3), 524–537.

Pudney, S. (2010). Drugs policy – what should we do about cannabis? *Economic Policy* **61**, 165–211.

Register, C. and Williams, D. (1992). Labor market effects of marijuana and cocaine use among young men. *Industrial and Labor Relations Review* **45**(3), 435–448.

Room, R., Fischer, B., Hall, W., Lenton, S. and Reuter, P. (2010). *Cannabis policy: Moving beyond stalemate*. Oxford, England: Oxford University Press.

Semple, D., McIntosh, A. and Lawrie, S. (2005). Cannabis as a risk factor for psychosis: systematic review. *Journal of Psychopharmacology* **19**(2), 187–194.

United Nations (2011). *World Drugs Report 2011*. Vienna, Austria: United Nations Office on Drugs and Crime.

Werb, D., Fischer, B. and Wood, E. (2010). Cannabis policy: time to move beyond the psychosis debate. *International Journal of Drug Policy* **21**, 261–264.

Williams, J. and Skeels, C. L. (2006). The impact of cannabis use on health. *De Economist* **154**, 517–546.

Zarkin, G. A., Mroz, T. A., Bray, J. W. and French, M. T. (1998). The relationship between drug use and labour supply for young men. *Labour Economics* **5**(4), 385–409.

Impact of Income Inequality on Health

J Wildman and J Shen, Newcastle University, Newcastle Upon Tyne, UK

Introduction: What Are Health Inequalities?

Health inequalities are observed in all societies. Although some inequalities may be considered unavoidable, resulting from sociodemographic characteristics such as age, gender, and genes, many of these health inequalities are associated with socioeconomic characteristics that are potentially amenable to policy interventions and could be considered as avoidable. In Europe, measuring and understanding these differences have been the major part of the literature on 'health inequalities.' In the US, these types of analyses are often referred to as 'health disparities.' For the purpose of this section the term 'health inequalities' will be used. It will also be assumed that it is clear what is meant by 'health.' Health, in this section could refer to a range of outcomes such as coronary heart disease, remaining expected quality-adjusted life-years (QALYs), length of life lived, mortality, morbidity, etc.

Avoidable health inequalities are commonly defined as unfair systematic differences in health outcomes, although whether such inequalities are unfair may depend on the equity criterion applied. Inequalities are not generally considered unjust in cases where genes or the human body's natural capacity are largely at play, for example, women tend to live longer than men, or 20-year olds in general have better health than 60-year olds. However, the marked differences evident in the populations of some countries in mortality rates (and other health measures) between occupational classes, between regions, between races, and between the rich and the poor are all considered to be examples of avoidable and unfair health inequalities. Researchers across the disciplines of economics, sociology, epidemiology, and psychology have suggested various theories that could explain health inequalities, and among these, theories regarding the influence of material factors – especially income – have been fundamental to the research into health inequalities.

The Link between Income and Health

The association between levels of income and health is well documented, with research suggesting that income levels and health outcomes are positively correlated. At the individual level (and controlling for other factors such as age, gender, and socioeconomic characteristics), income is often found to be a significant predictor of health. However, the direction of causality is difficult to identify: Does higher (lower) income lead to better (worse) health, or does health affect income? Further, the causality may be direct, or indirect, with income and health affecting each other via mediating factors. It is also possible, although not likely, that there may be a third factor affecting both health and income, giving the impression of a relationship but without any causal link between them.

At the aggregate level, cross-country comparisons have shown that higher average income (gross domestic product (GDP) per capita) correlates with higher average health (in this case, measured as life expectancy). This is often known as the absolute income hypothesis (AIH). This evidence can be found for both cross-sectional and longitudinal studies. The way average health has improved along with average income is clearly demonstrated here: http://www.youtube.com/watch?v=jbkSRLYSojo

These data give a clear animated illustration of Preston curves that demonstrate a concave relationship between life expectancy and average income (measured as GDP per capita). This means that as average income increases, life expectancy increases at a decreasing rate, or still more simply, a proportionate increase in average income is associated with larger health gains at lower initial levels of average income than at higher initial levels of income.

However, this aggregate-level result does not seem to hold when GDP per capita reaches a certain level. In developed countries that have passed through the 'epidemiological transition,' where the main causes of death are chronic conditions rather than contagious diseases, there is little evidence of a relationship between income and health (countries are on the flatter part of the Preston curve). It is worth noting that the flatter part of the Preston curve only implies that there is no difference in average health by average income across societies, but there can still be variations in average health across income groups within societies. Based on this evidence, it has been proposed that absolute income is not the main determinant of health in developed countries.

Relative Income Hypothesis

In developing countries, individual absolute income seems to be the main determinant of individual health. If income or material factors are important for health, then continuing growth in income should result in increasing health. Thus, for example, if mortality risk (or any other health outcome) at the individual level is convex (as shown in Figure 1) so that the risk of death decreases at a decreasing rate as income increases, then health inequalities should decrease as countries became richer – for example, 'as they progress toward becoming developed countries. As income grows, those at the top end of the distribution see their health improve but at a slower rate than that of individuals at the lower end of the distribution (because of the steeper gradient for these individuals). So over time, health inequalities will disappear if all individuals see the benefits of income growth: if the relationship at the individual level becomes completely flat, then even if the income of the most wealthy grows at a faster rate than that of the least wealthy, health inequalities will diminish as long as the least wealthy experience some income growth.

 doi:10.1016/B978-0-12-375678-7.00209-1

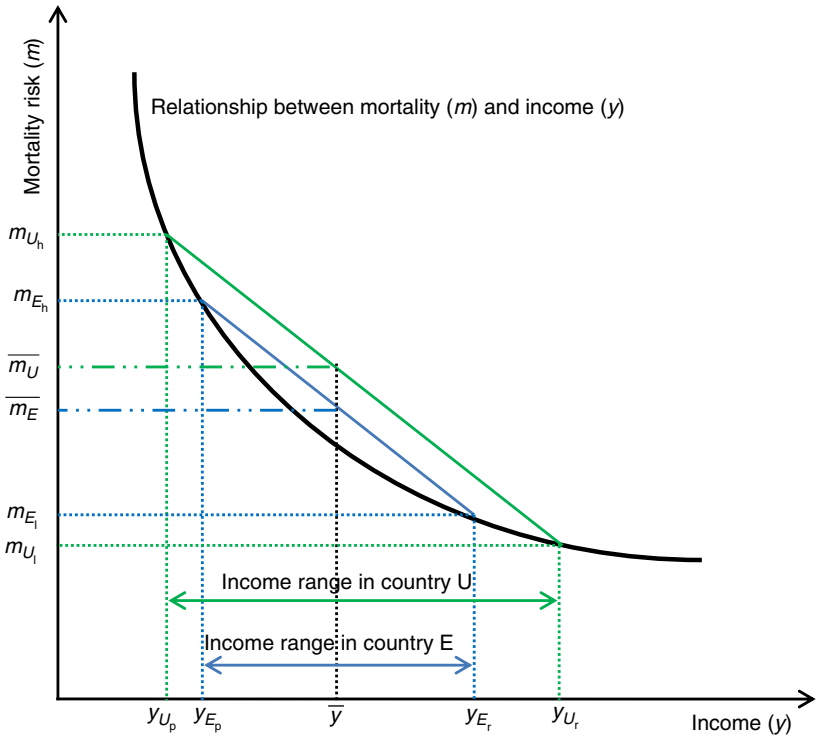


Figure 1 Effect of increased inequality of income on population mortality.

However, this is not what is observed: health inequalities persist in developed countries and are even increasing in some instances. Although this may be explicable by the AIH due to those in the lower socioeconomic groups not benefiting from income growth, it may be that an alternative explanation is more likely. The most influential alternative theory to date is the relative income hypothesis (RIH), which has been developed most notably by Richard Wilkinson. In his groundbreaking work, Wilkinson (1996) identifies the possibility that it is the distribution of income within a country or region rather than the absolute level of income that causes inequalities in health. The essence of the theory is that income inequality negatively affects individual health. In its strongest form, this means that all individuals, even the richest, experience worse health when there is high income inequality. Weaker forms involve only those individuals who are at the lower end of the socioeconomic spectrum experiencing worse health.

There is a large body of empirical evidence supporting the RIH using aggregate-level data. Measures of income inequality – the Gini coefficient is the most widely used – have been shown to have a significant negative association with health (measured either by life expectancy or by infant mortality): the larger a country's Gini coefficient, the higher the income inequality and the worse its health outcomes (Box 1).

Studies investigating the relationship between average health and the share of income (by quintiles or percentiles of the study population) or the percentage of the population in relative poverty, all seem to confirm that the wider a country's income distribution, the worse the average level of health in that country appears to be.

Box 1 What is the Gini coefficient?

The Gini coefficient is derived from the Lorenz curve. To derive a Lorenz curve, the population is ranked by income and the cumulative proportion of income is plotted against the cumulative proportion of the population. If income is equally distributed, this plots a 45° line (the line of perfect equality). If income is unequally distributed, it plots a line below the 45° line. The Gini coefficient measures the area between the Lorenz curve and the line of perfect equality.

Theories Behind the Relative Income Hypothesis

The RIH has its theoretical basis in, and is supported by, evidence from sociology and epidemiology. From a sociological perspective, social cohesion and social capital (such as trust, participation, and social inclusion) are the bases for the RIH, with income inequality acting as a proxy for either a lack of social cohesion or a lack of social capital. Countries with a wide distribution of income are assumed to have fewer social support mechanisms, thereby leading to higher crime rates and a diminished quality of social environment. As a result, the health of every individual in these unequal societies will be affected directly via disease development or hindered recovery, and indirectly through health-damaging behaviors (such as smoking or substance abuse) when individuals at the lower end of the social scale react to their adverse circumstances. The psychosocial effects of living in an unequal society could also support the RIH, as inequality may cause stress

or anger due to either a lack of social support networks or an inability to maintain a socially acceptable standard of living.

Epidemiology provides a large body of evidence of the impact of social status on health. Much of this work is summarized by Marmot in his book, Status Syndrome (2004). Studies suggest that health inequalities reach up the social scale because hierarchy and social regimentation are harmful to health: people at every level of the 'pecking order' suffer worse health than those above them. Income or income inequality acts as a proxy for the control in one's life. In wider hierarchies, those at the bottom suffer more than those at the top. Such a premise is partly based on the flight or fight syndrome – where the body produces a reaction in times of stress of whether to fight or take flight. With the body under stress, there are detrimental impacts on individual health.

Neither of these approaches is without problems. The sociological psychosocial approach looks only at psychological effects and appears to disregard the material, behavioral, or biological factors that may cause ill health. The epidemiological evidence is criticized because it often does not control for income as being part of the study. Without controlling for individual income, it is always possible that this confounds the relationship between health and income inequality.

Box 2 Health production function

The individual health production function is analogous to a firm's production function. As firms use combinations of capital and labor to produce the output, the individual uses combinations of goods to produce health. In its simplest for it can be considered that the only input to health is income. For a general relationship that gives:

$$H_i = f(Y_i)$$

This reads as individual health (H_i) being some unspecified function of individual income (Y_i). This would represent the AIH.

If health is measured as mortality risk and the function is decreasing, then the first derivative of the health production function would be negative (demonstrating that mortality risk falls as income increases). And if the function is nonlinear (convex), the second derivative would be negative (mortality risk falls at a decreasing rate – so the extra unit of income causes the mortality risk to fall, but by a smaller amount than that of the previous unit of income).

For the RIH, at the individual level, it is possible to specify a health production function of the form:

$$H_i = f(Y_i G)$$

This shows individual health (H_i) as being a function of both individual income (Y_i) and income inequality (G), measured at an appropriate level.

Aggregate-Level Data Problems

Much of the data used to investigate the relationship between income inequality and health are at the aggregate level, and although these 'aggregate-level data studies have provided considerable evidence in favor of the RIH, it is important to recognize their limitations. The key limitation of aggregate-level studies is the aggregation problem, which occurs when the existence of a nonlinear relationship between health (e.g. mortality risk) and income at the individual level leads to spurious results at the aggregate level.

The aggregation problem is best explained by using an example: Assume that absolute income is the only factor affecting individual health (AIH) and the relationship is nonlinear, so the relationship between mortality risk and income is convex – as income increases the risk of death decreases at a decreasing rate. This situation was described earlier and it is illustrated in Figure 1. The framework can be considered as a health production function (*see* Box 2). In this situation, income inequality has no impact on individual health – it is absolute income and not relative income that matters.

Now imagine that there are two countries (these are illustrated in Figure 1 as Evenland (E) and Unevenland (U)). In each country, there are only two groups: the rich (y_r) and the poor (y_p). The average level of income is the same in both countries ($\bar{y}$), but in Evenland, the difference between the incomes of the rich and the poor is smaller than in Unevenland.

The relationship between income and mortality risk (the graphical version of our convex health production function) is the same for both countries, which is the convex curve in Figure 1. Aggregating individuals and comparing average health in each country, shows that the average risk of mortality is lower in Evenland ($\overline{m_E}$) than it is in Unevenland ($\overline{m_U}$). From this aggregate-level evidence, it may be concluded that the distribution of income has a negative impact on health at the individual level; nevertheless, in fact, this would be a spurious conclusion because at the individual level, it has been assumed that there is no relationship between income distribution and health – it is only absolute income that matters.

In this example, the result $\overline{m_E} < \overline{m_U}$ can be explained solely by the AIH with no reference to the RIH. This occurs simply because of the nonlinear relationship between health and income at the individual level: the poor individuals in Unevenland are on a steeper part of the production function than those in Evenland. Conversely, the rich individuals in Unevenland are on a flatter portion of the production function than those in Evenland. The aggregation problem is demonstrated in its mathematical form by Gravelle *et al.* (2002).

Despite the aggregation problem, this example does show that more even distributions of income have better health on average than more unequal distributions. Again consider Figure 1. This time instead of thinking of two countries being compared consider the same country at two different time points – time E and time U. There are still two groups, the poor (p) and the rich (r), and between times E and U, there is a redistribution of income from the poor to the rich that leaves average income unchanged (at $\bar{y}$), but the income gap between the rich and the poor widens. Following the redistribution, the income of the poor falls from y_{E_p} to y_{U_p}, whereas the income of the rich increases from y_{E_r} to y_{U_r}. This leads to the mortality risk of the poor increasing from m_{E_h} to m_{U_h}, and the mortality risk of the rich decreasing from m_{E_l} to m_{U_l}. The increase in mortality risk for the poor outweighs the fall in risk for the rich, so $\overline{m_E} < \overline{m_U}$ and overall mortality risk increases. This result stems purely from the impact of individual income on

individual health (as predicted by the AIH) but clearly demonstrates why the distribution of income is important.

The aggregation problem highlights the need for individual-level studies to explore the RIH. Individual-level studies allow the exploration of the link between income inequality and health without having to deal with the aggregation problem. Lynch *et al.* (2004) have conducted a systematic review of the literature, investigating income inequality and health, and Jones and Wildman (2008) have considered the literature investigating relative deprivation and health. Although many of the results have been mixed – perhaps due to difficulties in a number of methodological and empirical issues, to be discussed in the next section – a recent meta-analysis does suggest a significant, if not causal, relationship between income inequality and health (Kondo *et al.*, 2009). It is likely that income inequality and health are related at the individual level; however, there are many unresolved issues before reaching a more definitive conclusion.

Unresolved Issues

The RIH presents a number of unresolved issues. Firstly, studies often use cross-sectional data that assume a contemporaneous relationship between income, income inequality, and health. This assumption raises an identification problem, which has not been dealt with adequately. When considering the mechanisms by which income inequality may affect health, such as stress generated by being of low social status, lack of social cohesion, or an inability to purchase status goods to 'keep up with the Joneses,' then the contemporaneous specification may not be detecting the true nature of the influence of income inequality on health. The impact of all these mechanisms on health takes time to develop; for example, being of low status may have a cumulative detrimental impact over time, so *ceteris paribus* the impact on health will be greater for older individuals. Longitudinal data are needed to examine the impact of income inequality over an individual's life course and whether the impact increases in severity over time.

Secondly, health can be measured across many different dimensions (e.g. expected lifetime QALYs, self-assessed health, long-standing limiting illness), but not all of these are sensitive to the effects of income inequality. If the psychosocial theory is correct, one would expect income inequality to have a greater effect on mental health measures than on measures of general health such as self-assessed health or physical health such as mortality or certain chronic illnesses. In addition, one may also expect a link between mental health and physical health, but the transitional effect from mental health to physical health may take time to develop. Furthermore, even though the observation of individuals over time provides the ability to control for unobservable heterogeneity, there are rarely data available that allow for the examination of the impact of inequality over the life course. So, if income inequality affects mental health, it may take even longer for the impacts to be revealed in more general health measures such as chronic illness or self-assessed health that are commonly collected in population surveys. To detect the psychosocial impact of income inequality on health, there is a need for longitudinal data with good measures of mental health.

Thirdly, investigating the RIH requires the construction of a relative income measure, namely, how an individual compares his or her income in society against a particular reference group; therefore, it is inevitable that investigations into the RIH are affected by the choice of a reference group against which individuals compare their income. There is no consensus in the literature on the reference group and there is no empirical solution to the problem – it is not possible to determine reference groups by observing behaviors because the choice of the reference group can itself be endogenous. Individuals may choose to compare their own income either with the average income of the country (region/town) they live in, or with the income of their peers, neighbors, people in the same age group, or any other plausible reference groups. Many of these reference group definitions have been used for research in this area.

A further issue for researchers is the measure of income inequality. The way in which this variable is constructed is yet another key element in understanding how income inequality affects health. The choice of measure can determine how income inequality appears to affect health. As noted above, the Gini coefficient is a commonly used measure of income inequality. Because this measure is an aggregate-level measure, there is only one Gini coefficient for any specific population wherein its use assumes that all individuals in that population are affected by income inequality in the same way. For example, in cross-national studies, each country has one Gini coefficient for any given year, and this means that there is no differential impact of income inequality for individuals within that country. Other methods have tried to create measures of income inequality that vary across individuals, and these measures are often considered to be measuring relative deprivation. Such measures compare an individual's income to a reference point, which may be the median or the highest income in an area. Such an approach acknowledges that income inequality may affect some individuals more than others as the relative income deprivation measure of someone being further away from, for example, the median income, is larger than that of someone being closer to it. This does raise an issue about the asymmetry of the inequality effect – individuals are negatively affected by having people above them in the income distribution, but they are unaffected by having people below them. It may be possible that individuals gain satisfaction from looking down on people in the income distribution, but this position has not been widely considered in the literature, partly because it is difficult to disentangle the positive effects of being above people from the negative effects of being below them for any given individual in a distribution (except the one at the very bottom end and the one at the top end of the income distribution).

Finally, there are theoretical or modeling issues that may be fundamental to examining the RIH. The measures of income inequality are often functions of individual income, which may cause multicollinearity in a regression while controlling for the effect of income. The income inequality measures may directly enter into the health production function or utility function, or enter indirectly through third factors, or both, and this may require a whole new theoretical framework for constructing the relationship between income inequality and

health. It could also be that as relative concerns allow such a broad range of behavior, their inclusion in choice models that theoretically consider individual behaviors may give them little or no predictive power. For these reasons, it is important to research and develop a proper theory underpinning the study of the RIH.

Conclusion

There is a substantial body of evidence linking inequalities in health with material factors, with income being considered as the most important factor. Recently, the RIH has been identified as an alternative explanation of health inequalities in developed nations. Initially, strong support for the RIH was provided by aggregate data studies, but these have been criticized because of the aggregation problem. Individual-level studies that overcome problems of aggregation have reported mixed results.

In recent years, research on income inequalities has widened its focus. Wilkinson and Pickett (2009) have considered the relationship between income inequality and a whole range of outcomes, including health and health behaviors (such as drug and alcohol addiction), social mobility, crime, well-being, and educational performances. This consideration of the relationship between income inequality and a wider range of outcomes suggests the importance of understanding the causal pathways at play.

Among both supporters and critics of the RIH, there appears to be a consensus calling for more research to model the effects of relative income on health from a broader perspective. Individuals live in societies and their behaviors need to be modeled and placed within a macro context in order to fully understand the relationship between income inequality and health; this would include individual-level characteristics and macro-level social factors such as social capital, social support mechanisms, and societal structures that cause inequalities. Developing a model to account for all these factors is the challenge for future research.

See also: Dominance and the Measurement of Inequality. Equality of Opportunity in Health. Health Status in the Developing World, Determinants of. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach. Unfair Health Inequality

References

Gravelle, H., Wildman, J. and Sutton, M. (2002). Income, income inequality and health: What can we learn from aggregate data? *Social Science and Medicine* **54**(4), 577–589.

Jones, A. M. and Wildman, J. (2008). Health, income and relative deprivation: Evidence from the BHPS. *Journal of Health Economics* **27**(304), 308–324.

Kondo, N., Sembajwe, G., Kawachi, I., et al. (2009). Income inequality, mortality, and self-assessed health. *British Medical Journal* **339**, b4471.

Lynch, J., Davey Smith, G., Harper, S., et al. (2004). Is income inequality a determinant of population health? Part 1. A systematic review. *Millbank Quarterly* **82**, 5–99.

Marmot, M. (2004). *Status syndrome*. London: Bloomsbury.

Wilkinson, R. (1996). *Unhealthy societies: The afflictions of inequality*. London: Routledge.

Wilkinson, R. and Pickett, K. (2009). *The spirit level: Why equality is better for everyone*. London: Penguin.

Further Reading

Deaton, A. (2003). Health, inequality, and economic development. *Journal of Economic Literature* **41**(1), 113–158.

Gravelle, H. (1998). How much of the relationship between population mortality and unequal distribution of income is a statistical artefact? *British Medical Journal* **314**, 382–385.

Wagstaff, A. and van Doorslaer, E. (2000). Income inequality and health: What does the literature tell us? *Annual Review of Public Health* **78**, 19–29.

Income Gap across Physician Specialties in the USA

G David and H Bergquist, University of Pennsylvania, Philadelphia, PA, USA
S Nicholson, Cornell University, Ithaca, NY, USA

Glossary

Relative value unit A value assigned to each service that a physician performs that reflects the time, intensity of effort, malpractice costs, and practice costs associated with the service.

Introduction

Despite their common medical school training and their shared title of 'physician,' there are many differences between physicians who enter different fields of medical practice. The most obvious difference is the variation in knowledge set and patient population that comes with each specialty. For example, pediatricians take care of children, whereas geriatricians take care of the elderly, and the types of conditions and diseases treated by these two types of physicians are completely disparate. Even among doctors who treat patients of the same age, there can be vast differences, as seen, for example, between psychiatrists (who often use counseling to treat unseen diseases of the mind, emotion, and personality) and radiation oncologists (who, with a requisite knowledge of physics, use radiation therapy to treat cancerous tumors).

However, the differences across medical fields go beyond scope of practice. With different areas of specialization come a range of patient interactions; whereas gynecologists almost always have face-to-face interactions with their patients, anesthesiologists most frequently see unconscious patients, and pathologists and radiologists rarely (if ever) see their patients at all. Similarly, the practice settings where physicians in different specialties work are quite variable: a family practitioner typically works out of an office, a hospitalist works in hospital medical/surgical units, an intensivist works amidst the machinery of a critical care unit, and a surgeon works in the operating room. Along with these different settings and patients come different schedules and work hours. Although a dermatologist may have typical office hours (Monday through Friday, 9 a.m. to 5 p.m.), a surgeon will typically start much earlier (5 or 6 a.m.) and frequently run late into the evening, and emergency medicine physicians have to work nights and weekends to staff the emergency room 24 h a day, 365 days a year. Another difference across medical fields is the degree of specialization. Whereas a family doctor will treat patients of all ages and genders, and thus must be expected to recognize (and treat) a vast range of conditions and diseases, a neonatologist, cardiac electrophysiologist, or gastroenterologist who specializes in colonoscopy will focus on a relatively specific subset of patients and conditions. Along with varying degrees of specialization come different lengths of training programs. A general internal medicine physician can practice after 3 years of residency training, whereas a pediatric neurosurgeon requires a 7-year neurosurgery residency, followed by a pediatric neurosurgery fellowship of at least 1 year.

Another significant difference between medical specialties, and the focus of this article, is average income. For example, in the US, physicians who practice in the primary care specialties (e.g., general internal medicine, family practice, and pediatrics) earn substantially less than physicians in non-primary care specialties (e.g., dermatology, radiology, and orthopedic surgery), with some higher-income specialty physicians earning more than three times as much as their primary care contemporaries. These income differences also exist in other developed countries. For example, orthopedic surgeons earn twice as much per year as primary care physicians in Australia and the UK, and more than 50% more in Canada, France, and Germany.

To provide a glimpse of the variety in physician specialty income in the US, data from several waves of the annual Physician Compensation and Production Survey between 1995 and 2010 have been used in this article. The survey was conducted by the Medical Group Management Association (MGMA), spanning more than 2300 medical organizations and multiple specialties. Specialty classifications have been used based on Modern Healthcare salary surveys and Sigsbee (2011).

Figure 1 reports physicians' median compensation in 1995 versus 2010 across 18 medical specialties. The dotted line represents the case where the 2010 compensation level equals the inflation-adjusted 1995 compensation level (using the consumer price index). Points below this line represent specialties for which median salary grew at a slower pace relative to inflation. For example, median compensation between 1995 and 2010 for Obstetrics and Gynecology grew 31.4% whereas inflation was 41.6%. Figure 1 highlights both the dispersion in compensation level within period as well as the widening of the gap over time. Even in 1995, the median compensation for anesthesiology, cardiology, radiology, and orthopedic surgery was nearly double the median compensation for family practice, internal medicine, pediatrics, and psychiatry. By 2010, these differences were more pronounced with orthopedic surgeons earning close to three times more than family practitioners. It is interesting to note that some specialties experienced greater growth than others. For example, the fastest growth in median compensation occurred for dermatology and gastroenterology.

Figure 2 tracks average annual compensation for 10 specialties, for which data were available between 1995 and 2009. The income is plotted for five high-compensation specialties (radiology, anesthesiology, cardiology, urology, and oncology) and five low-compensation specialties (hospitalist (added to

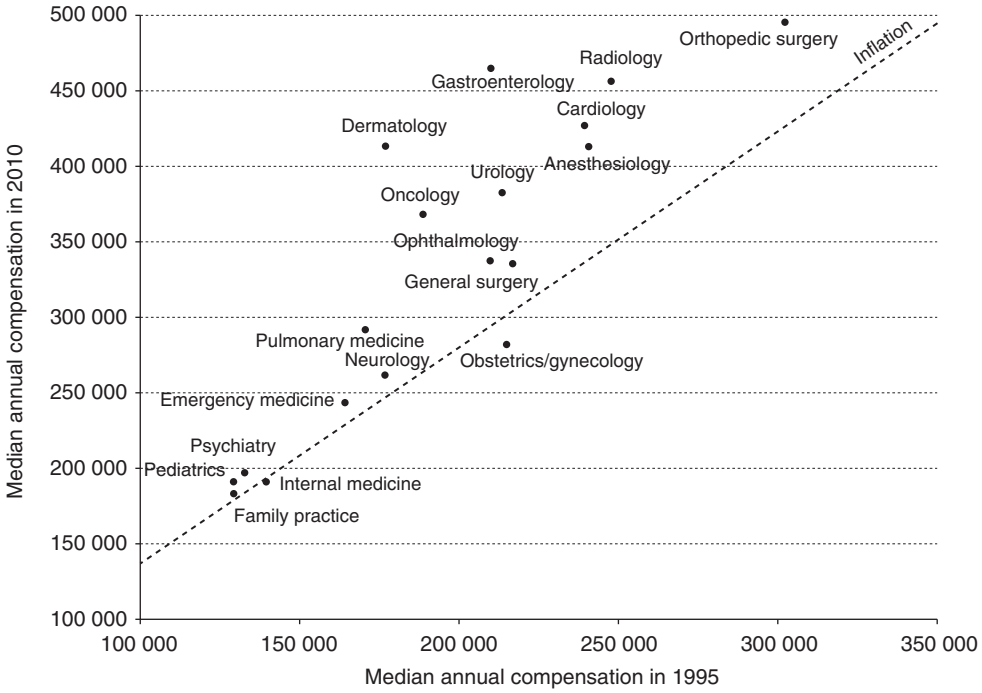


Figure 1 Median physician compensation in 1995 versus 2010 by specialty. Based on data from the 1995 and 2010 MGMA Physician Compensation and Production Surveys.

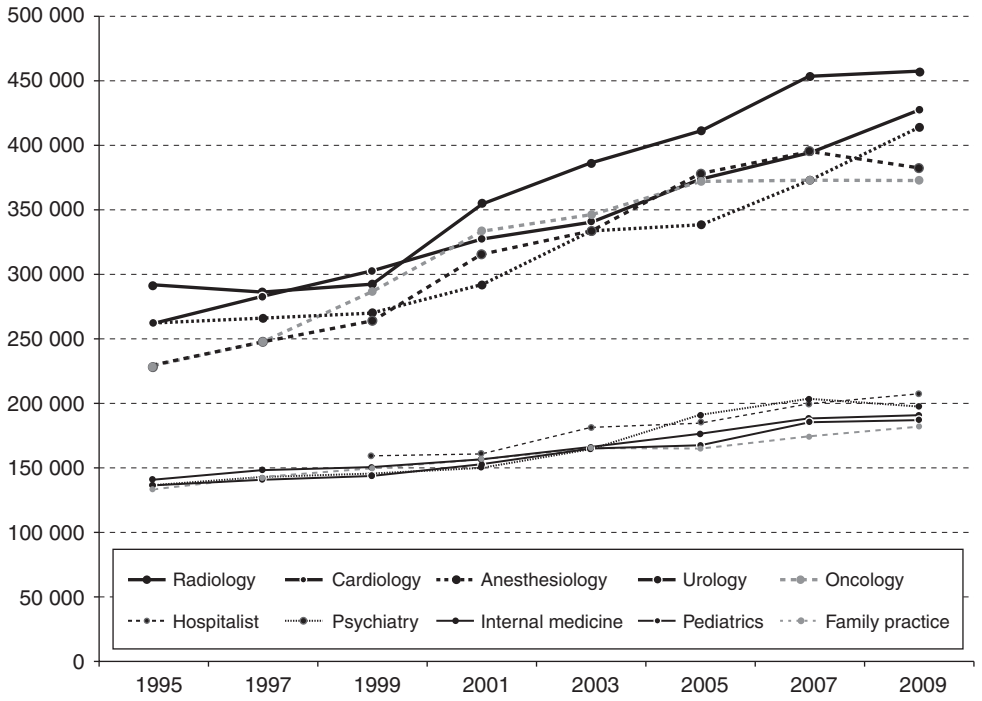


Figure 2 Trends in average annual compensation for selected specialties (1995–2009). Based on data from the 1995, 1997, 1999, 2001, 2003, 2005, 2007, and 2009 MGMA Physician Compensation and Production Surveys.

the survey in 1999), psychiatry, internal medicine, pediatrics, and family practice). Similar to Figure 1, both the difference at baseline (1995) and the difference in growth rates across high- and low-compensation specialties are apparent.

Specialists in most other developed countries receive much higher income than their primary care counterparts, although there are a few exceptions. In 2004, specialists earned at least 50% more than primary care physicians in Canada, Austria,

France, Luxembourg, and the Netherlands (in the US in that year, specialists earned 62% more than primary care physicians, on average) (Fujisawa and Lafortune, 2008). Primary care physicians in Japan earn more than specialists. This is probably because specialists are employed by hospitals, whereas primary care physicians tend to own their own practices, and physicians in Japan can provide a small number of hospital beds in their practice.

Aside from the potential discontent that this income gap may breed among different physicians, from a social or government policy perspective, this difference in expected income may have undesirable consequences. Research has shown that the US and Canadian medical students in general, when selecting their career specialty, are responsive to income differences. However, research has also shown that increased specialization leads to higher medical expenditures, without necessarily improving quality, mortality, satisfaction, or other important metrics of medical care. Thus, if specialists continue to have higher expected incomes than generalists or primary care physicians, the country's healthcare system may continue down a path of higher cost, lower value care, and a shortage of generalist physicians.

Although the ongoing existence of this salary disparity is undisputed, the reasons for its existence are subject to continued investigation and debate. For many in the medical community, the explanation for this income gap is simple: the major government and private payers of medical services (e.g., National Health Service in the UK, Medicare in Canada, US Medicare, US Medicaid, and private health insurers) have decided to reimburse specialists at higher rates than primary care doctors. In 1992, Medicare adopted the resource-based relative value scale system for the US, which is designed to set reimbursement rates according to the relative value of and resource requirements of different services, and most major insurers have followed suit. This payment system generally reimburses specialists at much higher rates, even for patient visits of comparable duration (Bodenheimber *et al.*, 2007). The relative value units used by the Centers for Medicare and Medicaid Services in the US to label each type of physician visit and procedure are updated periodically, under the recommendation of the Relative Value Scale Update Committee, which is dominated by specialty (i.e., nonprimary care) physicians, who make up 85% of its voting members.

Although it may be true that reimbursement rates for physicians are set exogenously by payers, it is not clear how such price setting would establish a persistent income gap across physician specialties. As mentioned two paragraphs above, studies have shown that expected income has a strong influence on specialty physician labor supply. If this is true and if there were no major differences between medical specialties aside from their expected income, one would expect all medical students to choose specialties with higher expected incomes (e.g., radiology, orthopedic surgery, anesthesiology, and cardiology), leading to a massive shortage of physicians in specialties with lower expected incomes (e.g., general internal medicine, family practice, psychiatry, and pediatrics). A large enough shortage would increase the demand for physicians in primary care and other low-paying specialties, which would eventually lead health insurers to offer increased income to attract medical students to these specialties, ultimately equalizing the expected incomes of different physician specialties. However, in reality, although there is a perceived shortage of primary care physicians, there is an enduring disparity in physician incomes across specialties, indicating that there are underlying causes or forces preventing the equilibration of expected income for physicians.

Two potentially different phenomena are required for this physician income gap to emerge and persist. First, there must be factors that cause medical students to sort into different medical fields, despite the difference in expected income. Second, there must be a reason (or reasons) that the income gap across specialties is allowed to continue and expand. That is, there must be some factors preventing prices from clearing the physician labor market. This article considers these two elements as different possible explanations and their supporting evidence for the observed income gap are evaluated.

Potential Explanations for the Income Gap

As evident from Figure 1, the income gap between different physician specialties has been persistent and has widened over time. However, the reason that this gap exists and persists is less obvious and is subject to continued debate and research. In reviewing the different hypotheses that attempt to explain the persistent income gap, this article will consider both mechanisms by which physicians sort themselves into different specialties independent of income (i.e., the reason the income gap is established) and also mechanisms that limit physicians' ability to concentrate in the highest paying specialties (i.e., the reason the income gap is maintained). The conclusion is that although individual preferences can help explain why incomes differ across specialties to begin with, the most important reason why these differences persist and have grown over time is that barriers prevent medical students from entering high-income specialties. Because physicians who are already in a specialty largely control the number of students who are allowed to enter that specialty, this raises the possibility that certain specialties are behaving as a cartel.

Preferences and Compensating Differentials

Aside from expected income, there are many other differences across medical specialties, including scope of practice, level of patient interaction, and regularity of working hours. Given this variety across medical fields, one might expect that a student's choice of medical specialty will be multifactorial and will depend on more than just expected income. The idea that job features and amenities may compensate for lower income may provide a possible explanation behind the physician salary gap and its persistence. If preferences for specialty choice are motivated by nonincome-related factors because individuals place less importance on expected income and more importance on, for example, the scientific content of their field, then some physicians will knowingly and purposefully choose lower-paying specialties, all in accordance with their individual valuation of other nonmonetary dimensions of the specialty.

In support of this hypothesis, many researchers have examined the influence of personality and personal preference on medical students' choice of career specialty. Many studies have found associations between different personality types and specific medical fields. As an example, research has shown that the traits of 'rule-consciousness' and 'tough-mindedness' predicted differences between physicians in general surgery and family practitioners (Borges, 2001). Furthermore, other studies have found that predictability of working hours and lifestyle are key factors driving medical students' choice of career specialty. For example, although pediatricians are paid less than trauma surgeons on average, it is often the difference in schedule structure (where general pediatricians work standard business hours that may extend into the evening as patient visits run over, whereas trauma surgeons work nights and weekends, but only in shifts with a well-defined beginning and end) that motivates students' selection of one specialty over the other. Similarly, one might expect that a student who is passionate about treating children would choose to be a pediatrician, even though (as the student knows) most pediatric specialists are paid notably less than adult specialists covering the same area of expertise (e.g., cardiology, neurology, intensive care, etc.).

Other factors that vary across specialties and individual preferences include specialties' level of intellectual content, number of challenging diagnostic problems, availability of research opportunities, likelihood (and severity) of malpractice litigation, and prestige relative to other specialties. Indeed, research has shown each of these different factors can play important roles in shaping medical students' choice of career specialty. It is important to note that not only is it possible for students to have different preferences for any given career attribute (e.g., one student may prefer a field that is diagnostically challenging, whereas another student may prefer a field that is diagnostically simpler) but it is also likely that students assign different degrees of importance to different attributes (e.g., predictable working hours is very important to some students, whereas other students' top priority is working in a specialty with more research opportunities). When considering all these different factors and how they might influence physician specialty selection, it is not surprising that differences in specialty income may be allowed to exist and persist. Certain specialties must have nonmonetary attributes that appeal to a large percentage of medical students, and this appeal has persisted (and perhaps grown) over time.

Ability Differences

In addition to preferences, another individual characteristic that varies between people and might explain the income disparity is ability, although it does not appear to play a strong role in explaining income differences for physicians. Although the admission process for medical school is rigorous and extremely selective, medical students still vary in ability. Here, ability may refer to IQ, memory (e.g., learning speed, and capacity), physical skills (e.g., dexterity and endurance), or personality type or temperament. Realizing the diversity that exists across medical specialties, one might hypothesize that some specialties require abilities that are relatively rare among medical students whereas other specialties require more common abilities. This article will assume that specialties that require greater ability (or more uncommon ability) are intrinsically more difficult or challenging, as it is not clear otherwise why they would demand greater skill or talent. As in many other professions, workers with the highest abilities have rare attributes, skills, or talents, which would demand a salary premium over other workers, so the difficult or challenging specialties will offer higher salaries to attract high-ability workers.

Another type of ability to consider is a physician's capacity for dealing with risk. Owing to their patient population and scope of practice, some specialties require physicians to act in higher risk situations. For example, most would agree that the likelihood and severity of patient harm from a physician mistake is greater in the fields of neurosurgery, anesthesiology, or interventional cardiology than in family practice or sports medicine. Moreover, specialists often accept more responsibility and thus risk compared to generalists. For example, it is commonplace for a family practitioner or general pediatrician to refer the patient to a specialist or an expert. The specialist, on the contrary, often represents the 'end of the line,' so the ultimate task of diagnosing and treating the patient often falls on the specialist. In this position, the specialist accepts more responsibility and risk (if the diagnosis or treatment is incorrect), so one might expect this additional responsibility to justify a salary premium.

As with the case of personal preferences, differences in individual ability can potentially explain both why students sort themselves into different medical specialties and why the income gap between specialties is able to endure. If the more challenging specialties (that require greater ability) pay more and all medical students prefer greater income, all students will prefer to work in the more difficult fields. However, if entrance into a specialty, which is typically dictated by acceptance to a residency program, is determined according to ability, then only the highly skilled or talented students will be able to work in the more difficult specialties. Given the demanding application and interview process that is required for admission to residency programs, which consider test scores, clinical evaluations, and letters of recommendation, there is clearly a process that prevents low-ability students from achieving positions in high-ability specialties, thus allowing persistence of the income disparity.

In the literature, theoretical economic work has supported this hypothesis that large income differences may reflect even relatively small differences in ability. Furthermore, research has shown that some medical specialists score higher than others in terms of, among other traits, intelligence and self-sufficiency. However, other works have found little evidence that differences in ability are responsible for the large gap in physician income (Bhattacharya, 2005).

The National Resident Matching Program and the Association of American Medical Colleges collect data on the medical students who match into different residency programs in the US, and these data indicate differences in ability between the students entering different specialties (NRMP and AAMC, 2009). For example, there are significant differences in scores on Step 1 of the US Medical Licensing Examination between

some higher-paying specialties (plastic surgery, neurosurgery, dermatology, and radiology) and lower-paying specialties (family practice, pediatrics, psychiatry, and physical medicine). Of course, these data do not indicate a causal relationship, but nevertheless the association between higher-paying specialties and students with higher test scores (a measure of ability) is noteworthy.

This being said, it is not clear whether or not the specialties with higher incomes are actually more challenging or demanding of greater physician ability than specialties with lower incomes. For example, is ophthalmology or dermatology more challenging than emergency medicine or neurology? Without any evidence of this, it is not clear that differences in medical students' or physicians' abilities are the reason that different specialties have different expected incomes. As long as students all want to maximize income and all residency programs want to attract students with the highest level of ability, residency programs for high-paying specialties will be able to select the most skilled and talented students, regardless of the reason(s) that different specialties have different expected incomes.

Workload and Effort

A straightforward explanation for why physicians in some specialties have higher salaries is that their specialties may require greater labor input. Taking this logic to the extreme, it may be that all physicians are paid approximately the same hourly wage, but those who work longer hours accumulate a greater total income. Furthermore, the effect of hours worked on income might be even greater if the marginal value of time increases as the number of hours worked increases. That is, comparing a physician who works 60 h per week to a physician who works 40 h per week, one might expect that the wage for the marginal hour should be higher for the former, because leisure time is more valuable to someone who spends more time working. This hypothesis of increased income with increased workload, if true, would provide a mechanism for both physician sorting into different specialties and the maintenance of the income gap, based on individual preferences for income and leisure. Given equivalent hourly rates, those physicians who choose to work longer hours (i.e., choose specialties that demand more time) are knowingly sacrificing leisure to receive higher pay, whereas those who place a higher value on leisure will willingly forego higher income.

Although this concept is intuitively plausible, it is not supported by evidence. In fact, it would be more likely that labor input is responsive to the hourly wage than vice versa. Put differently, exogenous variation in hourly wage across specialties would induce physicians with identical labor-versus-leisure preferences to vary in their labor supply. Thus, without the ability to verify the authors' assumptions of physician preferences, it cannot be determined if the income disparity results from variation in the value that different physicians place on their leisure (even with similar hourly wages), or from variation in hourly rates, which translates mechanically into an income gap even when physicians exhibit similar labor-versus-leisure preferences. To this end, research has shown that the number of physicians choosing a specialty is more responsive to changes in the number of relative hours worked than to changes in relative income earned. However, regardless of the assumptions, studies have found that only a small fraction of the income gap can be attributed to differences in the number of hours worked, indicating that hourly rates are almost certainly not equivalent across specialties (Bhattacharya, 2005).

Length of Training

Another hypothesis commonly believed to explain the income disparity across physician specialties is the difference in required training. The reasoning behind this belief is frequently offered as an explanation for why physicians, on average, compared to other professionals are among the best-paid members of most societies. Looking within medicine and comparing different types of physicians, specialists undergo more training than generalists, and additional years in residency and fellowship create potentially important opportunity costs for specialists (in terms of lost time and wages). Thus, the hypothesis states that specialists are paid more to compensate for the additional costs of their extended training. Without this increase in expected income, physicians would not be willing to incur the additional costs of training required for specialization. Therefore, if medical students have different time preferences for income, with some unwilling to take on short-term costs of training for the long-term gain of increased future income, they will sort themselves into specialties with different expected incomes.

In support of this hypothesis, research has indicated that medical students tend to prefer specialties with shorter residency training programs. However, other studies have shown that a relatively small portion of the income gap between different physician specialties can be explained by differences in training time; that is, students' choice of specialty is mostly unresponsive to expected length of training (Bhattacharya, 2005). Furthermore, although some specialties (e.g., gastroenterology) provide a favorable return to specialization, other specialties (e.g., rheumatology) actually provide an unfavorable return to specialization (e.g., compared to staying in general internal medicine). As another example, the postgraduate training requirements for geriatricians and dermatologists are typically equivalent, but the expected income of geriatricians is often less than half that of dermatologists.

Aside from direct costs of longer training, other short-term financial considerations may motivate students to choose a specialty with a shorter training requirement, even if it means lower expected long-term income. The majority of graduating medical students has extensive debt, mostly from accumulated loans for undergraduate and graduate education, and although such loans can be deferred while students are in school, when the students graduate and enter residency programs, they must begin repaying these loans. Given the amount of debt that some students have (more than US$200 000), the size of loan repayments can be substantial, causing significant financial stress during residency. Other than educational loans, some students might expect other

large financial burdens during residency, for example, supporting a new or growing family. Furthermore, although facing mounting financial demands, young physicians may have decreased access to private financial markets, as they are no longer eligible for educational loans. Any combination of these reasons can make residency training a particularly stressful financial time for young doctors, and this predicted stress may motivate students to choose specialties that minimize the length of residency, allowing them to become a practicing physician sooner. (Even the lowest paying jobs for practicing doctors pay at least three to four times more than a resident's salary.) However, most of the specialties with shorter residency programs have lower lifetime expected incomes than the specialties with longer residency training, so students who choose shorter residencies are typically choosing lower long-term expected salaries. Thus, differences in specialty length of training combined with debt and short-term financial considerations may explain why some students choose lower-income specialties and why the physician specialty income gap continues to exist. Furthermore, other work has shown that not only just the amount of student debt but also the type of debt (e.g., subsidized vs. unsubsidized loans) is a significant variable in students' choice of specialty.

Even for those medical students who are not carrying student debt or expecting increased financial stress during residency, there may be other motivations to choose a specialty with shorter residency training. For example, students might have very different future discount rates. Whether for financial reasons or otherwise, one could imagine students placing greater importance or value on the upcoming 5–10 years than on the more distant future; that is, a student might drastically discount all considerations that are more than 5–10 years away. In this time horizon, a medical specialty with a shorter residency program might appear more ideal than other specialties. For example, over the first 10 years of postgraduation, a student who enters family practice (3 years of residency) can expect to earn more than a student who pursues a career in surgery (more than 7 years of residency) because the income of the family practitioner will be much higher than that of a surgery resident. Thus, heterogeneity in time preferences and subjective discount rates may help explain why income-maximizing students choose specialties with very different lifetime expected incomes.

Although most of the literature has explored variation in the length of formal training across specialties as a potential explanation for the income gap, no attention has been given to aspects of on-the-job training across specialties. Surgical and procedural specialists have to keep up with changing equipment, technology, and procedures and are required to invest considerable amount of time in order to do so. Therefore, an argument could be made that physicians in more dynamic fields require a premium for keeping up with the latest technologies and procedures. Nevertheless, it could be argued that generalists and nonprocedural specialists have just as many journals to read and new guidelines to keep up with. Moreover, most states necessitate continuing medical education (CME), but do not make distinctions across specialties; hence, specialists do not have to perform more CME than generalists.

Variation in Training Focus across Medical Schools

Despite the theories and supporting studies that connect student debt to medical specialty choice, other researchers have shown that medical students entering primary care fields do not have significantly more (or less) debt than students entering nonprimary care fields. A medical student's choice of career specialty is a complicated, multifactorial decision. Not only is that decision influenced by the interplay of personal preferences and specialty characteristics but also the type of environment in which medical students are educated may also shape their choice of specialty. Medical schools can be viewed as producers of medical students, and although there is some standardization across medical schools, there can be substantial variation in the inputs that schools supply to this production process. Inputs in the medical student education process for any given medical specialty include, among other factors, preclinical curriculum, clinical rotation requirements, availability of mentors, and the presence of a residency training program. Through the variable use of these different inputs, some schools will produce more students who choose to pursue careers in primary care or other specialties. For example, the percentage of US graduating students who enter family practice varies (over a 10 year mean) from 1.7% to 34.9% depending on their medical school. Although the income gap across medical specialties may be established for exogenous reasons, the role played by medical schools may help explain how the income disparity is maintained. When students first enter medical school, they are less likely to exhibit strong preferences toward a given medical specialty, because they may have limited understanding of the different fields and little awareness of income differentials across fields. Therefore, students are likely to select their medical school regardless of the specialty mix it typically produces, and hence to be sorted into specialties with different income profiles.

Research has verified that medical schools do influence students' choices of career specialties. For example, the differential production of generalists versus specialists has been examined by studies that characterize the population of students who choose to enter family medicine residences. Supporting the hypothesis that medical schools may have different 'production functions' for medical students, research has shown that students from publicly funded schools are more likely to choose family medicine than students from privately funded medical schools. In some years, although some medical schools (including Johns Hopkins University, New York University, and Washington University in St. Louis) had no graduates who pursued careers in family medicine, at other schools (including University of Arkansas, the Medical College of Georgia, and University of Minnesota) greater than 22% of graduates entered family medicine residencies.

Institutional Barriers to Entry

Bhattacharya (2005) examined the role of different factors in explaining the disparity in physician income across specialties, and finds that only approximately half of the increase in expected income from specialization can be attributed to differences in hours of work, length of training, and skill or ability. Although individual preferences and their implications

for career path selection may explain some of the income disparity, barriers that prevent medical students from entering higher-income specialties offer another plausible explanation.

The Accreditation Council for Graduate Medical Education (ACGME) is the organization responsible for accrediting residency programs in the US, and thus it determines how many residency positions are available for training new physicians. Regulation of medical education and training is common in most developed countries. For example, the Medical Council of India, the Korean Institute of Medical Education, the General Medical Council (UK), the Netherlands–Flemish Accreditation Organization, and the Japan University Accreditation Organization approve curricula and accredit medical schools in their respective countries.

Broadly speaking, the restricted number of residency positions is a substantial factor (if not the most important factor) limiting the number of physicians who can enter professional practice, but it also plays a role in determining the number of physicians in different specialties. The ACGME oversees and sets policies for Residency Review Committees (RRCs), which are specialty specific and tasked with reviewing and accrediting hospital residency programs in their target specialties. In this position, an RRC essentially has complete control over the flow of physicians into a specialty because medical students who attend programs that are not certified by the ACGME are not eligible to take the licensing exam, and thus not able to practice in the US (Nicholson, 2003). Therefore, incumbents in a specialty determine how many new physicians may be trained in that specialty, which in turn will influence future earnings in that specialty.

Thus, regardless of the reasons why expected incomes vary across medical fields, the constrained number of available residency positions for each specialty prevents all medical students from entering higher-paying specialties, thus allowing the income gap to be sustained. High-income specialties in the US tend to have more residents who are trying to enter than there are positions available. For example, between 1991 and 2009, the ratio of the number of medical students who were trying to enter a specialty to the available number of first-year residency positions exceeded 1.40 in orthopedic surgery in all but 1 year, and between 1997 and 2009 the ratio exceeded 1.60 in dermatology in all but 1 year. Barriers to entry exist in other countries as well. Medical school graduates in Greece often wait several years for a nonprimary care residency position opening.

Concluding Remarks

In developed countries, specialists, or nonprimary care physicians, earn considerably more than primary care physicians, and these income differences have persisted over time. This article reviews and assesses the support for different hypotheses regarding nonmonetary reasons why physicians may sort themselves into different specialties (i.e., the reason an income gap is established), and also hypotheses that help explain why the income gap persists.

Specialists can earn more than primary care physicians if the former medical fields require scarce abilities, have unattractive nonmonetary attributes (e.g., undesirable working environment), and require relatively long training. This will be particularly true if medical students have different time preferences and debt levels. If these factors persist over time, such as medical students' preferences for the nonmonetary attributes of primary care, then the higher income of specialists relative to primary care physicians can also persist. The empirical support is strongest for the hypothesis that occupational attributes other than expected income do matter when medical students choose a specialty, and therefore do help explain income differences across specialties.

However, Bhattacharya (2005) finds that student preferences explain approximately one-half of the specialty premium, with entry barriers to high-income specialties possibly explaining the balance. Thus, regardless of the reasons why expected incomes vary across medical fields to begin with, constraints on the number of available residency positions in high-income specialties prevent medical students from entering high-income specialties and driving down specialist income, and thus allow the income gap to persist. Because physicians who are already practicing in a specialty largely control the flow of new physicians into that specialty, this raises the question of whether certain high-income specialties are behaving as cartels. Making it easier for medical students to enter high-income specialties would reduce income differences across specialties.

See also: Health Labor Markets in Developing Countries. Occupational Licensing in Health Care. Primary Care, Gatekeeping, and Incentives. Specialists

References

Bhattacharya, J. (2005). Specialty selection and lifetime returns to specialization within medicine. *Journal of Human Resources* **40**(1), 115–143.

Bodenheimber, T., Berenson, R. A. and Rudolf, P. (2007). The primary care – specialty income gap: Why it matters. *Annals of Internal Medicine* **146**(4), 301–306.

Borges, N. J. (2001). Personality and medical specialty choice: Technique orientation versus people orientation. *Journal of Vocational Behavior* **58**(1), 22–35.

Fujisawa, R. and Lafortune, G. (2008). The remuneration of general practitioners and specialists in 14 OECD countries: What are the factors influencing variations across countries? *OECD Health Working Papers No. 41*, Paris, France: OECD.

National Resident Matching Program (NRMP) and Association of American Medical Colleges (AAMC) (2009). Charting outcomes in the match: Characteristics of applicants who matched to their preferred specialty in the 2009 main residency match. Available at: http://www.nrmp.org/data/chartingoutcomes2009v3.pdf (accessed 16.05.11).

Nicholson, S. (2003). Barriers to entering medical specialties. *NBER Working Paper*, #9649. Cambridge, MA: National Bureau of Economic Research.

Sigsbee, B. (2011). The income gap: Specialties vs primary care or procedural vs nonprocedural specialties? *Neurology* **76**(10), 923–926.

Relevant Websites

http://www.nrmp.org/
National Resident Matching Program.

http://www.oecd.org/health/health-systems/oecdhealthdata2012.htm
OECD.

Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis

M Asaria, R Cookson, and S Griffin, University of York, York, UK

Introduction

Health sector programs often have important policy objectives relating to the reduction of unfair health inequality, as well as the improvement of total population health. Health inequality reduction objectives are particularly common in public health decision-making, for example, in relation to screening and vaccination programs, and are sometimes also relevant to decisions regarding the introduction and delivery of new medicines, surgical procedures, and other health technologies.

Standard economic evaluation methods, however, focus solely on identifying cost-effective interventions to maximize health. The distributional cost-effectiveness analysis (DCEA) framework described in this article builds on standard cost-effectiveness methods by extending them to incorporate distributional impacts on health. Like the standard cost-effectiveness analysis (CEA) framework, this framework focuses exclusively on health benefits and opportunity costs falling on the health sector budget. It focuses on the health impacts of health sector programs, assuming that there are no important impacts on the distribution of income, education, or other determinants of health outside the health sector. It is therefore not suitable for evaluating cross-government public health program with important nonhealth benefits and opportunity costs falling outside the health sector budget.

The key steps in the DCEA framework outlined below are: estimating the baseline health distribution in the general population; modeling changes to this baseline distribution due to the health interventions being compared, and using this to estimate the mean change in health due to each intervention; adjusting the resulting modeled distributions for alternative social value judgments regarding fair and unfair sources of health variation; using these adjusted distributions to estimate the change in the level of unfair inequality due to each intervention; and finally combining the mean level of health and level of unfair inequality associated with each intervention by using an appropriately specified social welfare function to rank interventions, and decide as to which best fulfills the dual objectives of maximizing health and minimizing unfair health inequality.

Estimating the Baseline Health Distribution

The first step in DCEA is describing the baseline distribution of health in the general population, taking into account variation in both quantity and quality of life among different subgroups in the population as defined by relevant population characteristics. A natural health metric to use in this context is quality adjusted life expectancy (QALE) at birth, though other suitable health metrics can be used – such as disability adjusted life expectancy at birth or age-specific QALE – so long as they are on an interpersonally comparable ratio scale suitable for use within CEA. Mortality rates and morbidity adjustments differentiated by relevant population characteristics are required to estimate this distribution. Figure 1 shows the estimated baseline population health distribution in the UK in the year 2010 as measured in QALE at birth, taking into account differential mortality and morbidity by age, gender, and area level deprivation.

Estimating the Distribution of Health Changes Due to the Intervention

The next step in DCEA is to estimate the net impact of one or more interventions on the baseline distribution of health within the general population. This requires not only 'effectiveness' information on the direct health benefits of the health intervention on individuals receiving the intervention, but also information on the indirect health impacts of the intervention – in particular, the health opportunity costs due to displaced expenditure within the health sector budget – on both recipients and nonrecipients of the intervention. There are a number of factors that may vary by relevant population subgroup characteristics, which must be incorporated into the model to estimate correctly the impact of a health intervention on the population health distribution, including:

- Prevalence and incidence of the health condition, which will also help to analyze the differing maximum potential impact that the intervention could have on each population subgroup.
- Uptake of the intervention, which for more complex interventions may include differential uptake by subgroup at multiple stages of the patient pathway.
- Effectiveness of the intervention.
- Mortality and morbidity due to condition and comorbidities.
- Opportunity cost.

Under the assumption of a fixed overall health budget, any additional costs associated with the intervention will result in some displacement of activity. The distribution of the health

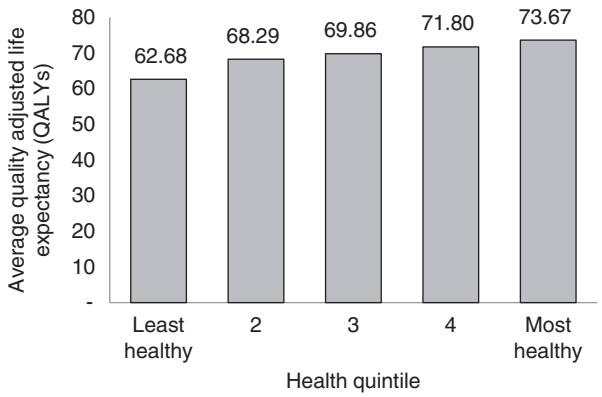


Figure 1 Baseline health distribution.

 doi:10.1016/B978-0-12-375678-7.01415-2

opportunity costs due to this displacement on both recipients and nonrecipients of the intervention in the population needs to be characterized by subgroup to give the overall distribution of health losses due to the intervention. A simple and convenient assumption is that the distribution is neutral – i.e., all subgroups share equally in the health opportunity cost of displaced health sector activity. However, this assumption may not be accurate, and ideally, one would want evidence on the likely distribution of health opportunity cost.

Once the distribution of health gains and health opportunity costs of an intervention for each population subgroup have been estimated, these distributions can be combined to produce a distribution of net health changes by subgroup and applied to the baseline health distribution to give an estimate of the impact of the intervention on the overall health distribution.

Measuring the Level of Inequality in the Estimated Health Distributions

The overall health distributions associated with each intervention can be assessed in terms of the level of health inequality they comprise. There are a number of commonly used indices for measuring inequality in the distribution of income, which can also be applied to health when measured on a ratio scale such as quality adjusted life expectancy. These indices are based on a common set of fundamental principles:

- Principle of transfers: The most universally recognized concept of what is meant by inequality in a distribution is the weak principle of transfers, also known as the Pigou–Dalton transfer principle. It broadly states that the transfer of health from a more healthy to a less healthy person reduces inequality so long as the amount of health transferred is less than the difference in health between them. It is of course not possible directly to transfer current health from one person to another (except in rare cases such as organ transplant); but one can think of indirect transfers in terms of gains or losses in people's expected future lifetime experience of health. This concept of inequality is useful for comparing alternative distributions of a fixed total pot of health. The next two concepts discuss how inequality measures react to a change in the size of the pot.
- Scale independence: Scale independence focuses attention on concern for relative inequality between individuals – their 'fair shares' of the total pot – rather than the size or scale of absolute differences between individuals. It states that any equal proportional change in each individual's level of health should not change the measure of health inequality. Although this is relatively uncontroversial when applied to changes in the scale used to measure health, it is harder to justify when looking at real differences in health. A commonly used tool to describe relative inequality in a distribution is the Lorenz curve, this plots the cumulative proportion of individuals ordered by their health on the x axis against their cumulative share of total health on the y axis. The difference between the Lorenz curve and the 45° line of equality represents the level of relative inequality in the distribution. Common relative inequality measures such as the Gini coefficient are based on measuring this difference. There are also relative inequality measures such as the Atkinson index that allow for the specification of a level of inequality aversion to adjust the sensitivity of the measure to inequalities in different parts of the distribution, and which also allow explicit formulation of tradeoffs with sum total health within a social welfare function framework.
- Translation independence: Translation independence focuses on concern for absolute inequality between individuals. It states that any equal absolute change in each individual's level of health should not change the measure of health inequality. Simple measures such as absolute gaps and slope indices are widely used to quantify absolute inequality. There are also absolute inequality measures such as the Kolm index – an absolute inequality equivalent to the Atkinson index – which allow the specification of an absolute inequality aversion parameter and the modeling of explicit tradeoffs with sum total health.

Although all reasonable inequality measures satisfy the principle of transfers, a measure cannot fully satisfy both scale independence and translation independence. For example, if everyone in a health distribution gains 25 years in life span the absolute gap between any two individuals remains the same, a relative gap between two individuals living 60 and 50 years respectively of 20%, however, declines into a relative gap of only 13%, with these individuals living 85 and 75 years after the gain in life span. When selecting inequality indices to rank distributions, it is important to recognize these distinctions and identify those that most closely represent the concept of inequality of relevance in the context of the decision being evaluated.

Adjusting for Social Value Judgments Regarding Fair and Unfair Sources of Inequality

The purpose of DCEA is to identify the health intervention that results in the best improvement in both average health and unfair health inequality in the population. The distributions of health estimated thus far represent all variation in health in the population. However, some variation in health may be deemed 'fair' or, at least 'not unfair,' perhaps because it is due to individual choice or unavoidable bad luck. The health distributions should therefore be adjusted to include only any health variation that is deemed 'unfair' before measuring the level of inequality. The DCEA framework allows multiple sources of unfair health inequality – for example, by income, education, ethnicity, geography, and other factors – to be analyzed in the same model. If decisionmakers are interested in one particular source of unfair health inequality, this can also be analyzed separately, or by decomposing the influence of this factor on overall unfair inequality. To make these adjustments for unfair sources of health variation, the association between relevant population characteristics and the estimated health distributions must be modeled. Social value judgments then need to be made regarding whether or not health variation associated with each of the population

characteristics is deemed fair. The modeled associations combined with these social value judgments are used to isolate unfair variation in the distribution, using either the methods of direct or indirect standardization. Inequality measures can then be used to assess the level of unfair inequality in the estimated health distributions associated with each health intervention and hence to rank health interventions by their impact on minimizing this inequality.

Social Welfare Functions and Distributional Dominance

Once both the mean level of health and the fairness adjusted distribution of health associated with each of the interventions have been estimated, social welfare functions (SWF) can be used to compare interventions. Several properties are considered useful when constructing a SWF. In describing these properties, one can use the terminology h_{iA} to represent the health of individual i in health distribution A, U_i to represent an individual utility function for individual i, and W to represent social welfare:

- Individualistic: This means the SWF is a function of the individual utilities, i.e., the SWF has the form: $W=W(U_1, U_2..., U_n)$.
- Nondecreasing: This states that if every individual has at least as good health in distribution A as in distribution B, then overall distribution A is at least as good as distribution B.
- Additive: This means that the social welfare function can be written as a sum of the individual utility functions, i.e., the SWF has the form: $W(h_1, h_2,..., h_n)=U_1(h_1)+U_2(h_2)+...+U_n(h_n)$.
- Symmetric: This means that the SWF treats individual utilities anonymously, i.e., the SWF has the form: $W=W(U_1, U_2,..., U_n)=W(U_2, U_1,..., U_n)=...=W(U_n, U_2,..., U_1)$.
- Concave: This means that when evaluating changes to social welfare lower weight is applied to increases in health to those with higher health than to those with lower health, where the welfare weight is defined as: $U'(h_i)=dU(h_i)/dh_i$.

These properties can be used to derive rules to help determine which of two health distributions are preferable. By using these dominance rules, the exact nature of the SWF need not be specified but can instead be described by broad characteristics that encompass whole classes of SWFs, under any of which the welfare rankings of particular interventions would be the same. The following rules are listed in order from least restrictive to most restrictive that allow a partial ordering of health distributions:

- Rule 1 – Pareto Dominance: For any individualistic, increasing and additive SWF, if $h_{iA} \geq h_{iB}$ for all i and $h_{iA} > h_{iB}$ for at least one i, then distribution A is preferred to distribution B, where subscript i represents the same individual in each distribution.
- Rule 2 – Reranked Pareto Dominance: If additionally, the SWF is also symmetric, then the same condition applies, only that now subscript i represents the individual with equivalent health ranking in each distribution rather than necessarily the same individual in both distributions.
- Rule 3a – Atkinson's Theorem: If additionally, the SWF is strictly concave and distributions A and B have equal mean health, then distribution A is preferred to distribution B if, and only if, the Lorenz curve for distribution A lies wholly inside the Lorenz curve for distribution B.
- Rule 3b – Shorrocks' Theorem: if Lorenz curves cross and the mean health in distribution A is greater than that in distribution B, then distribution A is preferred to distribution B if, and only if, the generalized Lorenz curve for distribution A lies wholly inside the generalized Lorenz curve for distribution B, where the generalized Lorenz curve is derived by multiplying the Lorenz curve for the distribution by the mean of the distribution.

These dominance rules may be used to rank the estimated distributions associated with the health interventions being compared and hence to rank the interventions in terms of social welfare. These rules do not, however, allow for trading off between health inequality and overall health and hence will only provide a partial ranking of interventions when rankings on these two objectives do not coincide.

Social Welfare Indices

Where interventions cannot be ranked based on distributional dominance rules, the SWF needs to be fully specified by defining the nature of the inequality aversion that it will embody to create social welfare indices. The principle underlying the interpretation of these indices is that if health is distributed unequally then, given an aversion to inequality, more overall health would be required to produce the same level of social welfare than if health were distributed equally. Social welfare is represented in these measures using the concept of 'equally distributed equivalent' health: the common level of health in a hypothetical equal distribution of health that results in the same level of social welfare as the actual unequal distribution of health. Two common alternatives specifications for the nature of inequality aversion expressed in social welfare indices are constant relative and constant absolute levels of inequality aversion, yielding the Atkinson and Kolm indices of social welfare respectively:

- Constant relative inequality aversion: This means that a constant proportionate change in health results in a constant proportionate change in welfare weight, i.e., function U(.) takes the form:

$$U(h_i) = \frac{{h_i}^{1-\varepsilon}}{1-\varepsilon}, \varepsilon \neq 1$$

$$U(h_i) = \ln h_i, \varepsilon = 1$$

Summing across this population gives the Atkinson index of social welfare:

$$h_{ede} = \left[\frac{1}{n}\sum_{i=1}^{n}[h_i]^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}$$

where the parameter ε, which can take any value from zero to infinity, specifies the level of societal inequality aversion.

The higher the ε, the further the index tilts toward concern for health improvement among less healthy individuals rather than more healthy individuals. A value of zero represents a classic 'utilitarian' view that all that matters is sum total health and not inequality in the distribution of health. Although as the value approaches infinity, the index comes to represent the 'maximin' view that all that matters is the health of the least healthy individual, irrespective of the health of all other individuals. The proportion of mean health that can be sacrificed to achieve equality will increase as the level of inequality aversion rises.

- Constant absolute inequality aversion: This means that a constant absolute change in health results in a constant proportionate change in welfare weight, i.e., function U(.) takes the form:

$$U(h_i) = -\frac{1}{\alpha}e^{-\alpha h_i}$$

Summing across the population, this gives the Kolm leftist index of social welfare:

$$h_{ede} = -\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha h_i}\right)$$

where the parameter α specifies the level of societal inequality aversion, with higher α values making the index more sensitive to changes at the lower end of the health distribution. The value of this index represents the absolute amount by which average health could be reduced to achieve equal health for all. As with the Atkinson index, the amount of mean health that could be sacrificed to achieve an equal distribution rises with the level of inequality aversion.

The ranking of health distributions using social welfare indices will always be consistent with that produced by the distributional dominance rules where these apply. Where distributional dominance does not apply, rankings may be sensitive to the type and level of inequality aversion embodied by the SWF and these should be chosen with care.

Comparing and Ranking Interventions

Having fully specified the SWF, all interventions can be compared and ranked on the combined objectives of maximizing health and minimizing unfair health inequalities in the population. Conclusions on which intervention is best may be sensitive to alternative social value judgments made both in the fairness adjustment process and in the specification of the type and level of inequality aversion. These social value judgments should ideally be made by the appropriate stakeholders through a deliberative decision-making process, and the robustness of conclusions to alternative plausible social value judgments should be explored.

Conclusion

DCEA is a framework for incorporating equity concerns into the standard methods of CEA. A number of social value judgments regarding which inequalities are deemed to be unfair and the nature and strength of inequality aversion need to be made when using the framework to evaluate and rank alternative health interventions. The framework makes these social value judgements explicit and transparent, and lends itself well to checking the sensitivity of conclusions drawn to alternative plausible social value judgments.

There are a number of alternative methods proposed in the literature for including health inequality concerns in economic evaluation. These typically involve either weighting health gains differently for different groups in the population or weighting overall health gains directly against overall changes in heath inequality. Both these types of method can be replicated using the DCEA framework by imposing the relevant restrictions on the fairness adjustment process and on the form and parameters of the social welfare function.

An important emerging source of empirical literature on incorporating health inequality impacts into economic evaluation in low and middle income countries is the 'extended cost-effectiveness analysis' work being developed by Dean Jamieson, Ramanan Laxminarayan, and colleagues as part of the Disease Control Priorities 3 project (www.dcp-3.org). Their approach to distributional analysis is similar in spirit to the approach outlined in this article, although simplifying the analysis by (1) focusing on a single distributional variable (wealth quintile group) rather than analyzing multiple distributional variables, (2) setting aside the issue of opportunity costs falling on the health budget by assuming that the intervention is funded by the tax system, and (3) presenting results as a disaggregated 'dashboard' of costs and consequences by social group rather than using inequality indices and social welfare functions to analyze tradeoffs between improving health and reducing unfair health inequality explicitly. However, their approach takes a broader perspective than standard CEA by incorporating financial risk protection benefits as well as health benefits. It therefore points the way toward the next great methodological challenge in this area: developing methods of 'distributional cost-consequence analysis' and 'distributional cost-benefit analysis' for incorporating health inequality impacts into economic evaluation of cross-government interventions with important nonhealth benefits and opportunity costs.

See also: Dominance and the Measurement of Inequality. Economic Evaluation of Public Health Interventions: Methodological Challenges. Efficiency and Equity in Health: Philosophical Considerations. Ethics and Social Value Judgments in Public Health. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Measuring Equality and Equity in Health and Health Care. Unfair Health Inequality

Further Reading

Adler, M. (2012). *Well-being and fair distribution: Beyond cost-benefit analysis.* New York: Oxford University Press.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory* **2**(3), 244–263.
Cowell, F. (2011). *Measuring inequality*, 3rd ed. Oxford University Press.
Culyer, A. J. and Wagstaff, A. (1993). Equity and equality in health and health care. *Journal of Health Economics* **12**(4), 431–457.
Dolan, P. and Tsuchiya, A. (2009). The social welfare function and individual responsibility: Some theoretical issues and empirical evidence. *Journal of Health Economics* **28**(1), 210–220.
Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics* **28**(1), 73–90.
Kolm, S. C. (1976). Unequal inequalities. I. *Journal of Economic Theory* **12**(3), 416–442.
O'Donnell, O., Van Doorslaer, E., Wagstaff, A. and Lindelow, M. (2008). *Analysing health equity using household survey data: A guide to techniques and their implementation*. Washington, DC: World Bank.
Roemer, J. E. (1998). *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Sen, A. K. (1973). *On economic inequality*. Oxford, UK: Oxford University Press.
Sen, A. K. (2002). Why health equity? *Health Economics* **11**(8), 659–666.
Shorrocks, A. F. (1983). Ranking income distributions. *Economica* **50**(197), 3–17.
Verguet, S., Laxminarayan, R., Jamison, D. (2012). Universal public finance of tuberculosis treatment in India: An extended cost-effectiveness analysis. Disease control priorities in developing countries, 3rd ed. *Working Paper No. 1*. Available at: http://www.dcp-3.org/resources/universal-public-finance-tuberculosis-treatment-india-extended-cost-effectiveness-analysis (accessed 19.06.13).
Wagstaff, A. (1991). QALYs and the equity-efficiency trade-off. *Journal of Health Economics* **10**(1), 21–41.
Williams, A. (1997). Intergenerational equity: An exploration of the 'fair innings' argument. *Health Economics* **6**(2), 117–132.
Williams, A. and Cookson, R. (2000). Equity in health. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, ch. 35, vol. 1, pp. 1863–1910. Amsterdam: Elsevier.
Williams, A. and Cookson, R. A. (2006). Equity-efficiency trade-offs in health technology assessment. *International Journal of Technology Assessment in Health Care* **22**(1), 1–9.

Relevant Websites

http://www.york.ac.uk/che/research/equity-health-care/economic-evaluation-of-equity/
Centre for Health Economics, University of York.
http://www.fao.org/easypol/output/browse_by_training_path.asp?pub_id=303&id_elem=303&id=303&id_cat=303
Food and Agriculture Organisation of the United Nations.
http://en.wikipedia.org/wiki/Atkinson_index
Wikipedia.
http://en.wikipedia.org/wiki/Income_inequality_metrics
Wikipedia.
http://en.wikipedia.org/wiki/Lorenz_curve
Wikipedia.
http://en.wikipedia.org/wiki/Social_welfare_function
Wikipedia.
http://en.wikipedia.org/wiki/Stochastic_dominance
Wikipedia.

Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview

R Cookson and S Griffin, University of York, York, UK
E Nord, Norwegian Institute of Public Health and the University of Oslo, Norway

Glossary

Cost-value analysis A variant of cost-effectiveness analysis in which *Quality-Adjusted Life-Years* are replaced by social values of a topically relevant kind that take explicit account of the severity of the condition for which the intervention is intended.

Discrete choice analysis A procedure used in experimental economics in which subjects choose real or simulated discrete (i.e., 'on' or 'off') options and thereby reveal (or 'state') their preferences over, for example, states of health.

Economic evaluation A general term for the economic evaluation of options.

Equity Equity is not necessarily to be identified with equality or egalitarianism, but relates in general to ethical judgments about the fairness of the distribution of such things as income and wealth, cost and benefit, access to health services, exposure to health-threatening hazards and so on. Although not the same as 'equality', for some people, equity frequently involves the equality of something (such as opportunity, health, access).

Equity weights The relative importance or value attached to different elements in a decision about what is fair. They may be numerical. In matters of vertical equity, the weights would make the desired adjustment to cost or outcome according to the differentiating features of individuals, such as their age or the severity of their illness.

Fair innings The name given to the idea that benefits to individuals who have not yet had a 'fair innings' (in terms of length of life in reasonable health) should receive a higher weight in cost-effectiveness analyses than those to people who have.

Fairness The ethical consideration of differences between people in terms of their health, access to health care, wealth, opportunities and so on. Fairness does not necessarily require equality since some differences may be regarded as fair ones as, for example, when they are deserved.

Horizontal equity Treating equally those who are equal in some morally relevant sense. Commonly met horizontal equity principles include 'equal treatment for equal need' and 'equal treatment for equal deservingness.'

Multi-criteria decision analysis A technique (often abbreviated as MCDA), akin to cost-effectiveness analysis (CEA), for helping decision makers to take decisions. It differs from CEA by explicitly helping decision makers to consider factors beyond standard welfare or health maximization.

Opportunity cost The value of a resource in its most highly valued alternative use. In a world of competitive markets, in which all goods are traded and where there are no market imperfections, opportunity cost is revealed by the prices of resources: The alternative uses forgone cannot be valued higher than these prices or the resources would have gone to such uses.

Person trade-off A method of assigning utilities to health states that works as follows: Subjects are asked a question of the following kind: 'If x people have health state A (described) and y have health B, and if you can only help (cure) one group, which group would you choose?' One of the numbers x or y is then varied until the subject finds the two groups equally deserving of their vote. The ratio x/y gives the 'utility' of state B relative to A.

Public health Similar to population health, drawing on social epidemiology to embrace the widest range of determinants of health in a society; a broader range of technologies for addressing them than is usually encompassed in public health medicine, such as population vaccination, safety at work, health education, and water purification. The wider range includes determinants such as better parenting for childhood development, better housing, even greater equality of income and wealth; and the broader range of institutional pathways and vectors of influence implied by the forgoing, such as schooling and schools, working and workplace.

Social welfare function A function that maps from the levels of utility attained by members of society to the overall level of welfare for society.

Vertical equity Treating unequally those who are unequal in some morally relevant sense. Commonly met vertical equity principles include 'higher contributions from those with greater ability to pay', 'more resource for those with greater need'.

Introduction

This article is a review of methods for incorporating concerns for fairness or equity in economic evaluation of health care and public health programs. By way of background, the next two sections review the role of equity concerns relative to concerns for efficiency and cost-effectiveness in actual health policy on the one hand and in health economics on the other hand. Section Concerns for Equity: Overview gives an overview of a number of different kinds of concerns for equity and highlights the most salient ones. In section Methods for Incorporating Concerns for Fairness into Economic Evaluation, the various methods for incorporation of equity concerns in economic evaluation are explained.

Equity in Health Policy

Health care decision makers are interested in equity in the finance and delivery of health care, and public health decision makers are interested in equity and inequality in health more broadly. The nature and importance of these equity objectives varies between countries, reflecting variation in concerns for fairness between different societies and over time. For example, in the US policy concerns for fairness in health care focus on offering all citizens a decent minimum of health care but, beyond that, tolerating substantial inequalities of access to health care and substantial risks of catastrophic household expenditure on health care. Although in most other high-income countries policy concerns about fairness in health care focus on minimizing catastrophic household expenditure on health care, minimizing socioeconomic inequality in health care, giving priority to the worse off, and securing equal access to people with equal need. Despite this heterogeneity between different societies, important concerns for fairness exist in all societies, which health sector decision makers need to reflect in their decision making.

Fairness concerns sometimes clash with efficiency concerns. For example, health care decision makers routinely face clashes between the efficiency concern to do as much good as possible with scarce resources and the fairness concern to give priority to the most severely ill patients. Such clashes are seen, for example, in relation to dialysis machines, intensive care for preterm babies, and new drugs for end-of-life cancer patients. In each case, these forms of care are often not cost effective by conventional standards, implying that health decision makers could do more good by diverting scarce resources to other more cost effective forms of care. Yet decision makers often choose to fund these cost ineffective forms of care, reflecting important concerns for fairness that lie outside the conventional calculusof economic evaluation. Clashes of this kind are likely to become more frequent and more intense over time, even in high-income countries, as cost-increasing medical innovation increasingly drives a wedge between what is technologically possible and what publicly funded health systems can afford.

Similarly, public health decision makers in all countries routinely face clashes between improving population health and reducing socioeconomic inequality in health. For example, smoking cessation programs, physical activity programs, and other public health programs that seek to change lifestyle behavior are typically more effective in higher socioeconomic groups – and hence tend to increase socioeconomic health inequalities. Decision makers may therefore seek to redesign such programs to encourage participation among lower socioeconomic groups. In doing so, however, they may incur additional costs and limit the scope for improving health among socioeconomically advantaged populations, thus potentially reducing the sum total gain in population health. Clashes of this kind are fundamental and perennial issues in public health. The nature, size, and persistence of health inequalities are well-known. Yet policy makers still do not know how to reduce them. For example, despite a series of concerted attempts by the UK government in the 2000s to tackle health inequality, the 2010 Marmot Report found a gap of 14 years in disability-free life expectancy between the most and least deprived twentieths of small areas of England. Equity concerns of this kind are likely to become sharper over time, as global economic growth continues to be driven by technological innovation and other factors favoring high-skill workers. Applied economic evaluation evidence is needed about the costs and benefits of alternative programs for tackling health inequalities, to identify what works and to measure not only effects on average health outcomes but also effects on the socioeconomic distribution of health outcomes.

Figure 1 presents a simple stylized example of the two kinds of trade-off described above. Program 1 maximizes total health – it yields a gain of 2 health units for both groups – whereas program 2 results in a more equal distribution of health – it yields a gain of 3 health units for the worse off group B, but nothing for the better off group A. If group B is a severely ill group and the health units represent quality of life on a 0–100 scale, then this is a trade-off between total health and priority to the most severely ill patient group. If group B is a socioeconomically disadvantaged group and health units are life-years, then this is a trade-off between total health and socioeconomic inequality in life expectancy.

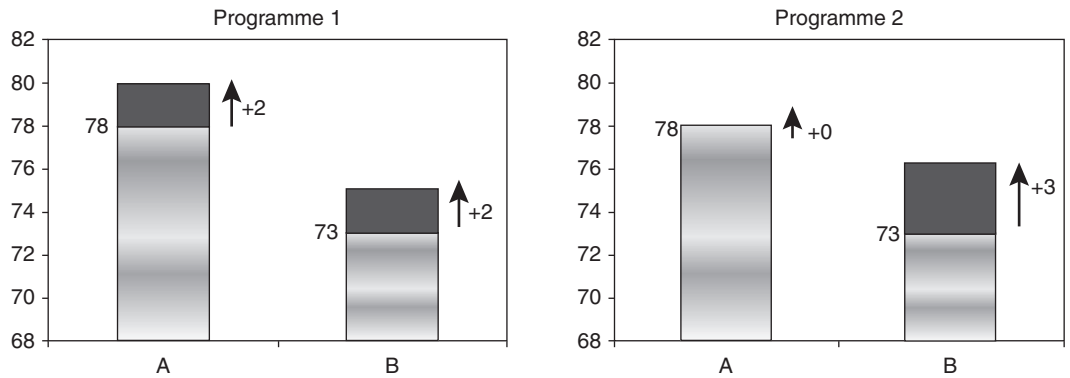


Figure 1 Trade-offs between total health and equal distribution of health. Imagine you are asked to choose between two programs, which will increase health in a population consisting of two groups. The programs cost the same. The areas in black represent increases in health. Program 1 delivers a larger total health benefit than program 2, whereas program 2 gives priority to the worse off group B and results in more equal levels of health. Which would you choose? Worse off group B might be a more socioeconomically disadvantaged group with lower life expectancy. Alternatively, it might be a more severely ill group with lower quality of life.

Equity in Health Economics

Health economists have made progress in the economic evaluation of health programs in recent decades. In the 1970s, the very idea of 'cost-effectiveness' was controversial among the medical community, methods were developmental, and applied economic evaluations were rarely used to inform real health care resource allocation decisions. Since then, health economists have developed sophisticated methods of economic evaluation that are now routinely used around the world to inform decisions about the funding of new health care technologies for particular groups of patients.

However, progress has focused on addressing concerns for efficiency, or cost-effectiveness, defined in terms of maximizing population health within a fixed health budget. Less attention has been devoted to addressing concerns for equity or fairness. Considerable methodological research has been done, and a variety of different theories and methods have been proposed for incorporating concerns for fairness into the economic evaluation of health programs. However, most of these methods remain developmental and even the most finished ones are still almost never used in the applied economic evaluations used to inform real resource allocation decisions.

The specific methods that have been proposed are reviewed in section Methods for Incorporating Concerns for Fairness into Economic Evaluation. Before that, it is useful to summarize the most frequently mentioned concerns for equity.

Concerns for Equity: Overview

In health care, concerns for equity often relate to the general principle that health care should be distributed in relation to need. This general principle can be divided into a vertical equity principle of greater treatment for greater need and a horizontal equity principle of equal treatment for equal need. Box 1 lists some potential concerns for fairness in health care that are often raised in relation to economic evaluations of new health care technologies. The first set of concerns, about prioritized patient subgroups, raises issues about which patients are in most 'need' – i.e., concerns for 'vertical equity'. The second category is about wishes not to discriminate between patients with the same degree of 'need' – i.e., concerns for 'horizontal equity'. The third set of concerns, about nonpatient benefits and nonhealth benefits, raise issues about how far 'need' relates to the needs of carers and dependents, as well as the needs of the patient, and how far 'need' relates to nonhealth needs as well as health needs. The fourth set of concerns, about industrial factors, raises issues about how far wider social policy objectives can be traded off against the equity principle of distribution according to need. The latter two categories can be thought of as concerns for efficiency, broadly construed to incorporate nonhealth benefits as well as health benefits, as opposed to concerns for fairness. However, they are important issues of social value judgment in health care resource allocation that go beyond concern for efficiency narrowly construed in the sense of health maximization.

In public health, as opposed to health care, concerns for equity often focus on reducing inequalities in population health – such as differences in life expectancy between socioeconomic groups. However, distinguishing between 'fair' and 'unfair' health inequality is problematic. Political and economic theorists have proposed numerous rival theories of what counts as 'fair'. Key dilemmas include how far decision makers should be concerned with:

- health inequality versus priority to improving the health of the worst off;
- inequality of income and other social determinants of health versus inequality in health;

Box 1 Potential societal concerns for fairness in health technology assessment

Prioritized patient subgroups
- The least healthy (e.g., severity of illness, poor current health, and poor prognosis)
- The socially disadvantaged (e.g., income, race/ethnicity, and vulnerable minority groups)
- Children and adolescents
- Life saving (i.e., permanently restored to normal life expectancy)
- Life extension near end of life (i.e., temporary relief from terminal illness)
- Type of illness and 'dread' (e.g., cancer)
- Health service responsibility (e.g., hospital infection)
- Unavailability of alternative treatment

Nondiscrimination
- Equal treatment of patients with different age; disability; gender reassignment; marriage and civil partnership; pregnancy and maternity; race; religion or belief; sex, and sexual orientation
- Equal treatment of patients with different potentials for health benefit
- Equal treatment of patients with different costs of treatment

Nonpatient and nonhealth benefits
- Impact on carers' health and wellbeing
- Impact on dependents' wellbeing
- Impact on productivity
- Impact on responsiveness and patient experience

Industrial factors
- Innovation and dynamic efficiency
- Promoting domestic industry
- Orphan drugs (i.e., prohibitive development cost due to rarity of condition)

- absolute inequality (e.g., gaps) versus relative inequality (e.g., ratios);
- inequality between groups versus inequality within groups (e.g., between individuals);
- univariate health inequality (i.e., the 'pure' distribution of health) versus bivariate health inequality (i.e., the joint distribution between health and one unfair determinant of health, such as income) versus multivariate health inequality (i.e., the joint distribution between health and multiple unfair determinants of health);
- avoidable versus unavoidable health inequality;
- compensable versus incompensable health inequality; and
- inequality of achieved health versus inequality of opportunity for health.

Each dilemma raises difficult value-laden issues of definition and measurement (see separate entries on 'Techniques for measuring equity in health and health care', and 'Field of inequality of opportunity in health').

Methods for Incorporating Concerns for Fairness into Economic Evaluation

The most ambitious approaches to incorporating concerns for fairness into applied economic evaluations are formal numerical value functions that take both efficiency and equity into account. The authors first review these. Less ambitious approaches include systematic characterization of relevant health equity concerns, multicriteria decision analysis, and estimation of the opportunity costs of equity. The authors return to these later on.

Formal Numerical Value Functions

In formal numerical value functions, trade-offs between efficiency and equity are expressed at a cardinal level of measurement, allowing for the overall value of an intervention or program to be estimated at that same level of measurement and thus made directly comparable with intervention or program costs. Formal value functions can in principle be applied in a fairly 'algorithmic' fashion, in the sense of requiring decision makers to use a single all-purpose set of social value judgments about equity, which leaves little room for deliberation and consultation with stakeholders about the appropriate set of value judgments to apply in each particular case. With suitable sensitivity analysis, however, formal value functions can also in principle be used in a more 'deliberative' fashion, in the sense of helping decision makers and stakeholders to deliberate their way toward a suitable set of value judgments. In this more 'deliberative' role, formal numerical value functions can help answer the questions: what are the implications of different value judgments for decision making in this case, and what implications might such value judgments have for other decisions in other contexts?

The social welfare function

One approach is to value health programs as a mathematical function of the distribution of health among individuals or groups in the relevant population. The standard theoretical framework that economists have used is a social welfare function, which takes as its arguments individual health or group average health, which might be measured, for example, using expected lifetime quality-adjusted life-years (QALYs). The social welfare function is typically increasing in health, reflecting concern for efficiency in the sense of health maximization. However, the social welfare function need not be a simple linear sum of individual or group average health. Instead, it may give more weight to improvements in health for some individuals or groups than others, depending on societal concerns for fairness.

One social welfare function is the isoelastic or Cobb-Douglas function, first proposed in a health context by Adam Wagstaff and subsequently used by Paul Dolan and Aki Tsychyia and others to empirically estimate 'equity weights' based on surveys of public views. In the simplest case of two individuals (or groups), this function takes the following form:

$$W = [\alpha h_1^{-r} + (1-\alpha)h_2^{-r}]^{-1/r}$$
$$h_1, h_2 \geq 0, \quad 0 \leq \alpha \leq 1, \quad r \geq -1, \quad r \neq 0$$

where h_1 and h_2 are respectively the health of person 1 and person 2 (or the average health of group 1 and group 2). This function can be visualized as a set of social indifference curves that pick out a socially preferred point on the health possibility frontier, see Figure 2. Points to the southeast of the maximin point are 'Pareto efficient', in the sense that the health of one person cannot be improved without reducing the health of another person. The social indifference curves pick out the best or fairest of these multiple 'Pareto efficient' points along the health frontier. Two parameters determine the shape of the social indifference curves. First, a general inequality aversion parameter, r, reflecting general aversion to health inequality between all individuals or groups. The magnitude of this parameter reflects the degree of curvature of the social indifference curves. Zero inequality aversion implies straight line 'utilitarian' style indifference curves, with $r = -1$, as illustrated by the blue-dashed lines, which pick out the health maximizing point in Figure 2. Complete inequality aversion implies L-shaped 'Leontief' or 'Rawlsian' style indifference curves, as r approaches infinity, as illustrated by the green-dashed lines, which pick out the maximin point in Figure 2. Second, a special priority parameter, α, reflecting priority to individuals or groups with a special equity-relevant characteristic (e.g., low socioeconomic status). This parameter would pivot the social indifference curves about the 45 degree line of equality. When additional individuals or groups are added into the analysis, additional special priority parameters can be added to allow for additional equity-relevant characteristics (e.g., ethnicity, disability, responsibility etc.), whereas the general inequality aversion parameter will apply to all individuals or groups in the analysis.

Another type of formal approach consists of weighting QALYs achieved by an intervention or program by the characteristics of the people who get the health gains, or possibly by the characteristics of the program, rather than by the resulting distribution of health. Different programs may then be compared with respect to value in terms of 'equity-weighted QALYs'.

One approach of this kind was proposed by Alan Williams in the so-called 'extended fair innings argument'. According to

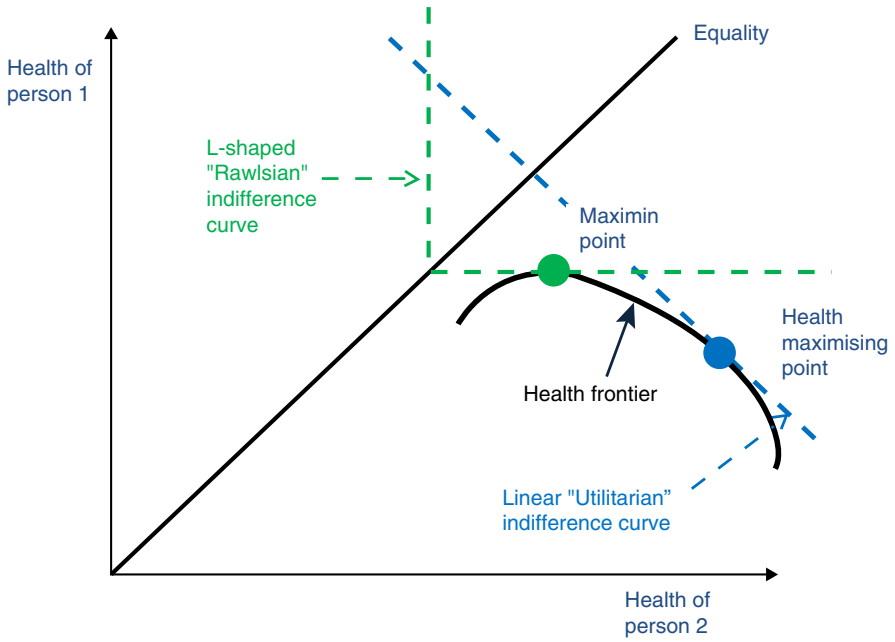


Figure 2 Health possibility frontier with social indifference curves.

Williams, all individuals are entitled to a normal 'fair share' of quality-adjusted life expectancy (the 'fair innings'). This implies that health gains to individuals below the fair innings norm – for example, the poor and disabled – should receive greater weight than health gains to individuals above the norm. In Williams' approach, this is achieved by multiplying QALY gains by separate fair innings weights. A similar model was proposed by Han Bleichrodt and colleagues, whose 'rank-dependent utility' theory weights individuals according to their rank in the distribution of expected lifetime health. Other approaches focusing on equality in life time health have been proposed by Magnus Johannesson and Ole Frithjof Norheim (*see* Further Reading).

A problem with models that focus on equality in life time health is the implication that health gains to older individuals who have already enjoyed a long and healthy life should receive lower weight than health gains to younger individuals expected to have a shorter or less healthy life. This could, for instance, mean that pain relief in an elderly person would receive lower priority than pain relief in a young person with a life expectancy below that of the elderly person, assuming the pain relief in both cases has the same cost per QALY gained.

Another approach to equity weighting of QALYs is so-called 'cost-value analysis' (CVA, mentioned below). In this approach, health gains in terms of QALYs are valued more the more severe the condition of the target group is. The approach furthermore discriminates less strongly than the conventional QALY maximization approach does between gains for people with equal severity of illness with different capacities to benefit from treatment – for example, due to differences in disease, age, or comorbidity. Erik Nord and colleagues showed in 1999 that these two features may – as with concerns for fair innings – be achieved by application of separate equity weights. However, the main approach in CVA is to replace conventional utilities by 'societal values' that – in a coordinate diagram with utilities on the x-axis and societal values on the y-axis – form a curve that is convex toward the y-axis and compresses moderate and mild problems toward the upper end of the 0–1 scale (mentioned above).

In the Netherlands, a government guideline from 2009 indicates that willingness to spend public money in order to gain a QALY will range from 10 000 euros for conditions of little severity to 80 000 euros for conditions of great severity. This 'graded willingness to pay' is effectively the same as assigning severity weights to QALYs. In the Dutch context, severity is measured in terms of 'proportional shortfall', which builds on 'absolute shortfall'. Absolute shortfall is the difference between a patient's expected remaining QALYs and the number of remaining QALYs in average individuals of the same gender and age. Proportional shortfall is absolute shortfall relative to the number of remaining QALYs in average individuals of the same gender and age.

Preference data

All the above models require data on societal preferences regarding trade-offs between efficiency and equity. Only on the basis of such data can the models be of practical use.

Preferences for such trade-offs are normally elicited from samples of the general population. They can be elicited in various ways.

In estimating parameters in a social welfare function, one possible approach is to ask subjects to compare different possible health scenarios for a set of social groups. The scenarios might vary, for example, with respect to average health in terms of quality-adjusted life expectancy (QALE) and the distribution of QALE between groups. Subjects may be asked to compare scenarios pairwise, and their willingness within each pair to trade-off equality for gains in average health (and vice versa) may be observed. Through statistical techniques the central tendency of such trade-offs may then be used to

estimate parameters in the social welfare function. This approach has not been widely applied, but was used by Paul Dolan and Aki Tsuchyia in 2009 in a small-scale methodological study.

Another approach is the person trade-off technique as described by Nord in 1995. This is commonly used to obtain equity weights for QALYs. The basic format is that subjects are asked to compare health gains for different groups of people that differ on some variable that is considered relevant for equity reasons. For instance, subjects are asked to consider a group A of 10 people who can obtain an improvement from 0.6 to 0.8 on a 0–1 utility scale. Another group B of *N* people can obtain an improvement from 0.8 to 1.0. All else equal, the health gain in terms of QALYs is equally large for each individual in the two groups. But people in group A are worse off. QALYs to them may therefore be valued more highly than the same number of QALYs to people in group B. To measure the strength of preference for the more severely ill group, subjects are asked how many people there would have to be in group B for that program to be considered equally worthy of funding as the program for group A. If the mean response is 20 B ~ 10 A, the implication is that the improvement from 0.6 to 0.8 is valued twice as highly as the improvement from 0.8 to 1.0. Weights for age and duration of benefits can be obtained in a similar fashion. In the last three decades, results from a number of person trade-off studies in different countries have been published. In principle, the results may be used as guidance in construction of models of equity-weighted QALYs. In practice, however, little use has hitherto been made of these data, partly because of uneasiness about their accuracy. An exception is Norway, where the Norwegian Medicine Agency since 2000 has recommended that conventional cost-utility analyses in terms of QALYs be supplemented by analyses using person trade-off-based health state values.

A third measurement approach was introduced by Paul Dolan in 1998. He asked subjects to compare a health gain from utility level 0.2 to 0.4 for a person A with a gain from level 0.4 to level X for a different person B. What would X have to be for subjects to consider the two health gains equally worthy of funding? On average, subjects answered 0.8, which suggests that they thought person A deserved a 'severity weight' of 2 compared to person B.

Finally, a fourth approach is to use pairwise choices between different groups with different health gains. This is similar to the person trade-off approach, except that individual subjects do not directly state their strength of preference but instead this is indirectly inferred from between-subject and/or within-subject patterns of pairwise choices using statistical modeling methods. This 'stated preference' or 'discrete choice experiment' approach has for instance been used in the UK by Rachel Baker and colleagues.

Other Approaches to Incorporating Concerns for Equity

Systematic characterization of relevant health equity concerns

This approach merely aims to foster a more systematic approach to identifying and characterizing the equity considerations at stake and to presenting relevant qualitative and quantitative background information that decision makers may find helpful. It might be useful, for instance, to develop a 'checklist' of potentially relevant equity concerns, based on precedent from past decisions and deliberation among stakeholder groups. Where a particular concern on the checklist is deemed relevant to the decision in hand, it would then be useful to present background information about the importance of this concern to the decision in hand. This might include qualitative information about stakeholder views and prior decision precedents; it might also include quantitative information, which puts the relevant equity concern into perspective – for instance, about how large and important the decision-relevant health inequality is compared with other health inequalities.

Multicriteria decision analysis

Multicriteria decision analysis aims not only to provide a qualitative 'checklist' of equity concerns but also to give each concern a numerical score and weight so as to arrive at an overall ranking of decision options. This approach has been advocated by Rob Baltussen and Louis Niessen in the context of both local and national health care planning in low and middle income countries, and has from time to time been used in health care priority setting exercises conducted in high-income countries. An advantage of this approach over an informal 'checklist' is that the scoring and weighting process can facilitate stakeholder engagement, transparency, and consistency. However, this approach does not integrate fairness concerns within economic evaluation – rather it takes the results of economic evaluation as one of many parameter inputs into a broader quantitative assessment. Furthermore, methods for the scoring and weighting of criteria currently lack the analytical rigor and evidential basis of methods for economic evaluation: much of the scoring and all of the weighting is typically done using decision maker or stakeholder opinion.

Health opportunity cost of equity

A third approach aims to estimate the health opportunity cost of a particular equity concern – for example, in terms of QALYs forgone by pursuing a more 'equitable' option compared with the QALY maximizing option. Every departure from health maximization on grounds of equity has an opportunity cost in terms of sum total health forgone. The size of that opportunity cost is a test of how important that equity concern is deemed to be.

This approach can be implemented using the standard methods of cost-effectiveness analysis, using the cost per QALY threshold to represent the health opportunity costs of unknown displaced programs. Or, if displaced programs can be identified and evaluated, mathematical programming can be used based on data rather than assumptions about the opportunity costs and equity characteristics of displaced program. Either way, one can compute the opportunity cost of equity by computing the difference in total net QALY benefit between 'more efficient' and 'more equitable' programs. When the equity concern relates to health inequality, this approach can be extended by calculating a health opportunity cost per unit reduction in health inequality. One could even imagine establishing a 'cost-equality threshold'

in terms of a benchmark cost per unit reduction in health inequality from previously evaluated programs.

An advantage of the opportunity cost approach is that it can be used to address any kind of equity consideration and not just concerns about health inequality. For example, during the 1990s, the UK Standing Medical Advisory Committee advised local health authorities against adopting a racially selective policy on screening for sickle cell anemia – which is more prevalent in certain ethnic minority groups – even though this may have been the most cost-effective strategy. Franco Sassi and colleagues showed in 2001 that imposing the equity constraint of nondiscrimination imposed a health sacrifice in terms of cost-effectiveness.

A limitation of the opportunity cost approach, however, is that it only looks at the cost of the equity concern, not the benefit. It measures the equity–efficiency trade-off implied by a particular decision (a factual matter) but does not value the trade-off that policy makers ought to make (a moral matter). That is, it does not help the decision maker decide how large a sum total sacrifice (in terms of health and/or nonhealth benefits) is worth making in order to pursue a particular equity consideration. This runs the risk of lack of transparency and inconsistency across decisions, because decision makers are then free to make implicit judgments about how much health sacrifice is worth making in pursuit of a particular equity principle – and to vary those judgments from one decision to another without giving any explicit justification.

Conclusion

Health economists have developed a substantial and growing body of theoretical tools and empirical methods for incorporating concerns for fairness into economic evaluation. However, these methods have not yet been taken up and applied in routine economic evaluations used to inform resource allocation decisions. There are two main barriers to this. First, concerns for fairness are contested and context specific. Second, research requirements for measuring population preferences for fairness are often greater than those for measuring health and valuations of health. Data on preferences for fairness are therefore much more limited than valuation data used in estimating efficiency.

These barriers to progress are not insurmountable. As the authors have shown, there are ways of specifying equity objectives in such a way they can be quantified in economic evaluation. As pressures for transparency and accountability in public life increase, and as clashes between equity and efficiency concerns in health care and public health become ever more apparent and insistent, policy makers may be persuadable to articulate more specific health equity goals. To the extent that these equity goals are context specific, it may be possible to harness deliberative approaches to facilitate stakeholder 'buy-in' to particular equity goals and analytical approaches in particular decision-making contexts. Furthermore, data sources are increasingly rich as are the methods available for analyzing them. Person trade-off data already yield some useful information for equity weighting of QALYs. Methods of evidence synthesis are available for combining patient level data from a network of randomized control trials along with observational data sources. These methods could be exploited to generate information on the distribution of health effects between equity-relevant patient groups. Econometric methods are available for identifying causal effects and subgroup heterogeneity in causal effects, by exploiting observational data from surveys, administrative databases and trials – including record-linkage studies that link together all three types of data. A key challenge for the next generation of health economists is to harness these data and methods in ways that fit the contours of societal concerns for fairness and deliver analytical insights that health sector decision makers find convincing and useful.

See also: Cost–Value Analysis. Efficiency and Equity in Health: Philosophical Considerations. Equality of Opportunity in Health. Health and Health Care, Need for. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach. Measuring Vertical Inequity in the Delivery of Healthcare. Quality-Adjusted Life-Years

Further Reading

Baltussen, R. and Niessen, L. (2006). Priority setting of health interventions: The need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation* **4**, 14.

Bleichrodt, H., Diecidue, E. and Quiggin, J. (2004). Equity weights in the allocation of health care: The rank-dependent QALY Model. *Journal of Health Economics* **23**, 157–171.

Cookson, R., Drummond, M. and Weatherly, H. (2009). Explicit incorporation of equity considerations into economic evaluation of public health interventions. *Journal of Health Politics, Policy, and Law* **4**, 231–245.

Culyer A. J. and Bombard Y. (2011). An equity checklist: A framework for health technology assessments. Centre for Health Economics. CHE Research Paper 62, University of York.

Dolan, P., Shaw, R., Tsuchiya, A. and Williams, A. (2005). QALY maximisation and people's preferences: A methodological review of the literature. *Health Economics* **14**(2), 197–208.

Dolan, P. and Tsuchiya, A. (2009). The social welfare function and individual responsibility: Some theoretical issues and empirical evidence. *Journal of Health Economics* **28**, 210–220.

Epstein, D. M., Chalabi, Z., Claxton, K. and Sculpher, M. (2007). Efficiency, equity, and budgetary policies: Informing decisions using mathematical programming. *Medical Decision Making* **27**(2), 128–137.

Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics* **28**, 73–90.

Johannesson, M. (2001). Should we aggregate relative or absolute changes in QALYs? *Health Economics* **10**, 573–577.

Nord, E. (1995). The person trade-off approach to valuing health care programs. *Medical Decision Making* **15**, 201–208.

Nord, E., Pinto, J. L., Richardson, J., Menzel, P. and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics* **8**(1), 25–39.

Norheim, O. F. (2001). Gini impact analysis: measuring pure health inequity before and after interventions. *Public Health Ethics* **3**(3), 282–292, doi:10.1093/phe/phq017.

Sassi, F., Archard, L. and Le Grand, J. (2001). Equity and the economic evaluation of healthcare. *Health Technology Assessment* **5**, 3.

van de Wetering, E. J., Stolk, E. A., van Exel, J. A. and Brouwer, W. B. F. (2013). Balancing equity and efficiency in the Dutch basic benefits package using the principle of proportional shortfall. *European Journal of Health Economics* **14**(1), 107–115, doi:10.1007/s10198-011-0346-7.

Wagstaff, A. (1991). QALYs and the equity-efficiency trade-off. *Journal of Health Economics* **10**(1), 21–41.
Williams, A. (1997). Intergenerational equity: An exploration of the 'fair innings' argument. *Health Economics* **6**(2), 117–132.

Relevant Websites

www.iseqh.org
International Society for Equity in Health.

www.instituteofhealthequity.org/
Marmot Review and Institute of Health Equity.

www.statisticalconsultants.co.nz/weeklyfeatures/WF26.html
Social Welfare Functions.

www.who.int/social_determinants/en/
WHO Commission on the Social Determinants of Health.

www.worldbank.org
World Bank (Analyzing Health Equity Using Household Survey Data by Owen O'Donnell, Eddy van Doorslaer, Adam Wagstaff, and Magnus Lindelow).

Infectious Disease Externalities

M Gersovitz, Johns Hopkins University, Baltimore, MD, USA

Glossary

Endogeneity An economic variable is said to be endogenous if it is a function of other parameters or variables in a model.

Equity Equity is not necessarily to be identified with equality or egalitarianism, but relates in general to ethical judgments about the fairness of the distribution of such things as income and wealth, cost and benefit, access to health services, exposure to health-threatening hazards, and so on. Although not the same as 'equality', for some people, equity frequently involves the equality of something (such as opportunity, health, and access).

Externality An externality is a consequence of an action by one individual or group for others. There may be external costs and external benefits. Some are pecuniary, affecting only the value of other resources (as when a new innovation makes a previously valuable resource obsolete); some are technological, physically affecting other people (communicable disease is a classic example of this type of negative externality); and some are utility effects that impinge on the subjective values of others (as when, e.g., one person feels distress at the sickness of another, or relief at their recovery).

Herd immunity The effective stoppage of the spread of a disease when a particular percentage of a population is vaccinated. This critical percentage varies according to the disease, the interactions between members of the population and the vaccine, but 90% is not uncommon.

Market imperfections Markets in health care are notable for 'failing' on a number of grounds, including asymmetry of information between producers (medical professionals of all kinds) and consumers (patients actual and potential); distorted agency relationships, failure of patients to behave in accordance with the axioms of rational choice theory; incomplete markets, especially those for risk; monopoly; and externalities and the presence of public goods.

Public good The technical meaning of a 'public good' in economics is a good or service that it is not possible to exclude people from consuming once any is produced. Street lighting and national defense are classic examples. Public goods are nonrival in the sense that providing more for one person does not entail another having any less of it. Some externalities have the character of publicness, such as the comfort one may have when others are protected from ill-health.

Rationality Technically, in economics, rationality means behaviour in conformity with axioms such as: completeness (either A is preferred to B, or B to A or an individual is indifferent between them – where the As and Bs are objects of choice); transitivity (if A is preferred or indifferent to B and B is preferred or indifferent to C, then A is preferred or indifferent to C); continuity (there is an indifference curve such that all points to its north-east are preferred to all points to its south-west); convexity (the marginal rate of substitution is negative); and nonsatiation (more is always preferred).

Utility Variously defined in the history of economics. Two dominant interpretations are hedonistic utility, which equates utility with pleasure, desire-fulfilment, or satisfaction; and preference-based utility, which defines utility as a real-valued function that represents a person's preference ordering.

Introduction

Infectious diseases are caused by pathogenic microorganisms, such as viruses, bacteria, parasites, or fungi. For almost any infectious human disease, what one person does about it affects the probability that other people get infected. Some infectious diseases spread from person to person through direct physical contact as in the case of sexually transmitted infections. People can also shed an infectious agent into the air, water, onto food, or other surfaces where other people come into contact with it and become infected, as with respiratory or diarrheal infections. Some infectious agents have life cycles that involve stages in both the human host and in a vector organism such as a mosquito. Thus in the case of malaria, an infected mosquito transfers the malaria parasite to an uninfected person through feeding, but an uninfected mosquito can likewise become infected by an infected person, making it possible for the mosquito to infect someone else. Infected people do not always play this role in infecting other people because humans may be dead-end hosts. For example, people infected by roundworms with trichinosis pose no risk to others as long as the larvae in their flesh are not eaten by suitable host animals that are subsequently eaten by other people.

To this point, one person puts another at risk because the first person is infected. Although operative for most infectious diseases, this mechanism is not the only one that affects the risks of infection faced by others. People may put others at risk of infection without being infected themselves. People who do not spray their own houses with insecticide to kill mosquitoes and other disease vector organisms put their neighbors at risk regardless of whether they themselves are or become infected or not.

In all these situations, people face choices. At an abstract level people are making choices about prevention including immunization and about therapy. For any actual disease, these choices are about a wide variety of day-to-day actions.

Of course, epidemiologists and other researchers on human health are well aware how infections spread and, in particular, that the actions of people affect the risks that others face. Epidemiologists use terms such as herd immunity

and community or mass effects to denote the ways that a lessening of the infection risk for some people lessens the risk for others. For the most part, it is from these disciplines that economists and others learn about the pathways of infection and what can be done to prevent infections or to mitigate them once they have occurred. Mathematical epidemiology provides algebraic models of the dynamics of infectious diseases, the starting point for an economic theory of infectious diseases and their control. Even without endogenous behavior by utility-maximizing individuals, these models are nonlinear and dynamic, capable of exhibiting complicated even chaotic behavior.

Basic Nature of the Externality

Unlike epidemiologists, economists predict behavior and devise policy using the hypothesis of rational decision making by self-interested individuals who pursue objectives subject to constraints. To the extent that people are selfish, they ignore the consequences to others put at risk by their actions or failure to act. It is the discrepancy between the choices made by this type of individual and the choices that are desirable for society as a whole taking into account all the consequences of an individual's actions that defines the externality and gives precision to this central concept in the economics of infectious disease control.

If individuals do too little of something from society's perspective, the classic solution to the problem of such an externality is to subsidize the activity – and in the reverse situation to tax the activity. With more than one activity going on simultaneously, it is desirable to think in terms of a package of interventions. For instance, if there is a preventive activity and a therapeutic one, the government should intervene to influence both and it is natural to ask how these interventions should be coordinated. If an infection is transmitted from one person to another, and if a person once infected recovers to be again susceptible, then the optimal package is to subsidize prevention and therapy at equal rates. The externality arises because people spend too much time in the state of being infected and it is socially just as desirable to give them incentives to stay out of this state as to get out of it once they are in it. This finding underlines that both prevention and therapy are associated with externalities. For other diseases from which people do not recover but rather die, or from which they recover to be immune, the package has different qualitative properties. In the case of vectors there may be many different types of prevention in terms of their roles in the model, with consequently different rates of subsidy.

Not all formulations of dynamic models of infectious disease lead to externalities or at least ones that justify government interventions. For instance, consider a simple model in which immunization always confers complete immunity, people if infected stay that way forever and never die, there are no newly born susceptibles, and all individuals have the same preferences (including attitudes toward risk and time) and susceptibility to infection. In this model, everyone gets immunized at the same time. This time is determined by the overall infection rate which determines the risk of infection and therefore the benefit of immunization. Once everyone is immunized, there is no one left to benefit from other people protecting themselves against being infected and therefore, no reason to move the time at which everyone who has remained susceptible gets immunized. Consequently there is no justification for government intervention to offset an externality. But this example is not very general and its importance is in emphasizing that it is being infected, rather than being susceptible and potentially infectible, that generates the externality. In general, even in models for which the only choice is to be immunized or not, there will be a justification for an optimal subsidy to immunization because one or other of the assumptions stated above do not obtain.

There is even the possibility of positive externalities if individuals increase activity that puts them at risk of infection. An example of this result occurs when there is more than one (homogeneous) group in which the groups mix together. First, consider a high-activity (and therefore high-risk) group mixing randomly only with its own members. The infection rate will be high. Now consider a second, low-activity group that increases its level of random contacts from none to one, some of which are with the high-activity group. Any member of the low-activity group who becomes infected does not infect anyone else because they have no more contacts. But by diverting high-activity people from having contact with other high-activity people, the prevalence of infection overall may fall and if the effect is strong enough, the infection may even disappear. The example illustrates not just the possibility of positive externalities but also the danger of thinking in terms of average activity levels without regard for the variability in activity levels in the face of a highly nonlinear process.

Policies to Offset Externalities

The general expectation, however, is that people do too little from a social perspective to avoid being infected, either by making too little effort to avoid becoming infected or to recover once infected. In principle, these problems could be fixed by subsidies, but in practice subsidies may be infeasible so that the first best as seen by society is unattainable. To internalize the externalities associated with infectious diseases optimally, subsidies have to be targeted at outcomes such as the probabilities of becoming infected or recovering from infection. If each probability depended only on inputs that could be subsidized then these inputs could be targeted. But in practice such probabilities depend on many inputs, both marketed goods and services such as insecticides, bed nets, medicines, or the services of health professionals, and nonmarketed inputs such as time and effort by the person involved who may also suffer side effects in the case of therapies. All these inputs may be brought together in activities that may be spread over time and space and expensive to monitor, and therefore hard or impossible to subsidize. Some health-related activities are even private and intimate. Consequently, policy may not be able to achieve targeting at the probabilities but rather only at some of the inputs not all of which are necessarily used exclusively to affect the probabilities, hence situations of the second best.

Examples of imperfect targeting abound. For instance, one can subsidize hand soap but not the outcome of sanitized

hands. Soap may be used for other purposes than health-related hand washing such as clothes washing that are then subsidized as well with a loss of economic efficiency. If people find washing hands disagreeable but its social benefits are large enough, it may be necessary in theory to pay them to wash their hands but it may be impossible in practice to do more than give soap away free. Paying people to take soap is not the same thing as getting them to use it to wash their hands. In the case of freely provided bed nets for protection against malaria, it has been claimed that they have been diverted for other uses such as fishing, but a recent review has found almost no such evidence. In the case of sexually transmitted infections, it is safe sex acts that should be subsidized, but typically what has been done is subsidizing or giving away condoms, which is not the same thing as ensuring their use. In the case of tuberculosis, programs of directly observed therapy short course (DOTS) pay for patients to be supervised to make it more likely that they take their medicines. People who do not comply and do not recover continue to infect others, and may even develop drug-resistant infections through incomplete adherence to the therapeutic protocol and then infect other people who in turn are more difficult to cure even if they comply. In principle, people could be paid to maintain their uninfected status as regards human immunodeficiency virus (HIV) or other infectious diseases if it is possible and cheap to test infection status. But it will often be much more difficult to implement subsidies to correct the externalities of infectious diseases than to deal with other types of externality such as vehicular pollution or congestion, which themselves pose difficult enough challenges to the implementation of the first best even under ideal conditions.

A failure of the government to intervene, either completely or partially, has implications for the effect on welfare of changes in the parameters of the system. The outcome can be immiserization, a perverse transformation of a seemingly beneficial change into an actual decrease in welfare. For instance, there is the question of how welfare responds to a lowering in the cost to individuals of being infected because of a more effective treatment. If the externality has been internalized by first-best government interventions, welfare is always increased by such a change even though the infection rate likely rises. But if the externality is not internalized, the direct effect of the reduction in the cost of infection (corresponding to the only effect if the externality were internalized) may be overwhelmed by a worsening of the externality. The reason immiserization may occur is that people make choices about prevention and therapy that are socially suboptimal because they disregard their effect on the welfare of others. A decrease in the private cost of infection could worsen this discrepancy between the socially desirable choices and privately rational choices about prevention and therapy, and on balance welfare declines even though the direct effect of the decrease in the cost of infection is to increase welfare.

Instead of, or in addition to subsidies, governments use methods of coercive physical control such as quarantine of people who may be incubating an infection, isolation of people known to be infected, and culling of domestic animals that may play a role in the infection of people. Thailand has successfully used administrative measures such as tracing clients who attend clinics for sexually transmitted infections back to brothels where condoms are not used and then pressuring brothel owners to ensure that condoms are used under threat of closure. DOTS has aspects of a subsidy and physical control depending on how one interprets the way it promotes compliance with the drug protocol. It does not mandate compliance subject to coercive sanctions but its supervision could either be thought to facilitate compliance by lowering its cost, for instance by providing a reminder, or to raise the cost of not complying by hectoring and nagging. In either case, it influences people one-on-one, rather than through a general subsidy of something people purchase.

People subject to policies of physical restriction are usually not fully compensated for the costs to themselves and so the policies are often resisted and dodged. In the case of isolation, people may have access to therapy so there is that benefit to them which promotes compliance. During the severe acute respiratory syndrome (SARS) epidemic in Taiwan, quarantined people were brought food and had odd jobs done for them to lessen their costs of compliance. In other cases, compensation may help induce compliance although it is important to ensure that it does not result in perverse effects such as the needless slaughter of animals by making such activity profitable.

Need for Persistent Policies

In addition to specifying how to target subsidies, program design has to address whether interventions need to be permanent or temporary. If it is optimal for the infection to remain endemic at some level, then subsidies will have to be permanent because there will be an ongoing discrepancy between the socially and privately desirable levels of prevention and therapy. Beginning from an infection rate that is different from the final one, the discrepancy between the social and private incentives to undertake prevention and therapy will be changing over time and consequently so will the optimal levels of subsidies as the infection rate settles toward its long-run endemic level. If, however, the infectious disease can be eradicated, then by definition further subsidies will not be necessary and programs can be ended. Indeed, it is this hope combined with the end to all the costs borne by individuals that makes eradication for all its difficulties such an attractive goal.

In the absence of scientific breakthroughs of an almost magical sort, however, eradication is not likely in the near future for most infectious diseases. One reason it may nonetheless be possible to lessen expenditures over time is if part of the reason for subsidies is to pay for the dissemination of information about the infection and how to respond to it. Information dissemination may be implicit, as when someone learns about the benefits of prevention or therapy by trying them out. Information dissemination may also have an externality component if people learn from others and without compensating the people from whom they learn for their own costs of acquiring and providing this information. There is also an externality associated with information if a lack of information leaves people acting against their own interest in ways that also have costs to others. Once the message is out, however, it may need little subsequent repetition so that it

may indeed be possible to wind down expenditure on information. Information dissemination by itself does not, however, deal with the ongoing hard core of discrepancy between the private and social benefits of prevention and therapy.

Sometimes noneconomists argue that people will take 'ownership' of measures to control the spread of infections and thenceforth subsidies can be lowered or ended altogether. If by ownership one means that once people are informed about the existence of a disease, its modes of transmission and the possibilities of prevention and therapy, they will do things differently, then such a view is partially consistent with an externality-based argument. If not, however, it is hard to understand what the argument means other than a somewhat naive faith in the power of habit formation as once subsidies are removed, behavior will likely revert to its original self-interested and socially suboptimal form.

Span of the Externality

In the case of infectious disease, people do not generate risks and external costs (and possibly benefits) equally for everyone in the whole world. It makes sense to think of the span of the externality, i.e., the range of people who may suffer costs external to someone else's choices. People who are directly exposed to risk by someone are more likely to be close to the person putting them at risk. This closeness may be because the people put at risk have important social relationships with the people who are infected, such as family, sexual partners and friends, or because they are in close geographical proximity such as people who live, work or shop in the same neighborhoods, or commute on the same routes. Of course, someone's failure to avoid infection can have worldwide implications through a chain of infection, as in the case of emerging infections like HIV, SARS, or avian influenza.

Naturally, what it means to be geographically close depends on the mode of transmission of the disease and intervention, something that needs documentation on a case-by-case basis. For instance, insecticide-treated bed nets protect people who sleep under them from malaria by providing a barrier. But they also kill mosquitoes (and other disease-transmitting insects) that make contact with the nets. In effect, the people sleeping under them serve as bait. The consequence is that these insects do not have the chance to bite other people who are not under nets, effects that seem to prevail up to 300 m from the people using the nets, a clear external benefit to the non-users.

Close relationships such as family or sexual partners raise several issues. At the simplest, people in this type of relationship may know about each other's infection status through observing symptoms or medication, or through knowing who could have infected them as in the case of a sexually transmitted infection. Information of this sort in turn raises questions of strategy, in which susceptible people take actions with regard to specific people. There may be conflict over the use of condoms or testing. Families may dissolve over the infection of some of its members and the threat they pose to others. This potential for conflict raises the question of altruism versus self-interest. To what extent does someone act to avoid infecting others? If people are entirely altruistic, caring about the well-being of everyone who is affected by their decisions, then there are no externalities. In other situations they may be forced to take account of the risks they pose to others. Here one sees very starkly, possibly as a matter of life or death, the many possible considerations that arise in families.

Tuberculosis provides a good example of these family issues. It is often fatal and casually transmitted – a terrifying combination. As a result, relatives do indeed force infected members to leave the household. Understanding the motivations within the household is especially important in the case of DOTS. One focus of debate among DOTS professionals is who should be the supervisor that ensures that the infected person complies. Cost is an issue because specialized personnel – especially medical personnel – are expensive and either they or the patient have to travel for compliance to be observed and the protocol extends over many months with daily medication. Another alternative is supervision by a family member. Here it is important to identify motivated supervisors who will get the job done. There can be several motivations: Altruistic concern for the infected family member, fear of contraction of infection, or self interest in having the infected family member return to contributing to the family by earning income or doing chores. But by the very nature of the fact that not all costs external to the infected individual occur within the family, it is unlikely that family members will always be sufficiently motivated to serve the broader social interest.

The span of the externality is important not just in determining who infects whom. It also helps think about what level of government should be dealing with the internalization of the externality. The government should encompass the people who generate and experience the external costs, otherwise the government itself will lack the motivation to internalize the externality. It is a simple principle but one that is difficult to apply when the infection spreads globally. At a global level there is no supranational government that can compel action on health and even international organizations such as the World Health Organization (WHO) depend on the cooperation of their member countries and have no independent authority. National governments may not want to share information or admit WHO or other foreign teams to investigate outbreaks and, in general, they have made no commitment to do so. This type of issue has arisen in the surveillance of avian influenza in some Asian countries during the 2000s. Conflict between different national interests also arises. For example, rich countries may decide to ban dichlorodiphenyltrichloroethane (DDT) for environmental reasons even though DDT if used for antimalarial spraying of dwellings in poor countries can be highly beneficial and without significant environmental costs if it is not diverted to agricultural use.

Conclusion

Taken together, what is known suggests a robust role for the externality in understanding the dynamics of infectious diseases and how to control them. But it is only one set of economic considerations in the design of policies. Insurance markets are notorious for posing their own set of market

imperfections and are highly relevant to health where the risks are large and people are fearful. Issues involving equity also deserve important attention.

See also: Infectious Disease Modeling. Sex Work and Risky Sex in Developing Countries. Vaccine Economics. Water Supply and Sanitation

Further Reading

Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans*. Oxford: Oxford University Press.

Bock, N. N., Sales, R. -M., Rogers, T. and DeVoe, B. (2002). A spoonful of sugar...: Improving adherence to tuberculosis treatment using financial incentives. *International Journal of Tuberculosis and Lung Disease* **5**, 96–98.

Eisele, T. P., Thwing, J. and Keating, J. (2011). Claims about the misuse of insecticide-treated mosquito nets: Are these evidence-based? *PLOS Medicine* **8**, 1–3.

Gersovitz, M. (2011). The economics of infection control. *Annual Review of Resource Economics* **3**, 277–296.

Gersovitz, M. (2013). Mathematical epidemiology and welfare economics. In Manfredi, P. and d'Onofrio, A. (eds.) *Modeling the interplay between human behavior and the spread of infectious diseases*. New York: Springer.

Hawley, W. A., Phillips-Howard, P. A., ter Kuile, F. O. et al. (2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *American Journal of Tropical Medicine and Hygiene* **68** (supplement 4), 121–127.

Hsieh, Y. H., King, C.-C., Chen, C. W. S. et al. (2005). Quarantine for SARS, Taiwan. *Emerging Infectious Diseases* **11**, 278–282.

Keeling, M. J. and Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton: Princeton University Press.

Khan, M. A., Walley, J. D., Witter, S. N., Shah, S. K. and Javeed, S. (2005). Tuberculosis patient adherence to direct observation: Results of a social study in Pakistan. *Health Policy and Planning* **20**, 354–365.

Lagarde, M., Haines, A. and Palmer, N. (2007). Conditional cash transfers for improving uptake of health interventions in low- and middle-income countries: A systematic review. *Journal of the American Medical Association* **298**, 1900–1910.

Normile, D. (2005). Vietnam battles bird flu... and critics. *Science* **309**, 368–373.

Normile, D. (2007). Indonesia to share flu samples under new terms. *Science* **316**, 37.

Rojanapithayakorn, W. (2006). The 100% condom use program in Asia. *Reproductive Health Matters* **14**, 41–52.

Rosenberg, T. (2004). *What the world needs now is DDT*. New York: New York Times.

Relevant Websites

http://ccdd.hsph.harvard.edu/
Harvard Center for Communicable Disease Dynamics.

http://www.hpa.org.uk/
UK Health Protection Agency.

http://www.cdc.gov/ncezid/
US National Center for Emerging and Zoonotic Infectious Diseases.

http://www.who.int/topics/infectious_diseases/en/
WHO Website on Infectious diseases.

Infectious Disease Modeling

RJ Pitman, Oxford Outcomes Ltd, Oxford, UK

Glossary

Basic reproductive number (R_0) The number of secondary infectious hosts arising from one average primary infectious host in an entirely susceptible population.

Communicable disease Illness due to a specific infectious agent or its toxic products that arises through transmission of that agent or its products from an infected person, animal, or reservoir to a susceptible host, either directly or indirectly through an intermediate plant or animal host, vector, or the inanimate environment.

Incidence of infection The number of new infections arising in a defined period of time, typically expressed as a rate per 100 000 population per year.

Prevalence of infection The proportion of the population infected at one point in time.

Introduction

The first recorded mathematical model describing a communicable disease was constructed by the Swiss mathematician Daniel Bernoulli and read at the Royal Academy of Sciences in Paris in 1760. His model aimed to evaluate the impact on human life expectancy at birth if smallpox were to be eliminated as a cause of death through the use of variolation: the practice of deliberately infecting individuals with a mild form of smallpox in order to induce immunity to the disease. Bernoulli's work was used to inform the sale of annuities and so had an immediate economic impact.

The model constructed by Bernoulli assumed that the instantaneous probability of infection, or force of infection, remained constant over time and so was, what is now termed, a static model. This approach to infectious disease modeling, using a static force of infection, remained the norm in modeling for cost-effectiveness analysis until the turn of the twenty-first century.

Dynamic epidemiological modeling of communicable disease transmission started in 1906, when William Hamer, working on childhood infections including measles, postulated that the course of an epidemic depends on the rate of contact between susceptible and infectious individuals, defining the so-called 'mass action' principle of transmission for directly transmitted viral and bacterial infections. In doing so, he removed the assumption of a static force of infection and laid the foundations of modern transmission modeling.

In 1908, Hamer's initial discrete-time model was translated into a continuous time framework by Ronald Ross, who received the Nobel Prize in 1902 for identifying mosquitoes as the vector transmitting malaria. His work was further developed by Kermack and McKendrick, who, in 1927, recognized that a threshold population density was required before an epidemic could take place. The critical elements were now in place for the development of the models used today.

The first landmark textbook on mathematical modeling of epidemiological systems was published by Norman T. Bailey in 1975 and led to the recognition of the importance of epidemiological modeling in public health decision making.

There were, however, still two separate disciplines informing public health policymaking: health economics and epidemiology. Policymakers inevitably have to make decisions about fair and efficient allocation of limited resources and, as such, economic modeling, and cost-effectiveness models in particular, are of critical importance. Unfortunately, when analyzing interventions that targeted communicable diseases, most cost-effectiveness models were static in nature and ignored the developments in dynamic modeling that flowed from the foundations outlined above. At the same time, dynamic transmission models largely ignored the economic aspects of disease control.

The first model to bring these two schools together was published in 1994 by Rowley and Anderson and sought to model the impact and cost-effectiveness of HIV prevention efforts.

Over the past 20 years the fields of dynamic communicable disease modeling and cost-effectiveness modeling have developed rapidly and, when combined, are an indispensable tool used to inform health technology assessments and the formulation of public health policy to control these diseases.

This article is aimed at health economists, who would like an introduction to dynamic infectious disease modeling.

Communicable diseases are each caused by a pathogen, transmitted from one individual to another in whom they may or may not cause clinical symptoms. Such pathogens are typically bacteria (*Salmonella*), viruses (influenza), fungi (*Aspergillus*), protozoa (malaria), or prions (bovine spongiform encephalopathy) and exhibit a wide range of natural histories.

The specific details of the biological interaction between a pathogen and its host are fundamental to its epidemiology at the population level. The site of infection may influence the route of transmission, examples being direct airborne transmission between individuals, contaminative transmission via the fecal–oral route, or sexual transmission.

The site of infection also influences the host's ability to mount an effective immune response. Replication within sites that are not easily reached by the immune system is one way that pathogens, such as the herpes viruses responsible for cold sores, genital herpes, chicken pox, and shingles, can remain latent for decades. Such 'immunologically privileged' sites include cells of the nervous system and to a lesser degree the external mucosal surfaces in the nose. The latter are exploited

 doi:10.1016/B978-0-12-375678-7.01402-4

by the numerous rhinoviruses that collectively cause the common cold.

So what are the key features that set communicable diseases apart from noncommunicable conditions such as heart disease and why do they need special consideration in economic analyses? To illustrate some of these features, the rest of this article will focus on directly transmitted airborne infections.

Direct Airborne Transmission

With directly transmitted airborne pathogens, the hosts typically experience a period of immunological naïvety before they first become infected (Figure 1). This naïve period is typically measured in years. Such susceptible individuals may then become infected, after which there is a delay before becoming infectious, while the pathogen replicates to sufficiently high levels. These exposed, but as yet not infectious, individuals are said to be latently infected and may remain so for days (influenza) to years (tuberculosis), depending on the pathogen. A distinction should be noted between this latent period and an incubation period, the latter being the time from infection to the development of clinical symptoms. An individual may remain infectious for days (rhinovirus) to years (tuberculosis). A proportion of those infected may die; those that survive either remain infected or recover, often with the development of pathogen specific immunity that typically lasts for decades.

The rate at which individuals transition from one state to another dictates the dynamic pattern of temporal change in the prevalence and incidence of infection in a population.

Pathogen transmission is dependent on both biological factors, as described in the Section Introduction, and on the behavior of the host. Host behavior will influence the probability of a susceptible and an infectious individual coming into contact, whereas pathogen and host biology dictate the probability of such a contact resulting in the successful transmission of the pathogen.

The probability of meeting an infectious individual is in part dependent on the number of infectious individuals in the community, a number that is likely to change over time as susceptible individuals are infected and in turn become infectious, recover, or die. This feedback in the risk of infection is a key characteristic of communicable diseases and one of the principal features that distinguishes them from noncommunicable diseases. Feedback produces nonlinear interactions allowing the possibility for small interventions to have large, possibly counterintuitive outcomes and for different pathogens to exhibit a rich diversity of dynamic patterns of infection.

The basic reproductive number (R_0) is a pivotal concept in infectious disease epidemiology, and is defined as the number of secondary infectious cases arising from an average primary infectious case, in a totally susceptible population. In other words, if you start with a population in which no one is immune to a particular infection and a single infectious individual is introduced, how many people will they infect who themselves go on to become infectious.

Processes that drive the transmission dynamics of infectious diseases can be broadly divided into those factors which allow a disease to invade a population and those that enable it to persist there.

When an infection is introduced into an entirely susceptible population the occurrence, or not, of an epidemic depends on the basic reproductive number (R_0). If R_0 is greater than 1 then the number of infections can increase and an epidemic may ensue. If it is less than 1, then the infection is destined to die out.

Directly transmitted infections that induce long-lasting immunity in those that recover are responsible for many of the classic epidemics, characterized by a wave of cases. The chance of any one individual becoming infected will change over the

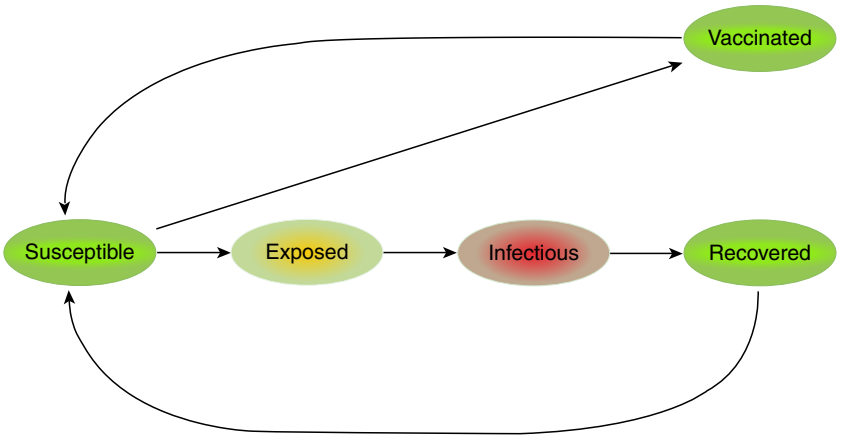


Figure 1 States of host infection and immunity for a typical directly transmitted airborne pathogen. The period of immunological naïvety typically last years to the point of first infection. Once infected, the exposed host may take days to years to become infectious, depending on the pathogen. An individual may then be infectious for days to years during which time a proportion may die; those that survive either remain infected or recover, often with the development of pathogen specific immunity that typically lasts for decades. Reproduced from Figure 2 in Pitman, R., White, L. and Sculpher, M. (2012). Estimating the clinical impact of introducing paediatric influenza vaccination in England and Wales. *Vaccine* **30**, 1208–1224.

course of such an epidemic wave. At the start, with an entirely susceptible population, any encounter with a newly arrived infectious case has the potential to result in transmission of the infection. The spread of the infection is therefore initially dependent on there being sufficient encounters which, on average, result in transmission, such that R_0 is greater than 1. This is easier to achieve in large, dense populations.

As the wave of infection sweeps through the population, infected individuals recover and are immune to further infection. As the proportion of immune individuals in a population increases, a diminishing percentage of people encountered by infectious cases will be susceptible to infection. As the epidemic wave reaches its peak, the average number of secondary infections per infectious individual falls to 1. As the population is no longer fully susceptible, this measure is known as the effective reproductive number (R). Ongoing transmission continues to deplete the susceptible pool, such that R falls below 1 and the number of infectious hosts starts to decline (Figure 2). If the number of infectious individuals is not to continue to decline, the pool of susceptibles must be replenished sufficiently quickly to maintain R at or above 1. This may be via the birth of new individuals, immigration, and the waning of acquired immunity over time. Persistence of an infection (endemicity) is therefore more likely in populations with high birth rates. Conversely, many common infections are absent from small isolated communities as birth rates are too low to supply new susceptibles sufficiently quickly for these infections to remain endemic.

Provided there is sufficient replenishment of susceptibles and all external factors remain constant, the number of infections will settle to a stable endemic state at a constant prevalence that corresponds to an effective reproductive number equal to 1. A consequence of this is that the proportion of the population that is naturally immune will also settle to a constant value that crucially is less than 100%. This is sometimes referred to as the critical proportion immune. Should the number of infections rise for any reason then infectious cases will be generated at a faster rate resulting in an increase in the proportion immune and a decline in the number of susceptibles, which will in turn downregulate the rate of infection, bringing the prevalence of infections back down to its equilibrium state. The converse is true should the number of infections fall.

Vaccination

The aim of vaccination is firstly to protect the vaccinee against infection; however, given sufficiently high uptake, vaccination also benefits the wider, unvaccinated population. This is a consequence of immunized individuals blocking chains of transmission sufficiently often to reduce the number of new infectious individuals produced by each infectious case. Vaccination, therefore, reduces the probability of encountering an infectious individual, thereby helping to protect the whole population. This population-wide protection is known as herd immunity and is the reason an infection may be eliminated from a population without having to vaccinate everyone.

Vaccination that immunizes a proportion of the population may also affect the temporal dynamics of an infection. This may be observed when a program that utilizes a new vaccine is first introduced into a population (Figure 3). Before the introduction of vaccination, immunity is naturally acquired by infection. When a vaccination program is introduced, vaccine-derived immunity supplements this preexisting naturally acquired immunity; the proportion immune now exceeds the equilibrium critical proportion leading to a fall in the incidence of infection and a decline in prevalence. With fewer infectious individuals to transmit the pathogen, the proportion of the population with naturally acquired immunity falls, reducing the overall proportion protected back down to below the critical proportion. This reduction in the proportion immune allows the rate of infection to increase again, leading to a partial rebound in the numbers infected and restoration of the critical

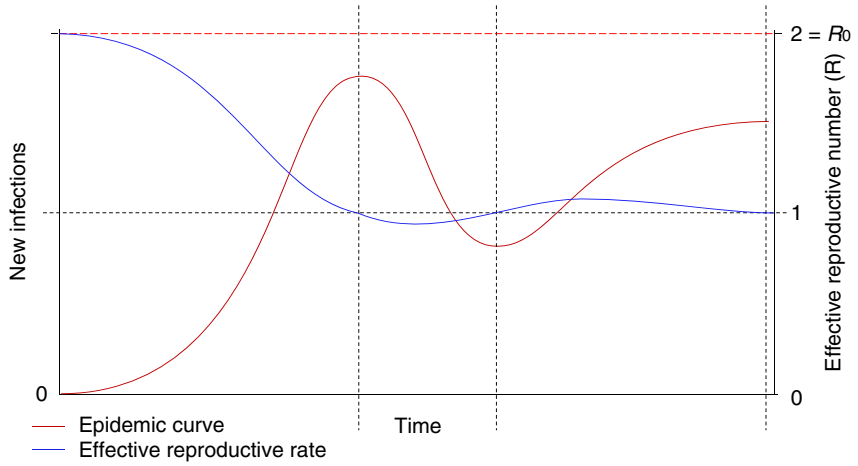


Figure 2 Schematic representation of the relationship between the incidence of infection and the effective reproductive number (R), see text for definitions. If an infectious case arrives in a totally susceptible population, provided R_0 is above 1, the infection will start to spread. As the proportion of the population susceptible starts to fall, R falls to 1, at which point incidence levels off. Continuing transmission further decreases the proportion susceptible, supressing R below 1 and reducing the incidence rate to below the rate at which new susceptible are generated, allowing the proportion susceptible to increase again. As a result, R increases to exceed 1 allowing the incidence rate to recover. If all else remains constant, R will eventually equilibrate to 1 and incidence will settle to a constant rate, equal to the rate of replenishment of susceptibles.

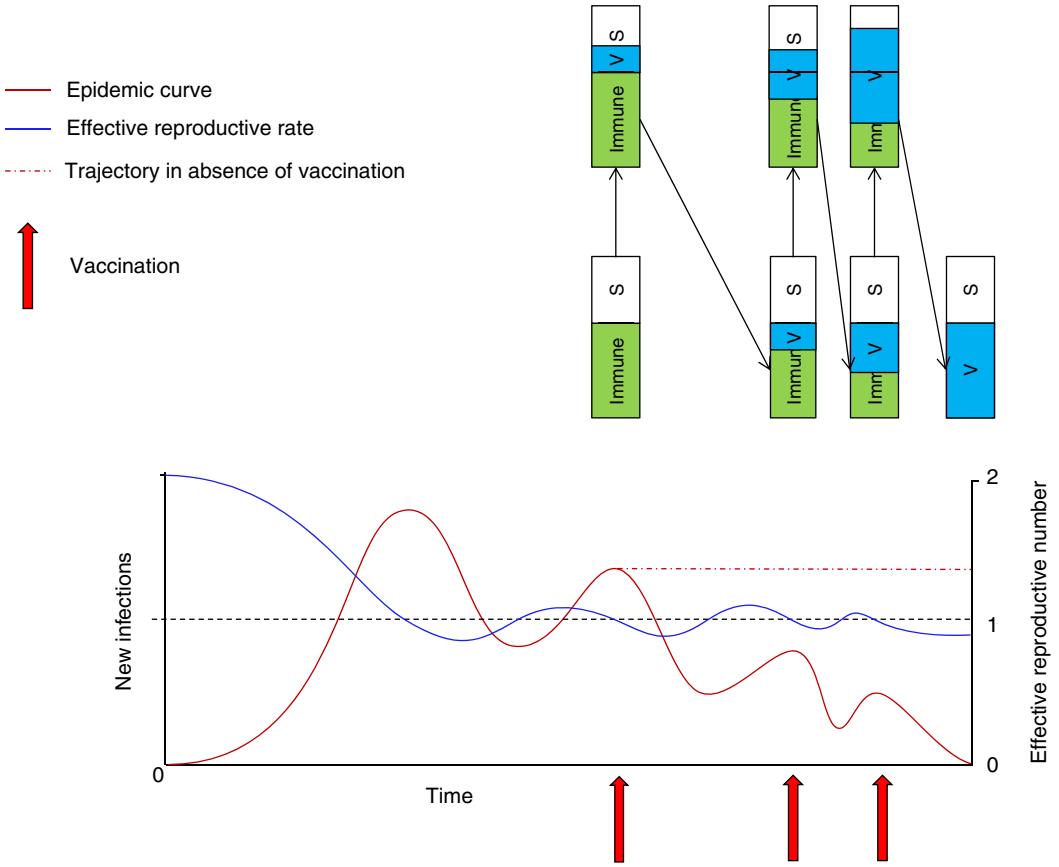


Figure 3 The effect of vaccination on a hypothetical directly transmitted infection. Rectangles represent the total population, green area is the proportion with naturally acquired immunity, blue have vaccine acquired immunity, white are susceptible. Rectangles align with the relevant point on the time axis. When in a stable endemic state, in an unvaccinated population, a constant proportion of the population is immune following a natural infection. Vaccination supplements this equilibrium proportion immune, reducing incidence and with it the proportion acquiring immunity through natural infection, until the equilibrium proportion immune is restored. The effective immunization of a proportion equal to the equilibrium proportion immune leads to local elimination.

proportion immune. The resulting transient low prevalence following program initiation is a well recognized phenomenon known as the honeymoon period.

Should a sufficiently large proportion of the population be vaccinated to account for the entire critical proportion, then local elimination of a pathogen may be achieved. Consequently, the critical proportion immune is also known as the critical proportion to vaccinate (V_c). In a randomly mixing (homogeneous) population, this is defined as $V_c = 1 - 1/R_0$.

Endemic persistence of an infection within a population is therefore dependent on the balance between the generation of immunity, resulting either from pathogen spread or from vaccination, and replenishment of susceptibles as a result of the loss of effective immunity, births, and immigration (Figure 4).

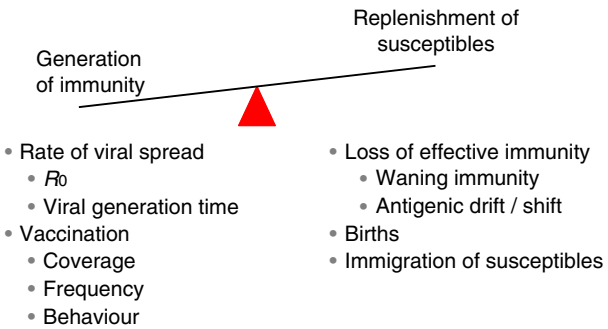


Figure 4 Endemic persistence of an infection is dependent on there being a balance between the rate at which immunity is generated and the rate of replenishment of susceptibles.

An Example Transmission Model

One way to simulate the flow of individuals between each of the stages of infection and immunity outlined above is to compartmentalize the population into corresponding subgroups (susceptible, exposed, infectious, recovered, and vaccinated). Movement between these compartments, including the dynamics of viral transmission, progression, and recovery, may then be described by the following set of linked

differential equations, for $a=0, 1, 2, \ldots, 100$ years of age:

$$\frac{dS_a}{dt} = \Lambda_a + \omega_v V_a(t) + \omega_i R_a(t) - S_a(t)[\mu_a + \psi_a + \lambda_a(t)]$$

$$\frac{dE_a}{dt} = \lambda_a(t) S_a(t) - E_a(t)[\mu_a + \gamma]$$

$$\frac{dI_a}{dt} = \gamma E_a(t) - I_a(t)[\mu_a + \rho]$$

$$\frac{dR_a}{dt} = \rho I_a(t) - R_a(t)[\mu_a + \omega_i]$$

$$\frac{dV_a}{dt} = \psi_a S_a(t) - V_a(t)[\mu_a + \omega_v]$$

where ω_v and ω_i are the rate of loss of vaccine induced and naturally acquired immunity, respectively. The natural death rate is given by μ_a, the average latent period by $1/\gamma$, and the mean duration of infectiousness as $1/\rho$. The age-dependent vaccination rate is signified by ψ_a and $\lambda_a(t)$ represents the age-dependent force of infection in the model:

$$\lambda_a(t) = \sum_{a'} \beta_{a,a'} I_{a'}(t)$$

where $\beta_{a,a'}$ is the transmission coefficient describing the rate of contact and per contact probability of transmission from individuals of age a' to those of age a and

$$\Lambda_a = \begin{cases} \text{birth rate}, & a = 0 \\ 0, & a > 0 \end{cases}$$

To arrive at an expression for R_0, first note that the incidence of infection at age a ($\zeta_a(t)$) is a function of both the force of infection and prevalence of susceptible hosts of age a:

$$\zeta_a(t) = \lambda_a(t) S_a(t)$$

This may be written in the form

$$\zeta_a(t) = \left(\sum\nolimits_{a'} \beta_{a,a'} I_{a'}(t)\right) S_a(t)$$

Now consider the simplified situation where age is ignored and the population assumed to mix homogeneously. Recalling the definition of R_0 as the number of secondary infectious hosts arising from one primary infectious host, in an entirely susceptible population:

$$S = N$$

$$I = 1$$

where N is the total population size. The basic reproductive number may now be expressed in the following form:

$$R_0 = \beta N D q$$

where D is the duration of infectiousness and q is the proportion of infections that become infectious. Note that for the model outlined above

$$D = \frac{1}{(\mu + \rho)}$$

$$q\Big| = \frac{\gamma}{(\mu + \gamma)}$$

This expression of R_0 may be adapted to give the number of infectious hosts of a particular age, arising from infectious individuals of the same or a different age and is usually expressed in matrix form, using the same notation $R_{a,a'}$ as for $\beta_{a,a'}$ above:

$$\begin{pmatrix} R_{0,0} & \cdots & R_{0,100} \\ & \ddots & \\ R_{100,0} & \cdots & R_{100,100} \end{pmatrix}$$

$$= \begin{pmatrix} \beta_{0,0} N_0 D_0 q_0 & \cdots & \beta_{0,100} N_0 D_{100} q_0 \\ & \ddots & \\ \beta_{100,0} N_{100} D_0 q_{100} & \cdots & \beta_{100,100} N_{100} D_{100} q_{100} \end{pmatrix}$$

This matrix is known as the 'next generation matrix,' $\mathbf{M}$, in which

$$D_{a'} = \frac{1}{(\mu_{a'} + \rho)}$$

$$q_a = \frac{\gamma}{(\mu_a + \gamma)}$$

The basic reproductive number for an age structured population may be calculated as the dominant eigenvalue of the next generation matrix, that is to say it is equal to the largest value of R_0 that satisfies the following equation:

$$\det|\mathbf{M} - R_0 \mathbf{I}| = 0$$

where $\mathbf{I}$ is the identity matrix.

Toward Further Realism

All models are, to a greater or lesser degree, caricatures of the real world. To be useful, such caricatures need to capture the essential details of the system being modeled. Models should therefore only be as complex as is required to address the question being asked. Unnecessary complexity reduces the transparency of a model and increases the number of parameters that must be estimated, each of which bringing with it its own uncertainty. The unnecessary proliferation of parameters also makes it harder to decide which model best fits any observed data that may be available. A model should therefore also only be as complex as can be supported by the available data.

Additional complexity may be justified, for example, where certain subgroups of the population need to be accounted for, such as the important risk groups that have a strong influence on the transmission dynamics of a pathogen or where certain types of behavior are similarly important. An example of the latter is the willingness of different subgroups to be vaccinated.

In temperate climates, certain directly transmitted infections, such as influenza, show a strong seasonal variation in incidence, tending to circulate more easily during the winter months. Although the precise reasons for this remain unclear, low temperatures that extend the time exhaled droplets take to evaporate and increased periods of time spent in poorly ventilated congregate settings have both been implicated.

One way to capture such phenomena is to utilize a periodic function such as a sine wave to emulate the seasonal fluctuation in the force of infection. Using the same notation as employed in the example model above, the force of infection may now be expressed in the following way:

$$\lambda_a(t) = z(t)\sum_{a'}\beta_{a,a'}I_{a'}(t)$$

where $z(t)$ is the sine wave function;

$$z(t) = 1 + h.\sin\left(\frac{2\pi(t-f)}{365}\right)$$

t being the number of days since the start of the simulation, whereas h controls the amplitude, and f the phase of the wave.

Choice of Model

The population-based approach to modeling communicable diseases, outlined above, is the method of choice when dealing with common infections, transmitted in large populations. In these circumstances it is a set of population averages that are being modeled and numbers are sufficiently large that they are not significantly affected by chance variations at the individual level. However, where populations are small or an infection is rare, such as at the very start or end of an epidemic, these chance, or stochastic, variations may have a profound effect on the course of events. Epidemics may fail to take off or may simply burn out due to chance. In these circumstances, individual-based models should be used.

Individual or agent-based models simulate each person in a population, recording for each of them their current state with regards to age, infection, immunity etc. Such models are well suited for simulating stochastic events and can capture a greater level of population heterogeneity than can population models; however, this flexibility comes at a cost. When used to simulate even moderately large populations, they have a high computational overhead necessitating the use of high-performance computers with memory capacity measured in terabytes. Consequently, it is often not possible for such models to simulate transmission over more than a single year, which has implications for the time horizon of an analysis.

Burden of Disease

Once an individual is infected, numerous factors may influence whether or not they develop disease, such as their age, physical fitness, and the presence of any comorbidities. The relationship between age and morbidity is of particular importance, as any change in the probability of infection can have an impact on the average age at which individuals first become infected. Widespread vaccination can reduce the prevalence of infection in a population and with it, the probability of encountering an infectious individual, leading to an increase in the average age of first infection. Very young babies may benefit from such a shift, as they are often at an increased risk of more serious disease; however, more perverse outcomes are also possible. If the average age of infection moves into the childbearing ages, this can have disastrous consequences with pathogens such as rubella, where the pathogen poses a significant risk if contracted during pregnancy. In such cases, high levels of vaccine coverage in the wider population must be maintained to produce a net reduction in the risk of morbidity in vulnerable age groups. Alternatively, vulnerable age groups can be targeted for vaccination.

To move from a model of infection incidence to one that captures disease burden, the age stratified probabilities of developing disease, given infection, need to be estimated by dividing the incidence of disease outcome by the incidence of infection, over a defined period of time. Disease outcomes of interest may include primary care consultations, outpatient visits, hospitalizations and death.

Cost-Effectiveness Analysis

Once the dynamic aspects of transmission have been accounted for, the cost-effectiveness analysis of interventions that target communicable diseases is conducted in much the same way as for any other intervention. There are, however, a few areas in which special consideration is required, particularly with regard to the time horizon of the analysis and the implementation of discounting.

Economic analyses recognize the fact that individuals prefer to receive the benefits of an intervention immediately and to defer the costs incurred till later. Such 'time preference' is the reason for discounting future costs and benefits, but can raise difficult questions when applied to public health interventions such as vaccination program that target communicable diseases. Such programs typically incur large upfront costs but accrue benefits over a much longer time scale.

As an example, certain strains of human papillomavirus (HPV) can cause genital warts relatively soon after infection, whereas other strains can induce cervical cancer typically decades later. Applying a standard discounting approach to the cost-effectiveness analysis of HPV vaccination, where costs and benefits receive equal discounting, would only account for the benefits of preventing genital warts, a condition that is relatively easily treated. The lifesaving benefits of preventing cervical cancer would be largely discounted away. Such considerations have led to a widespread debate over the most appropriate approach to discounting, a debate that is as yet unresolved. The reason for this indecision lies more in the lack of information on social attitudes to the value of public health interventions of this nature than in the technical challenges of constructing an appropriate model formulation.

A related challenge concerns the choice of time horizon: how far into the future should costs and benefits be counted in a cost-effectiveness analysis. Again, the upfront costs and deferred benefits of vaccination raise an issue. With an

ongoing vaccination program, any choice of time horizon will result in those individuals being vaccinated toward the end of the simulation accruing all of the costs associated with vaccination, but none of the benefits.

One solution is to extend the time horizon to the point where discounting renders further increases in costs and benefits insignificant. However, this approach is still problematic in circumstances such as those described above for HPV vaccination, where benefits accrue decades after vaccination. The uncertainty in projecting transmission dynamics so far into the future is also a potential hindrance to this approach.

Despite these challenges, numerous successful cost-effectiveness analyses have been conducted on interventions targeting communicable diseases, including vaccines to prevent influenza, pneumococcal disease, HPV, meningococcal group C (MenC), and varicella-zoster.

Growing interest in this field has also led to the publication of guidelines covering various aspects of communicable disease modeling (see Further Reading).

Further Reading

Historical

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its application*, 2nd ed. New York: Hafner.

Bartlett, M. (1949). Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)* **11**, 211–229.

Hamer, W. (1928). *Epidemiology old and new. London.* Kegan Paul.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of Royal Society* **115**, 700–721.

Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. Part I. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **92**(638), 204–226, doi:10.1098/rspa.1916.0007.

Ross, R. and Hudson, H. P. (1917a). An application of the theory of probabilities to the study of a priori pathometry. Part II. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **93**(650), 212, doi:10.1098/rspa.1917.0014.

Ross, R. and Hudson, H. P. (1917b). An Application of the theory of probabilities to the study of a priori pathometry. Part III. *Proceedings of the Royal Society B: Biological Sciences* **89**(621), 507, doi:10.1098/rspb.1917.0008.

Rowley, J. T. and Anderson, R. M. (1994). Modeling the impact and cost-effectiveness of HIV prevention efforts. *AIDS* **8**, 539–548.

General

Anderson, R. and May, R. (1991). *Infectious diseases of humans: Dynamics and control.* Oxford, New York, Tokyo: Oxford University Press.

Anderson, R. M. and May, R. M. (1979). Population biology of infectious diseases: Part I. *Nature* **280**, 361–367.

May, R. M. and Anderson, R. M. (1979). Population biology of infectious diseases: Part II. *Nature* **280**, 455–461.

Porta, M. and Last, J. M. (2008). *A dictionary of epidemiology*, 5th ed. Oxford, New York: Oxford University Press.

Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling.* Oxford, New York: Oxford University Press.

Guidelines

Beutels, P., et al. (2002). Economic evaluation of vaccination programmes: A consensus statement focusing on viral hepatitis. *Pharmacoeconomics* **20**, 1–7.

Jit, M. and Brisson, M. (2011). Modelling the epidemiology of infectious diseases for decision analysis: A primer. *Pharmacoeconomics* **29**, 371–386.

Pitman, R., Fisman, D., Zaric, G. S., et al. (2012). Dynamic transmission modeling: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5. *Medical Decision Making* **32**, 712–721.

Walker, D. G., Hutubessy, R. and Beutels, P. (2010). WHO guide for standardisation of economic evaluations of immunization programmes. *Vaccine* **28**, 2356–2359.

Methodological

Bilcke, J., Beutels, P., Brisson, M. and Jit, M. (2011). Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: A practical guide. *Medical Decision Making* **31**, 675–692.

Bos, J. M., Beutels, P., Annemans, L. and Postma, M. J. (2004). Valuing prevention through economic evaluation: Some considerations regarding the choice of discount model for health effects with focus on infectious diseases. *Pharmacoeconomics* **22**, 1171–1179.

Brisson, M. and Edmunds, W. J. (2003). Economic evaluation of vaccination programs: The impact of herd-immunity. *Medical Decision Making* **23**, 76–82.

Brisson, M. and Edmunds, W. J. (2006). Impact of model, methodological, and parameter uncertainty in the economic analysis of vaccination programs. *Medical Decision Making* **26**, 434–446.

Melegaro, A., Jit, M., Gay, N., Zagheni, E. and Edmunds, W. J. (2011). What types of contacts are important for the spread of infections?: Using contact survey data to explore European mixing patterns. *Epidemics* **3**, 143–151.

Mossong, J., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* **5**, e74.

Westra, T. A., et al. (2012). On discounting of health gains from human papillomavirus vaccination: Effects of different approaches. *Value Health* **15**, 562–567.

Specific diseases

Baguelin, M., Jit, M., Miller, E. and Edmunds, W. J. (2012). Health and economic impact of the seasonal influenza vaccination programme in England. *Vaccine* **30**, 3459–3462.

Brisson, M., de Velde, N. V., Wals, P. D. and Boily, M.-C. (2007). The potential cost-effectiveness of prophylactic human papillomavirus vaccines in Canada. *Vaccine* **25**, 5399–5408.

Choi, Y. H., Jit, M., Flasche, S., Gay, N. and Miller, E. (2012). Mathematical modelling long-term effects of replacing Prevnar7 with Prevnar13 on invasive pneumococcal diseases in England and Wales. *PLoS One* **7**, e39927.

Effelterre, T. V., et al. (2010). A dynamic model of pneumococcal infection in the United States: Implications for prevention through vaccination. *Vaccine* **28**, 3650–3660.

Jit, M., Chapman, R., Hughes, O. and Choi, Y. H. (2011). Comparing bivalent and quadrivalent human papillomavirus vaccines: Economic evaluation based on transmission model. *British Medical Journal* **343**, d5775.

Melegaro, A., et al. (2010). Dynamic models of pneumococcal carriage and the impact of the Heptavalent Pneumococcal Conjugate Vaccine on invasive pneumococcal disease. *BMC Infectious Disease* **10**, 90.

Pitman, R., White, L. and Sculpher, M. (2012). Estimating the clinical impact of introducing paediatric influenza vaccination in England and Wales. *Vaccine* **30**, 1208–1224.

Pitman, R. J., Nagy, L. D. and Sculpher, M. J. (2013). Cost-effectiveness of childhood influenza vaccination in England and Wales: Results from a dynamic transmission model. *Vaccine* **31**, 927–942.

Vynnycky, E., Pitman, R., Siddiqui, R., Gay, N. and Edmunds, W. J. (2008). Estimating the impact of childhood influenza vaccination programmes in England and Wales. *Vaccine* **26**, 5321–5330.

Wright, T. C., et al. (2006). Chapter 30: HPV vaccines and screening in the prevention of cervical cancer; conclusions from a 2006 workshop of international experts. *Vaccine* **24**(supplement 3), S251–S261.

Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap

AC Cameron, University of California – Davis, Davis, CA, USA

Glossary

False discovery proportion (FDP) The proportion of incorrectly rejected hypotheses.

False discovery rate (FDR) The expectation of the proportion of incorrectly rejected hypotheses.

Family-wise error rate (FWER) The probability of finding statistical significance in at least one test.

This article presents inference for many commonly used estimators – least squares, generalized linear models, generalized method of moments (GMM), and generalized estimating equations – that are asymptotically normally distributed. Section Inference focuses on Wald confidence intervals and hypothesis tests based on estimator variance matrix estimates that are heteroskedastic-robust and, if relevant, cluster-robust. Section Model Tests and Diagnostics summarizes tests of model adequacy and model diagnostics. Section Multiple Tests presents family-wise error rates and false discovery rates (FDRs) that control for multiple testing such as subgroup analysis. Section Bootstrap and Other Resampling Methods presents bootstrap and other resampling methods that are most often used to estimate the variance of an estimator. Bootstraps with asymptotic refinement are also presented.

Inference

Most estimators in health applications are m-estimators that solve estimating equations of the form

$$\sum_{i=1}^{N} g_i(\hat{\theta}) = 0 \qquad [1]$$

where θ is a $q \times 1$ parameter vector, i denotes the ith of N observations, $g_i(\cdot)$ is a $q \times 1$ vector, and often $g_i(\theta) = g_i(y_i, x_i, \theta)$ where y denotes a scalar-dependent variable and x denotes the regressors or covariates. For ordinary least squares, for example, $g_i(\beta) = (y_i - x_i'\beta)x_i$. Nonlinear least squares, maximum likelihood (ML), quantile regression, and just-identified instrumental variables estimators are m-estimators. So too are generalized linear model estimators, extensively used in biostatistics, that are quasi-ML estimators based on exponential family distributions, notably Bernoulli (logit and probit), binomial, gamma, normal, and Poisson.

The estimator $\hat{\theta}$ is generally consistent if $E[g_i(\theta) = 0]$. Statistical inference is based on the result that $\hat{\theta}$ is asymptotically normal with mean θ and variance matrix $V[\hat{\theta}]$ that is estimated by

$$\hat{V}[\hat{\theta}] = \hat{A}^{-1}\hat{B}\hat{A}^{-1'} \qquad [2]$$

where $N^{-1}\hat{A}$ and $N^{-1}\hat{B}$ are consistent estimates of $A = E\left[N^{-1}\sum_i H_i(\theta)\right]$, where $H_i(\theta) = \partial g_i(\theta)/\partial\theta'$ and $B = E[N^{-1}\sum_i\sum_j g_i(\theta)g_j(\theta)']$. The variance is said to be of 'sandwich form,' because $\hat{B}$ is sandwiched between $\hat{A}^{-1}$ and $\hat{A}^{-1'}$. The estimate $\hat{A}$ is the observed Hessian $\sum_i H_i(\hat{\theta})$, or in some cases the expected Hessian $E\left[\sum_i H_i(\theta)\right]\big|_{\hat{\theta}}$. By contrast, the estimate $\hat{B}$, and hence $\hat{V}[\hat{\theta}]$ in eqn [2], can vary greatly with the type of data being analyzed and the associated appropriate distributional assumptions.

Default estimates of $V[\hat{\theta}]$ are based on strong distributional assumptions, and are typically not used in practice. For ML estimation with density assumed to be correctly specified $B = -A$, so the sandwich estimate simplifies to $\hat{V}[\hat{\theta}] = -\hat{A}^{-1}$. Qualitatively similar simplification occurs for least squares and instrumental variables estimators when model errors are independent and homoskedastic.

More generally, for data independent over i, $\hat{B} = \frac{N}{N-q}\sum_i g_i(\hat{\theta})g_i(\hat{\theta})'$, where the multiple $N/(N-q)$ is a commonly used finite sample adjustment. Then the variance matrix estimate in eqn [2] is called the Huber, White, or robust estimate – a limited form of robustness as independence of observations is assumed. For OLS, for example, this estimate is valid even if independent errors are heteroskedastic, whereas the default requires errors to be homoskedastic.

Often data are clustered, with observations correlated within a cluster but independent across clusters. For example, individuals may be clustered within villages or hospitals, or students clustered within class or within school. Let c denote the typical cluster, and sum $g_i(\theta)$ for observations i in cluster c to form $g_c(\theta)$. Then $\hat{B} = \frac{C}{C-1}\sum_{c=1}^{C} g_c(\hat{\theta})g_c(\hat{\theta})'$, where C is the number of clusters, and the variance matrix estimate in eqn [2] is called a cluster-robust estimate. The number of clusters should be large as the asymptotic theory requires $C \to \infty$, rather than $N \to \infty$. The clustered case also covers short panels with few time periods and data correlated over time for a given individual but independent across individuals. Then the clustering sums over time periods for a given individual. Wooldridge (2003) and Cameron and Miller (2011) survey inference with clustered data.

Survey design can lead to clustering. Applied biostatisticians often use survey estimation methods that explicitly control for the three complex survey complications of weighting, stratification, and clustering. Econometricians instead usually assume correct model specification conditional on regressors (or instruments), so that there is no need to weight; ignore the potential reduction in standard error estimates that can occur with stratification; and conservatively

 doi:10.1016/B978-0-12-375678-7.00706-9

control for clustering by computing standard errors that cluster at a level such as a state (region) that is usually higher than the primary sampling unit.

For time series data, observations may be correlated over time. Then the heteroskedastic and autocorrelation consistent (HAC) variance matrix estimate is used; see Newey and West (1987). A similar estimate can be used when data are spatially correlated, with correlation depending on the distance and with independence once observations are more than a given distance apart. This leads to the spatial HAC estimate; see Conley (1999).

Note that in settings where robust variance matrix estimates are used, additional assumptions may enable more efficient estimation of θ such as feasible generalized least squares and generalized estimating equations, especially if data are clustered.

Given $\hat{\theta}$ asymptotic normal with variance matrix estimated using eqn [2], the Wald method can be used to form confidence intervals and perform hypothesis tests.

Let θ be a scalar component of the parameter vector θ. Since $\hat{\theta} \overset{a}{\sim} N[\theta, \hat{\mathrm{V}}[\hat{\theta}]]$, we have $\hat{\theta} \overset{a}{\sim} N[\theta, s^2_{\hat{\theta}}]$, where the standard error $s_{\hat{\theta}}$ is the square root of the relevant diagonal entry in $\hat{V}[\hat{\theta}]$. It follows that $(\hat{\theta} - \theta)/s_{\hat{\theta}} \overset{a}{\sim} N[0,1]$. This justifies the use of the standard normal distribution in constructing confidence intervals and hypothesis tests for sample size $N \to \infty$. A commonly used finite-sample adjustment uses $(\hat{\theta} - \theta)/s_{\hat{\theta}} \overset{a}{\sim} T(N - q)$, where $T(N-q)$ is the students T distribution with $(N-q)$ degrees of freedom, N is the sample size, and K parameters are estimated.

A 95% confidence interval for θ gives a range of values that 95% of the time will include the unknown true value of θ . The Wald 95% confidence interval is $\hat{\theta} \pm c_{.025} \times s_{\hat{\theta}}$, where the critical value $c_{.025}$ is either $z_{[.025]} = 1.96$, the .025 quantile of the standard normal distribution, or $t_{[.025]}$ the .025 quantile of the $T(N-q)$ distribution. For example, $c_{.025} = 2.042$ if $N - q = 30$.

For two-sided tests of $H_0{:}\theta = \theta^*$ against $H_a{:}\theta \neq \theta^*$, the Wald test is based on how far $|\hat{\theta} - \theta^*|$ is from zero. On normalizing by the standard error, the Wald statistic $w = (\hat{\theta} - \theta^*)/s_{\hat{\theta}}$ is asymptotically standard normal under H_0, though again a common finite sample correction is to use the $T(N-q)$ distribution. H_0 at the 5% significance level is rejected if $|w| > c_{.025}$. Often $\theta^* = 0$, in which case w is called the t-statistic and the test is called a test of statistical significance. Greater information is conveyed by reporting the p-value, the probability of observing a value of w as large or larger in absolute value under the null hypothesis. Then $p = \Pr[|W| > |w|]$, where W is standard normal or $T(N-q)$ distributed. $H_0{:}\theta = \theta^*$ is rejected against $H_0{:}\theta \neq \theta^*$ at level 0.05 if $p < .05$.

More generally, it may be interesting to perform joint inference on more than one parameter, such as a joint test of statistical significance of several parameters, or on functions(s) of the parameters. Let $\mathrm{h}(\theta)$ be an $h \times 1$ vector function of θ, possibly nonlinear, where $h \leq q$. A Taylor series approximation yields $\mathrm{h}(\hat{\theta}) \simeq \mathrm{h}(\theta) + \hat{\mathrm{R}}(\hat{\theta} - \theta)$, where $\hat{\mathrm{R}} = \partial \mathrm{h}(\theta)/\partial \theta'|_{\hat{\theta}}$ is assumed to be of full rank h (the nonlinear analog of linear dependence of restrictions). Given $\hat{\theta} - \theta \overset{a}{\sim} N[0, \hat{v}[\hat{\theta}]]$, this yields $\mathrm{h}(\hat{\theta}) \overset{a}{\sim} N[\mathrm{h}(\theta), \hat{\mathrm{R}}\hat{\mathrm{V}}[\hat{\theta}]\hat{\mathrm{R}}']$. The term delta method is used as a first derivative and is taken in approximating $\mathrm{h}(\hat{\theta})$.

Confidence intervals can be formed in the case that $h(\cdot)$ is a scalar. Then $h(\hat{\theta}) \pm c_{.025} \times [\hat{R}\hat{V}[\hat{\theta}]\hat{R}']^{1/2}$ is used. A leading example is a confidence interval for a marginal effect in a nonlinear model. For example, for $E[y|x] = \exp(x'\theta)$ the marginal effect for the jth regressor is $\partial \mathrm{E}[y|\mathrm{x}]/\partial x_j = \exp(\mathrm{x}'\theta)\theta_j$. When evaluated at $\mathrm{x} = \mathrm{x}^*$ this equals $\exp(\mathrm{x}^{*\prime}\hat{\theta})\hat{\theta}_j$ which is a scalar function $h(\hat{\theta})$ of $\hat{\theta}$; the corresponding average marginal effect is $\sum_i \exp(\mathrm{x}_i'\hat{\theta})\hat{\theta}_j$.

A Wald test of $H_0{:}\mathrm{h}(\theta) = 0$ against $H_a{:}\mathrm{h}(\theta) \neq 0$ is based on the closeness of $\mathrm{h}(\hat{\theta})$ to zero, using

$$w = \mathrm{h}(\hat{\theta})'[\hat{\mathrm{R}}\hat{\mathrm{V}}[\hat{\theta}]\hat{\mathrm{R}}']^{-1}\mathrm{h}(\hat{\theta}) \overset{a}{\sim} \chi^2(h) \qquad [3]$$

under H_0. H_0 at level 0.05 is rejected if $w > \chi^2_{.95}(h)$. An F version of this test is $F = w/h$, and is rejected at level 0.05 if $w > F_{.95}(h, N-q)$. This is a small sample variation, analogous to using the $T(N-q)$ rather than the standard normal.

For ML estimation the Wald method is one of the three testing methods that may be used. Consider testing the hypothesis that $\mathrm{h}(\theta) = 0$. Let $\tilde{\theta}$ denote the ML estimator obtained by imposing this restriction, whereas $\hat{\theta}$ does not impose the restriction. The Wald test uses only $\hat{\theta}$ and tests the closeness of $h(\hat{\theta})$ to zero. The log-likelihood ratio test is based on the closeness of $L(\hat{\theta})$ to $L(\tilde{\theta})$, where $L(\theta)$ denotes the log-likelihood function. The score test uses only $\tilde{\theta}$ and is based on the closeness to zero of $\partial L(\theta)/\partial \theta|_{\tilde{\theta}}$, where $L(\theta)$ here is the log-likelihood function for the unrestricted model.

If the likelihood function is correctly specified, a necessary assumption, these three tests are asymptotically equivalent. So the choice between them is one of convenience. The Wald test is most often used, as in most cases $\hat{\theta}$ is easily obtained. The score test is used in situations in which estimation is much easier when the restriction is imposed. For example, in a test of no spatial dependence versus spatial dependence, it may be much easier to estimate θ under the null hypothesis of no spatial dependence. The Wald and score tests can be robustified. If one is willing to make the strong assumption that the likelihood function is correctly specified, then the likelihood ratio test is preferred due to the Neyman–Pearson lemma and because, unlike the Wald test, it is invariant to reparameterization.

GMM estimators are based on a moment condition of the form $\mathrm{E}[\mathrm{g}_i(\theta)] = 0$. If there are as many components of $\mathrm{g}(\cdot)$ as of θ the model is said to be just identified and the estimate $\hat{\theta}$ solves $\sum_i \mathrm{g}_i(\hat{\theta}) = 0$, which is eqn [1]. Leading examples in the biostatistics literature are generalized linear model estimators and generalized estimating equations estimators. If instead there are more moment conditions than parameters there is no solution to eqn [1]. Instead make $\sum_i \mathrm{g}_i(\hat{\theta})$ as close to zero as possible using a quadratic norm. The method of moments estimator minimizes

$$Q(\theta) = \left(\sum_{i=1}^{N} g_i(\theta)\right)' \mathrm{W} \left(\sum_{i=1}^{N} \mathrm{g}_i(\theta)\right)$$

where W is a symmetric positive definite weighting matrix and the best choice of W is the inverse of a consistent estimate of the variance of $\sum_i \mathrm{g}_i(\theta)$.

The leading example of this is two-stage least-squares (2SLS) estimation for instrumental variables estimation in overidentified models. Then $\mathrm{g}_i(\beta) = z_i(y_i - \mathrm{x}_i'\beta)$, and it can be shown that the 2SLS estimator is obtained if $\mathrm{W} = (\mathrm{Z}'\mathrm{Z})^{-1}$. The estimated variance matrix is again of sandwich form eqn [2], though the expressions for $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$ are more complicated. For

instrumental variables estimators with instruments weakly correlated with regressors an alternative asymptotic theory may be warranted. Bound *et al.* (1995) outline the issues and Andrews *et al.* (2007) compare several different test procedures.

Model Tests and Diagnostics

The most common specification tests imbed the model under consideration into a larger model and use hypothesis tests (Wald, likelihood ratio, or score) to test the restrictions that the larger model collapses to the model under consideration. A leading example is test of statistical significance of a potential regressor.

A broad class of tests of model adequacy can be constructed by testing the validity of moment conditions that are imposed by a model but have not already been used in constructing the estimator. Suppose a model implies the population moment condition

$$H_0 : \mathrm{E}[\mathrm{m}_i(\mathrm{w}_i, \theta)] = 0 \qquad [4]$$

where w is a vector of observables, usually the dependent variable y, regressors x, and, possibly, additional variables z. An m-test, in the spirit of a Wald test, is a test of whether the corresponding sample moment

$$\hat{\mathrm{m}}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{m}_i(\mathrm{w}_i, \hat{\theta}) \qquad [5]$$

is close to zero. Under suitable assumptions, $\hat{\mathrm{m}}(\hat{\theta})$ is asymptotically normal. This leads to the chi-squared test statistic

$$M = \hat{\mathrm{m}}(\hat{\theta})'\hat{\mathrm{V}}_{\mathrm{m}}^{-1}\hat{\mathrm{m}}(\hat{\theta}) \overset{a}{\sim} \chi^2(rank(V_{\mathrm{m}})) \qquad [6]$$

if the moment conditions eqn [4] are correct, where $\hat{\mathrm{V}}_{\mathrm{m}}$ is a consistent estimate of the asymptotic variance of $\hat{\mathrm{m}}(\hat{\theta})$. The challenge is in obtaining $\hat{\mathrm{V}}_{\mathrm{m}}$. In some leading examples an auxiliary regression can be used, or a bootstrap can be applied.

Especially for fully parametric models there are many candidates for $\mathrm{m}_i(\cdot)$. Examples of this approach are White's information matrix test to test correct specification of the likelihood function; a regression version of the chi-squared goodness of fit test; Hausman tests such as that for regressor endogeneity; and tests of overidentifying restrictions in a model with endogenous regressors and an excess of instruments. Such tests are not as widely used as they might be for two reasons. First, there is usually no explicit alternative hypothesis so rejection of H_0 may not provide much guidance as to how to improve the model. Second, in very large samples with actual data any test at a fixed significance level such as 0.05 is likely to reject the null hypothesis, so inevitably any model will be rejected.

Regression model diagnostics need not involve formal hypothesis tests. A range of residual diagnostic plots can provide information on model nonlinearity and observations that are outliers and have high leverage. In the linear model, a small sample correction divides the residual $y_i - x_i'\hat{\beta}$ by $\sqrt{1-h_{ii}}$, where h_{ii} is the ith diagonal entry in the hat matrix $\mathrm{H}=\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}$. As H has rank K, the number of regressors, the average value of h_{ii} is K/n and values of h_{ii} in excess of $2K/N$ are viewed as having high leverage. This result extends to generalized linear models where a range of residuals have been proposed; McCullagh and Nelder (1989) provide a summary. Econometricians place less emphasis on residual analysis, compared with biostatisticians. If datasets are small then there is concern that residual analysis may lead to overfitting of the model. Besides if the dataset is large then there is a belief that residual analysis may be unnecessary as a single observation will have little impact on the analysis. Even then diagnostics may help detect data miscoding and unaccounted model nonlinearities.

For linear models, R^2 is a well understood measure of goodness of fit. For nonlinear models a range of pseudo-R^2 measures have been proposed. One that is easily interpreted is the squared correlation between y and $\hat{y}$, though in nonlinear models this is not guaranteed to increase as regressors are added.

Model testing and diagnostics may lead to more than one candidate model. Standard hypothesis tests can be implemented for models that are nested. For nonnested models that are likelihood based, one can use a generalization of the likelihood ratio test due to Vuong (1989), or use information criteria such as Akaike's information criteria based on fitted log-likelihood with a penalty for the number of model parameters. For nonnested models that are not likelihood based one possibility is artificial nesting that nests two candidate models in a larger model, though this approach can lead to neither model being favored.

Multiple Tests

Standard theory assumes that hypothesis tests are done once only and in isolation, whereas in practice final reported results may follow much pretesting. Ideally reported p values should control for this pretesting.

In biostatistics, it is common to include as control variables in a regression only those regressors that have $p<.05$. By contrast, in economics it is common to have a preselected candidate set of control regressors, such as key socioeconomic variables, and include them even if they are statistically insignificant. This avoids pretesting, at the expense of estimating larger models.

A more major related issue is that of multiple testing or multiple comparisons. Examples include testing the statistical significance of a key regressor in several subgroups of the sample (subgroup analysis); testing the statistical significance of a key regressor in regressions on a range of outcomes (such as use of a range of health services); testing the statistical significance of a key regressor interacted with various controls (interaction effects); and testing the significance of a wide range of variables on a single outcome (such as various genes on a particular form of cancer). With many such tests at standard significance levels one is clearly likely to find spurious statistical significance.

In such cases one should view the entire battery of tests as a unit. If m such tests are performed, each at statistical significance level α^*, and the tests are statistically independent, then the probability of finding no statistical significance in all m

tests is $(1-\alpha^*)^m$. It follows that the probability of finding statistical significance in at least one test, called the family-wise error rate (FWER), equals $\alpha=1-(1-\alpha^*)^m$. To test at FWER α, each individual test should be at level $\alpha^*=1-(1-\alpha)^m$, called the Sidak correction. For example, if $m=5$ tests are conducted with FWER of $\alpha=0.05$, each test should be conducted at level $\alpha^*=0.01021$. The simpler Bonferroni correction sets $\alpha^*=\alpha/m$. The Holm correction uses a stepdown version of Bonferroni, with tests ordered by p-value from smallest to largest, so $p_{(1)}<p_{(2)}<<p_{(m)}$, and the jth test rejects if $p_{(j)}<\alpha_j^*=\alpha/(m-j+1)$. A stepdown version of the Sidak correction uses $\alpha_j^*=1-(1-\alpha)^{m-j+1}$. These corrections are quite conservative in practice, as the multiple tests are likely to be correlated rather than independent.

Benjamini and Hochberg (1995) proposed an alternative approach to multiple testing. Recall that test size is the probability of a type I error, i.e., the probability of incorrectly rejecting the null hypothesis. For multiple tests it is natural to consider the proportion of incorrectly rejected hypotheses, the false discovery proportion (FDP), and its expectation E[FDP] called the FDR. Benjamini and Hochberg (1995) argue that it is more natural to control FDR than FEWR. They propose doing so by ordering tests by p-value from smallest to largest, so $p_{(1)}<p_{(2)}<\cdots<p_{(m)}$, and rejecting the corresponding hypotheses $H_{(1)},\ldots,H_{(k)}$, where k is the largest j for which $p(j)\leq\alpha j/m$, where α is the prespecified FDR for the multiple tests. If the multiple tests are independent then the FDR equals α.

In practice tests are not independent. Farcomeni (2008) provides an extensive guide to the multiple testing literature. A recent article on estimating the FDR when tests are correlated is Schwartzman and Lin (2011). Duflo *et al.* (2008) provide a good discussion of practical issues that arise with multiple testing and consider the FEWR but not the FDR. White (2001) presents simulation-based methods for the related problem of testing whether the best model encountered in a specification search has a better predictive power than a benchmark model.

Bootstrap and Other Resampling Methods

Statistical inference controls for the uncertainty that the observed sample of size N is just one possible realization of a set of N possible draws from the population. This typically relies on asymptotic theory that leads to limit normal and chi-squared distributions. Alternative methods based on Monte Carlo simulation are detailed in this section.

Bootstrap

Bootstraps can be applied to a wide range of statistics. The most common use of the bootstrap is considered here, to estimate the standard error of an estimator when this is difficult to do using conventional methods.

Suppose 400 random samples from the population were available. Then 400 different estimates of $\hat{\theta}$ can be obtained and the standard error of $\hat{\theta}$ is simply the standard deviation of these 400 estimates. In practice, however, only one sample from the population is available. The bootstrap provides a way to generate 400 samples by resampling from the current sample. Essentially, the observed sample is viewed as the population and the bootstrap provides multiple samples from this population.

Let $\hat{\theta}^*_{(1)},\ldots,\hat{\theta}^*_{(B)}$ denote B estimates where, for example, $B=400$. Then in the scalar case the bootstrap estimate of the variance of $\hat{\theta}$ is

$$\hat{V}_{\text{Boot}}\left[\hat{\theta}\right]=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_{(b)}-\overline{\hat{\theta}^*})^2 \qquad [7]$$

where $\overline{\hat{\theta}^*}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)}$ is the average of the B bootstrap estimates. The square root of $\hat{V}_{\text{Boot}}[\hat{\theta}]$, denoted $\text{se}_{\text{Boot}}[\hat{\theta}]$, is called the bootstrap estimate of the standard error of $\hat{\theta}$. In the case of several parameters

$$\hat{V}_{\text{Boot}}\left[\hat{\theta}\right]=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_{(b)}-\overline{\hat{\theta}^*})(\hat{\theta}^*_{(b)}-\overline{\hat{\theta}^*})'$$

and even more generally the bootstrap may be used to estimate the variance of functions $h(\hat{\theta})$, such as marginal effects, not just $\hat{\theta}$ itself.

There are several different ways that the resamples can be obtained. A key consideration is that the quantity being resampled should be independent and identically distributed (i.i.d).

The most common bootstrap for data (y_i,x_i) that are i.i.d. is a paired bootstrap or nonparametric bootstrap. This draws with replacement from $(y_1,x_1),\ldots,(y_N,x_N)$ to obtain a resample $(y_1^*,x_1^*),\ldots,(y_N^*,x_N^*)$ for which some observations will appear more than once, whereas others will not appear at all. Estimation using the resample yields estimate $\hat{\theta}^*$. Using B similarly generated resamples yields $\hat{\theta}^*_{(1)},\ldots,\hat{\theta}^*_{(B)}$. This bootstrap variance estimate is asymptotically equivalent to the White or Huber robust sandwich estimate.

If data are instead clustered with C clusters, a clustered bootstrap draws with replacement from the entire clusters, yielding a resample $(y_1^*,X_1^*),\ldots,(y_C^*,X_C^*)$. This bootstrap variance estimate is asymptotically equivalent to the cluster-robust sandwich estimate.

Other bootstraps place more structure on the model. A residual or design bootstrap in the linear regression model fixes the regressors and only resamples the residuals. For models with i.i.d. errors the residual bootstrap samples with replacement from $\hat{u}_1,\ldots,\hat{u}_N$ to yield residual resample $\hat{u}_1^*,\ldots,\hat{u}_N^*$. Then the typical data resample is $(y_1^*,x_1),\ldots,(y_N^*,x_N)$ where $y_i^*=x_i'\hat{\beta}+\hat{u}_i^*$. If errors are heteroskedastic one should instead use a wild bootstrap; the simplest example is $\hat{u}_i^*=\hat{u}_i$ with probability .5 and $\hat{u}_i^*=-\hat{u}_i$ with probability .5.

For a fully parameterized model one can generate new values of the dependent variable from the fitted conditional distribution. The typical data resample is $(y_1^*,x_1),\ldots,(y_N^*,x_N)$ where y_i^* is a draw from $F(y|x_i,\hat{\theta})$.

Whenever a bootstrap is used in applied work the seed, the initial value of the random number generator used in determining random draws, should be set to ensure replicability of results. For standard error estimation $B=400$ should be more than adequate.

The bootstrap can also be used for statistical inference. A Wald 95% confidence interval for scalar θ is $\hat{\theta}\pm1.96\times\text{se}_{\text{Boot}}[\hat{\theta}]$. An asymptotically equivalent alternative

interval is the percentile interval $(\hat{\theta}^*_{[.025]}, \hat{\theta}^*_{[.975]})$, where $\hat{\theta}^*_{[\alpha]}$ is the αth quantile of $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$. Similarly, in testing $H_0{:}\theta = 0$ against $H_a{:}\theta \neq 0$ the null hypothesis may be rejected if $|w| = |\hat{\theta}/\text{se}_{\text{Boot}}[\hat{\theta}]| > 1.96$, or if $\hat{\theta} < \hat{\theta}^*_{[.025]}$ or $\hat{\theta} > \hat{\theta}^*_{[.975]}$.

Care is needed in using the bootstrap in nonstandard situations as, for example, $\text{V}[\hat{\theta}]$ may not exist, even asymptotically; yet it is always possible to (erroneously) compute a bootstrap estimate of $\text{V}[\hat{\theta}]$. The bootstrap can be applied if $\hat{\theta}$ is root-N consistent and asymptotically normal, and there is sufficient smoothness in the cumulative distribution functions of the data-generating process and of the statistic being bootstrapped.

Bootstrap with Asymptotic Refinement

The preceding bootstraps are asymptotically equivalent to the conventional methods of section Inference. Bootstraps with asymptotic refinement, by contrast, provide a more refined asymptotic approximation that may lead to better performance (truer test size and confidence interval coverage) in finite samples. Such bootstraps are emphasized in theory papers, but are less often implemented in applied studies.

These gains are possible if the statistic bootstrapped is asymptotically pivotal, meaning its asymptotic distribution does not depend on unknown parameters. An estimator $\hat{\theta}$ that is asymptotically normal is not usually asymptotically pivotal as its distribution depends on an unknown variance parameter. However, the studentized statistic $t = (\hat{\theta} - \theta_0)/s_{\hat{\theta}}$ is asymptotically $N[0,1]$ under $H_0{:}\theta = \theta_0$, so is asymptotically pivotal. Therefore compute $t^* = (\hat{\theta}^* - \hat{\theta})/s_{\hat{\theta}^*}$ for each bootstrap resample and use quantiles of $t^*_{(1)}, \ldots, t^*_{(B)}$ to compute critical values and p-values. Note that t^* is centered around $\hat{\theta}$ because the bootstrap views the sample as the population, so $\hat{\theta}$ is the population value.

A 95% percentile t-confidence interval for scalar θ is $(\hat{\theta} + t^*_{[.025]}s_{\hat{\theta}}, \hat{\theta} + t^*_{[.975]}s_{\hat{\theta}})$, where $t^*_{[\alpha]}$ is the αth quantile of $t^*_{(1)}, \ldots, t^*_{(B)}$. A percentile-t Wald test rejects $H_0{:}\theta = \theta_0$ against $H_a{:}\theta \neq \theta_0$ at level 0.05 if $t = (\hat{\theta} - \theta_0)/s_{\hat{\theta}}$ falls outside the interval $(t^*_{[.025]}, t^*_{[.975]})$.

Two commonly used alternative methods to obtain confidence intervals with asymptotic refinement are the following. The bias-corrected method is a modification of the percentile method that incorporates a bootstrap estimate of the finite-sample bias in $\hat{\theta}$. For example, if the estimator is upward biased, as measured by estimated median bias, then the confidence interval is moved to the left. The bias-corrected accelerated confidence interval is an adjustment to the bias-corrected method that adds an acceleration component that permits the asymptotic variance of $\hat{\theta}$ to vary with θ.

Theory shows that bootstrap methods with asymptotic refinement outperform conventional asymptotic methods as $N \to \infty$. For example, a nominal 95% confidence interval with asymptotic refinement has a coverage rate of $0.95 + O(N^{-1})$ rather than $0.95 + O(N^{-1/2})$. This does not guarantee better performance in typical sized finite samples, but Monte Carlo studies generally confirm this to be the case. Bootstraps with refinement require a larger number of bootstraps than recommended in the previous subsection, as the critical values lie in the tails of the distribution. A common choice is $B = 999$, with B chosen so that $B + 1$ is divisible by the significance level 100α.

Jackknife

The jackknife is an alternative resampling scheme used for bias correction and variance estimation that predates the bootstrap.

Let $\hat{\theta}$ be the original sample estimate of θ, let $\hat{\theta}_{(-i)}$ denote the parameter estimate from the sample with the ith observation deleted, $i = 1, \ldots, N$, and let $\bar{\hat{\theta}} = N^{-1}\sum_{i=1}^{N}\hat{\theta}_{(-i)}$ denote the average of the N jackknife estimates. The bias-corrected jackknife estimate of θ equals $N\hat{\theta} - (N-1)\bar{\hat{\theta}}$, the sum of the N pseudovalues $\hat{\theta}^*_{(-i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(-i)}$ that provide measures of the importance or influence of the ith observation estimating $\hat{\theta}$.

The variance of these N pseudovalues can be used to estimate $\text{V}[\hat{\theta}]$, yielding the leave-one-out jackknife estimate of variance:

$$\hat{\text{V}}_{\text{Jack}}\left[\hat{\theta}\right] = \left[\frac{1}{N(N-1)}\sum_{i=1}^{N}(\hat{\theta}^*_{(-i)} - \bar{\hat{\theta}})(\hat{\theta}^*_{(-i)} - \bar{\hat{\theta}})'\right]$$

A variation replaces $\bar{\hat{\theta}}$ with $\hat{\theta}$.

The jackknife requires N resamples, requiring more computation than the bootstrap if N is large. The jackknife does not depend on random draws, unlike the bootstrap, so is often used to compute standard errors for published official statistics.

Permutation Tests

Permutation tests derive the distribution of a test statistic by obtaining all possible values of the test statistic under appropriate rearrangement of the data under the null hypothesis.

Consider scalar regression, so $y_i = \beta_1 + \beta_2 x_i + u_i$, $i = 1, \ldots, N$, and Wald test of $H_0{:}\beta_2 = 0$ based on $t = \hat{\beta}_2/s_{\hat{\beta}_2}$. Regress each of the $N!$ unique permutations of $(y_1, \ldots, y_N)$ on the regressors $(x_1, \ldots, x_N)$ and in each case calculate the t-statistic for $H_0{:}\beta_2 = 0$. Then the p-value for the original test statistic is obtained directly from the ordered distribution of the $N!$ t-statistics.

Permutation tests are most often used to test whether two samples come from the same distribution, using the difference in means test. This is a special case of the previous example, where x_i is an indicator variable equal to one for observations coming from the second sample.

Permutation methods are seldom used in multiple regression, though several different ways to extend this method have been proposed. Anderson and Robinson (2001) review these methods and argue that it is best to permute residuals obtained from estimating the model under H_0, a method proposed by Freedman and Lane (1983).

Conclusion

This survey is restricted to classical inference methods for parametric models. It does not consider Bayesian inference, inference following nonparametric and semiparametric

estimation, or time series complications such as models with unit roots and cointegration.

The graduate-level econometrics texts by Cameron and Trivedi (2005), Greene (2012) and Wooldridge (2010) cover especially sections Inference and Model Tests and Diagnostics; see also Jones (2000) for a survey of health econometrics models and relevant chapters in this volume. The biostatistics literature for nonlinear models emphasizes estimators for generalized linear models; the classic reference is McCullagh and Nelder (1989). For the resampling methods in section Bootstrap and Other Resampling Methods, Efron and Tibsharani (1993) is a standard accessible reference; see also Davison and Hinkley (1997) and, for implementation, Cameron and Trivedi (2010).

See also: Instrumental Variables: Methods. Models for Count Data. Models for Discrete/Ordered Outcomes and Choice Models. Panel Data and Difference-in-Differences Estimation. Primer on the Use of Bayesian Methods in Health Economics. Spatial Econometrics: Theory and Applications in Health Economics. Survey Sampling and Weighting

References

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* **43**, 75–88.

Andrews, D. W. K., Moreira, M. J. and Stock, J. H. (2007). Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics* **139**, 116–132.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–450.

Cameron, A. C. and Miller, D. A. (2011). Robust inference with clustered data. In Ullah, A. and Giles, D. E. (eds.) *Handbook of empirical economics and finance*, pp. 1–28. Boca Raton: CRC Press.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using Stata* First revised edition College Station, TX: Stata Press.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics* **92**, 1–45.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Duflo, E., Glennerster, R. and Kremer, M. (2008). Using randomization in development economics research: A toolkit. In Shultz, T. P. and Strauss, J. A. (eds.) *Handbook of development economics*, vol. 4, pp. 3896–3962. Amsterdam: North-Holland.

Efron, B. and Tibsharani, J. (1993). *An introduction to the bootstrap*. London: Chapman and Hall.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* **17**, 347–388.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics* **1**, 292–298.

Greene, W. H. (2012). *Econometric analysis*, 7th ed. Upper Saddle River: Prentice Hall.

Jones, A. M. (2000). Health econometrics. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1, pp. 265–344. Amsterdam: North-Holland.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd ed. London: Chapman and Hall.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98**, 199–214.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

White, H. (2001). A reality check for data snooping. *Econometrica* **68**, 1097–1126.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review* **93**, 133–138.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, 2nd ed. Cambridge, MA: MIT Press.

Information Analysis, Value of

K Claxton, University of York, York, North Yorkshire, UK

Policy Relevance

The general issue of balancing the value of evidence about the performance of a technology and the value of providing patients with access to a technology can be seen as central to a number of policy questions in many different types of health-care systems (HCS). For example, decisions about approval or reimbursement of new drugs are increasingly being made close to their launch when the evidence base to support their use is least mature and when there may be substantial uncertainty surrounding their cost effectiveness. In these circumstances, further evidence maybe particularly valuable as it will lead to better decisions about the use of the technology, which would improve patient outcomes and/or reduce resource costs. Therefore, it is useful to establish the key principles of what assessments are needed to decide whether there is sufficient evidence to support reimbursement or recommending the use of a new drug, whether it should be approved but additional evidence sort or whether its widespread use should be restricted until the additional evidence is available. Such assessments can help to inform the questions posed by coverage with evidence development and managed entry in many health-care systems including restricting approval to 'only in research' which is part of the UK National Institute for Heath and Clinical Excellence (NICE) statutes.

If there are constraints on the growth of health-care expenditure, then approving a more costly technology will displace other activities that would have otherwise generated improvements in health for other patients, as well as other socially valuable activities outside health care. If the objective of a HCS is to improve health outcomes across the population it serves then, even if a technology is expected to be more effective, the health gained must be compared to the health expected to be forgone elsewhere as a consequence of additional costs, i.e., whether the technology is expected to be cost effective and offer positive net health benefits (NHB) (other effects, e.g., on consumption, can also be expressed as their health equivalent). An assessment of expected cost effectiveness or NHB relies on evidence about effectiveness, impact on long-term overall health and potential harms, as well as additional health-care costs together with some assessment of what health is likely to be forgone as a consequence (the cost-effectiveness threshold).

Such assessments are inevitably uncertain and, without sufficient and good quality evidence, decisions about the use of technologies will also be uncertain. There will be a chance that the resources committed by the approval of a new technology may be wasted if the expected positive net health effects are not realized. Equally, rejecting a new technology will risk failing to provide access to a valuable intervention if the net health effects prove to be greater than expected. Therefore, if the social objective is to improve overall health for both current and future patients then the need for and the value of additional evidence is an important consideration when making decisions about the use of technologies. This is even more critical once it is recognized that the approval of a technology for widespread use might reduce the prospects of conducting the type of research that would provide the evidence needed. In these circumstances there will be a trade-off between the net health effects for current patients from early access to a cost-effective technology and the health benefits for future patients from withholding approval until valuable research has been conducted.

Research also consumes valuable resources which could have been devoted to patient care, or other more valuable research priorities. Also uncertain events in the near or distant future may change the value of the technology and the need for evidence (e.g., prices of existing technologies, the entry of new technologies and other evidence about the performance of technologies as well as the natural history of disease). In addition, implementing a decision to approve a new technology may commit resources which cannot subsequently be recovered if a decision to approve or reimburse might change in the future (e.g., due to research reporting). Therefore, appropriate research and coverage decisions will depend on whether the expected benefits of research are likely to exceed the costs and whether any benefits of early approval or reimbursement are greater than withholding approval until additional research is conducted or other sources of uncertainty are resolved. Methods of analysis which provide a quantitative assessment of the potential benefits of acquiring further evidence allow research and reimbursement decisions to be addressed explicitly and accountably.

The Value of Additional Evidence

The principles of value of information analysis have a firm foundation in statistical decision theory with closely related concepts and methods in mathematics and financial economics with diverse applications in business decisions, engineering, environmental risk analysis, and financial and environmental economics. There are now many applications in health, some commissioned to directly inform policy and others published in specialist as well as general medical and health policy journals. Most commonly these methods of analysis have been applied in the context of probabilistic decision analytic models used to estimate expected cost effectiveness of alternative interventions. However, the same type of analysis can also be used to extend standard methods of systematic review and meta-analysis. Indeed the principles or value of information analysis can also be used as a conceptual framework for qualitative assessment of how important uncertainty might be and the relative priority of alternative research topics and proposals.

Additional evidence is valuable because it can improve patient outcomes by resolving existing uncertainty about the cost effectiveness of the interventions available, thereby

informing treatment choice for subsequent patients. For example, the balance of existing evidence might suggest that a particular intervention is expected to be cost effective and offer the greatest NHB, but there will be a chance that others are in fact more cost effective, offering higher NHB to the HCS. If treatment choice is based on existing evidence then there will be a chance that other interventions would have improved overall health outcomes to a greater extent, i.e., there are adverse net health consequences associated with uncertainty. The scale of uncertainty can be indicated by the results of probabilistic analysis of a decision analytic model and/or based on the results of a meta-analysis of the evidence relevant to the choice between interventions. The expected consequences of this uncertainty can be expressed in terms of NHB or the equivalent HCS resources that would be required to generate the same net health effects. These expected consequences can be interpreted as an estimate of the NHB that could potentially be gained per patient if the uncertainty surrounding their treatment choice could be resolved, i.e., it indicates an upper bound on the expected NHB of further research.

Expected Value of Perfect Information

More formally, if there are alternative interventions (j), where the NHB of each depends on uncertain parameters that may take a range of possible values (θ), the best decision based on the information currently available would be to choose the intervention that is expected to offer the maximum net benefit (i.e., $max_j\ E_\theta\ NHB(j, \theta)$). If the uncertainty could be fully resolved (with perfect information), the decision maker would know which value θ would take before choosing between the alternative interventions. They would be able to select the intervention that provides the maximum NHB for each particular value of θ (i.e., $max_j\ NHB(j, \theta)$). However, when a decision about whether further research should be undertaken is made, the results (the true values of θ) are necessarily unknown. Therefore, the expected NHB of a decision taken when uncertainties are fully resolved (with perfect information) is then found by averaging these maximum net benefits over all the possible results of research that would provide perfect information (over the joint distribution of θ); $E_\theta\ max_j\ NB(j, \theta)$. The expected value of perfect information (EVPI) for an individual patient is simply the difference between the expected value of the decision made with perfect information about the uncertain parameters θ, and the decision made on the basis of existing evidence ($EVPI = E_\theta\ max_j\ NB(j, \theta) - max_j\ E_\theta\ NHB(j, \theta)$).

Once the results of research are available they can be used to inform treatment choice for all subsequent patients. Therefore, the potential expected benefit of research (EVPI) needs to be expressed for the population of patients that can benefit from it. The population EVPI will increase with the size of the patient population whose treatment choice can be informed by additional evidence and the time over which evidence about the cost effectiveness of these interventions is expected to be useful, but will tend to decline with the time that research is likely to take to be commissioned, conducted and report.

Time Horizons for Research Decisions

The information generated by research will not be valuable indefinitely, because other changes occur over time, which will have an impact on the future value of the information generated by research that can be commissioned today. For example, over time the prices of the alternative technologies are likely to change (e.g., patent expiry of branded drugs and the entry of generics versions) and new and more effective interventions become available which will eventually make current comparators obsolete, so information about their effectiveness will no longer be relevant to future clinical practice. Other information may also become available in the future which will also impact on the value of the evidence generated by research that can be commissioned today. For example, other evaluative research might be (or may already have been) commissioned by other bodies or HCS, that may resolve much of the uncertainty anyway. Also, this research or other more basic science may fundamentally change our understanding of disease processes and effective mechanisms. Finally, as more information about individual effects is acquired through greater understanding of the reasons for variability in patient outcomes, the value of evidence that can resolve uncertainty in expected or average effects for the patient population and/or it's subpopulations will decline (see Section Uncertainty, Variability, and Individualized Care). For all these reasons there will be a finite time horizon for the expected benefits of additional evidence, i.e., there will be a point at which the additional evidence that can be acquired by commissioning research today will no longer be valuable.

The actual time horizon for a particular research decision is unknown, because it is a proxy for a complex, and uncertain process of future changes. Nonetheless some judgment, whether made implicitly or explicitly, is unavoidable when making decisions about research priorities. Some assessment is possible based on historical evidence and judgments about whether a particular area is more likely to see earlier patent expiration, future innovations, other evaluative research, and the development of individualized care (e.g., where diagnostic technologies, application of genomics, and the development of evidence-based algorithms are rapidly developing). Information can also be acquired about trials that are already planned and underway around the world (e.g., various trial registries) and future innovations from registered patents and/or phase I and II trials as well as licensing applications, combined with historic evidence on the probability of approval and diffusion. For these reasons, an assessment of an appropriate time horizon may differ across different clinical areas and specific research proposals. The incidence of patients who can benefit from the additional evidence may also change over time, although not necessarily decline as other types of effective health-care change competing risks. However, in some areas recent innovations might suggest a predictable decline, e.g., the decline in the incidence of cervical cancer following the development of the HPV vaccine.

Research Prioritization Decisions

Two questions are posed when considering whether further research should be prioritized and commissioned: Are the

potential expected NHB of additional evidence (population EVPI) sufficient to regard the type of research likely to be required as potentially worthwhile; and should it be prioritized over other research that could be commissioned with the same resources? Of course, these assessments require some consideration of the period of time over which the additional evidence generated by research is likely to be relevant; as well as the time likely to be taken for proposed research to be commissioned, conducted and report.

One way to address the question is to ask whether the HCS could generate similar expected NHB more effectively elsewhere, or equivalently whether the costs of the research would generate more NHB if these resources were made available to the HCS to provide health care. Very recent work in the UK has estimated the relationship between changes in NHS expenditure and health outcomes. This work suggests that the NHS spends approximately £75 000 to avoid one premature death, £25 000 to gain one life year and somewhat less than £20 000 to gain one quality-adjusted life-year (QALY). Using these estimates proposed research that, for example, costs £2 million could have been used to avoid 27 deaths and generate more than 100 QALY elsewhere in the NHS. If these opportunity costs of research are substantially less than the expected benefits (population EVPI) then it would suggest that the proposed research is potentially worthwhile.

However, most research funders have limited resources (with constraints relevant to a budgetary period) and cannot draw directly on the other (or future) resources of the HCS. Therefore, even if the population EVPI of proposed research exceeds the opportunity costs it is possible that other research may be even more valuable. If similar analysis is conducted for all proposals competing for limited research resources it does become possible to identify a short list of those which are likely to be worthwhile and then select from these those that are likely to offer the greatest value.

Research and Reimbursement Decisions

It should be noted that the population EVPI represents only the potential or maximum expected benefits of actual research that could be conducted for two reasons: no research, no matter how large the sample size or how assiduously conducted can resolve all uncertainty and provide perfect information; and there are usually a large number of uncertain parameters that contribute to θ and are relevant to differences in NHB of the alternative interventions – most research designs will not provide information about all of them. Nonetheless EVPI does provide an upper bound to the value of conducting further research, so when compared with the opportunity cost of conducting research (e.g., the health equivalent of the resources required) it can provide a necessary condition for a decision to conduct further research while the intervention is approved for widespread use. It also provides a sufficient condition for early approval when approval would mean that the type of further research needed would not be possible or too costly to be worthwhile (e.g., because there would be a lack of incentives for manufacturers, or further randomized trials would not be regarded as ethical and/or would be unable to recruit). In these circumstances the population EVPI represents an upper bound on the benefits to future patients that would be forgone or the opportunity costs of early approval based on existing evidence.

What Type of Evidence?

The type of analysis described above indicates the potential value of resolving all the uncertainty surrounding the choice between alternative the interventions. However, it would be useful to have an indication of which sources of uncertainty are most important and what type of additional evidence would be most valuable. This can start to indicate the type of research design that is likely to be required, whether the type of research required will be possible once a new technology is approved for widespread use as well as indicating the sequence in which different studies might be conducted.

Expected Value of Perfect Parameter Information

The potential expected benefits of resolving the different sources of uncertainty that determine the NHB of the alternative interventions can be established using the same principles. For example, if the NHB of each intervention (j) depends on two (groups of) uncertain parameters ($\theta1$ and $\theta2$) that may take a range of possible values, the best decision based on current information is still to choose the intervention that is expected to offer the maximum net benefit (i.e., $max_j\ E_{\theta2,\theta1}\ NHB(j, \theta1, \theta2)$). If the uncertainty associated with only one of these groups of parameters ($\theta1$) could be fully resolved (i.e., with perfect parameter information), the decision maker would know which value $\theta1$ would take before choosing between the alternative interventions. However, the values of the other parameters ($\theta2$) remain uncertain so the best they can do is to select the intervention that provides the maximum expected NHB for each value of $\theta1$ (i.e., $max_j\ E_{\theta2|\theta1}\ NHB(j, \theta1, \theta2)$). Which particular value $\theta1$ will take is unknown before research is conducted so the expected NHB when uncertainty associated with $\theta1$ is fully resolved is the average of these maximum net benefits over all the possible values of $\theta1$, (i.e., $E_{\theta1}\ max_j\ E_{\theta2|\theta1}\ NHB(j, \theta1, \theta2)$). The expected value of perfect parameter information about $\theta1$ (EVPPI $_{\theta1}$) is simply the difference between the expected value of the decisions made with perfect information about $\theta1$, and a decision based on existing evidence (EVPI $= E_{\theta1}\ max_j\ E_{\theta2|\theta1}\ NHB(j, \theta1, \theta2) - max_j\ E_{\theta2,\theta1}\ NHB(j, \theta1, \theta2)$).

It should be noted that this describes a general solution for nonlinear models. However, it is computationally intensive because it requires an inner loop of simulation to estimate the expected NHB for each value of $\theta1$ ($E_{\theta2|\theta1}\ NHB(j, \theta1, \theta2)$), as well outer loop of simulation to sample the possible value $\theta1$ could take. The computational requirements can be somewhat simplified if there is a multilinear relationship between the parameters and net benefit. If the model is multilinear in $\theta2$, the parameters in $\theta2$ are uncorrelated with each other and $\theta1$ and $\theta2$ are independent then the inner loop of simulation is unnecessary (using the mean values of $\theta2$ will return the correct estimate of $E_{\theta2|\theta1}\ NHB(j, \theta1, \theta2)$).

Sequence of Research

This type of analysis can be used to focus research on the type of evidence that will be most important by identifying those parameters for which more precise estimates would be most valuable. In some circumstances, this will indicate which endpoints should be included in further experimental research. In other circumstances, it may focus research on getting more precise estimates of particular parameters that may not necessarily require experimental design and can be provided relatively quickly. This type of analysis can be extended to consider the sequence in which different types of study might be conducted, e.g., whether: no research; research about $\theta 1$ and $\theta 2$ simultaneously; $\theta 1$ first and then $\theta 2$ depending on the results of $\theta 1$ research; or $\theta 2$ first and then $\theta 1$ depending on the results of $\theta 2$ research, would be the most valuable research decision.

Informing Research Design

Identifying which sources of uncertainty are most important and what type of evidence is likely to be most valuable is useful in two respects. It can help to identify the type of research design that is likely to be required (e.g., an randomized controlled trial (RCT) may be needed to avoid the risk of selection bias if additional evidence about the relative effect of an intervention is required) and identify the most important endpoints to include in any particular research design. It can also be used to consider whether there are other types of research that could be conducted relatively quickly (and cheaply) before more lengthy and expensive research (e.g., a large RCT) is really needed (i.e., the sequence of research that might be most effective).

Estimates of EVPI and EVPPI only provide a necessary condition for conducting further research. To establish a sufficient condition to decide if further research will be worthwhile and identify efficient research design, estimates of the expected benefits and the cost of sample information are required.

The same value of information analysis framework can be extended to establish the expected value of sample information (EVSI) for particular research designs.

Expected Value of Sample Information

For example, a sample of n on θ will provide a sample result D. If the sample result was known the best decision would be to choose the alternative with the maximum expected net benefit when the estimates of the NHB of each alternative was based on the sample result (averaged over the posterior distribution of the net benefit given the sample result D). However, which particular sample result will be realized when the research reports is unknown. The expected value of acquiring a sample of n on θ is the found by averaging these maximum expected net benefits over the distribution of possible sample results, D, i.e., the expectation over the predictive distribution of the sample results D conditional on θ, averaged over the possible values of θ (the prior distribution of θ). The additional expected benefit of sample information (EVSI) is simply the difference between the expected value of a decision made with sample information and the expected value with current information.

The EVSI calculations require the likelihood for the data to be conjugate with the prior so there is an analytic solution to combining the prior distribution of θ with the predicted sample result (D) to form a predicted posterior. If the prior and likelihood are not conjugate, the computational burden of using numerical methods to form predicted posteriors is considerable. Even with conjugacy, EVSI still requires intensive computation if the relationship between the sampled parameters (end points in the research design) and differences in the NHB of the alternatives are nonlinear.

Optimal Sample Size and Other Aspects of Research Design

To establish the optimal sample size for a particular type of study these calculations need to be repeated for a range of sample sizes. The difference between the EVSI and the costs of acquiring the sample information is the expected net benefit of sample information (ENBS) or the societal payoff to research. The optimal sample size is simply the value of n that generates the maximum ENBS. As well as sample size the same type of analysis can be used to evaluate a range of different dimensions of research design such as which endpoints to include, which interventions should be compared, and the length of follow-up. The best design is the one that provides the greatest ENBS. The same type of analysis can also be used to identify whether a combination of different types of study might be required (an optimal portfolio of research). It should be recognized that the costs of research not only include the resources consumed in conducting it but also the opportunity costs (NHB forgone) falling on those patients enrolled in the research and those whose treatment choice can be informed once the research reports. Therefore, optimal research design will depend, among other things, on whether or not patients have access to the new technology while the research is being conducted and how long it will take before it reports (determined by length of follow-up and recruitment rates). It is also possible to take account of likely implementation of research findings in research design, e.g., if an impact on clinical practice depends on the trial reporting a statistically significant result for a particular effect size (and there are no other effective ways to ensure implementation) this will influence optimal sample size as well.

The Value of Commissioned Research

Research decisions require an assessment of the expected potential value of future research before the actual results that will be reported in the future are known. Therefore, using hindsight to inform research prioritization decisions is inappropriate for two reasons: (1) such an (*ex post*) assessment cannot directly address the (*ex ante*) question posed in research prioritization decisions; and (2) assessing the (*ex post*) value of research with hindsight is potentially misleading if used to judge whether or not the original (*ex ante*) decision to prioritize and commission it was appropriate. This is because the findings of research are only one realization of the uncertainty about potential results that could have been found

when the decision to prioritize and commission research must be taken.

It is useful and instructive, however, to reconsider the analysis set out above once the results of research become available by updating the synthesis of evidence, reestimating the NHB of the alternative interventions and updating the value of information analysis to consider whether the research was indeed definitive (the potential benefits of acquiring additional evidence does not justify the costs of further research) or whether more or different types of evidence might be required. Therefore, value of information analysis can also provide the analytic framework to consider when to stop a clinical trial, how to allocate patients between the arms of a trial as evidence accumulates (sequential and group sequential designs) and when other types of evidence might become more important as the results of research are realized over time.

Value of Implementation

Overall health outcomes can also be improved by ensuring that the accumulating findings of research are implemented and have an impact on clinical practice. Indeed, the potential improvements in health outcome by encouraging the implementation of what existing evidence suggests is the most cost-effective intervention may well exceed the potential improvements in NHB through conducting further research.

The distinction between these two very different ways to improve overall health outcomes is important because, although the results of additional research may influence clinical practice and may contribute to the implementation of research findings, it is certainly not the only, or necessarily the most effective, way to do so. Insofar as there are other more effective mechanisms (e.g., more effective dissemination of existing evidence) or policies (e.g., those that offer incentives and/or sanctions), than continuing to conduct research to influence clinical practice, rather than because there is real value in acquiring additional evidence itself, would seem inappropriate, because research resources could have been used elsewhere to acquire additional evidence in areas where it would have offered greater potential NHB.

Clearly, the potential health benefits of conducting further research will only be realized (health outcomes actually improve and/or resources are saved) if the findings of the research do indeed have an impact on clinical practice. Recognizing that there are very many ways to influence the implementation of what current evidence suggests, other than by conducting more research, is important when considering other policies to improve implementation of research findings instead of, or in combination with, conducting further research. However, the importance of implementing the findings of proposed research might influence consideration of its priority and research design in a number of ways. If it is very unlikely that the findings of proposed research will be implemented and other mechanisms are unlikely to be effective or used, then other areas of research where smaller potential benefits are more likely to be realized might be prioritized. If the impact of research on clinical practice is likely to require highly statistically significant results this will influence the design, cost, and time taken for research to report and therefore its relative priority. It maybe that a larger clinical difference in effectiveness would need to be demonstrated before research would have impact on clinical practice. This will tend to reduce the potential benefits of further research as well because large differences are less likely to be found than small ones.

Decisions Based on the Balance of Existing Evidence?

It should be recognized that restricting attention to whether or not the result of a clinical trial, a meta-analysis of existing trials, or the results of a cost-effectiveness analysis offer statistically significant results is unhelpful for a number of reasons: it provides only a partial summary of the uncertainty associated with the cost effectiveness of an intervention, nor does it indicate the importance of the uncertainty for overall patient outcomes or the potential gains in NHB that might be expected from acquiring additional evidence that could resolve it. Of course, failing to implement an intervention which is expected to offer the greatest NHB will impose unnecessary opportunity cost. This suggest that always waiting to implement research findings until the traditional rules of statistical significance are achieved (whether based on frequentist hypothesis testing or on Bayesian bench mark error probabilities) may well come at some considerable cost to patient outcomes and HCS resources.

However, once uncertainty and the value of additional evidence is recognized there are a number of issues that need to be considered before decisions to approve or reimburse a new technology can be based on the balance of accumulated evidence, i.e., expected cost effectiveness and expected NHB:

1. As already discussed, if early approval or reimbursement means that the type of research required to generate the evidence needed is impossible or more difficult to conduct then the expected value of additional evidence that will be forgone by approval needs to be considered alongside the expected benefits of early implementation.
2. Insofar as widespread use of an intervention will be difficult to reverse if subsequent research demonstrates that it is not cost effective (e.g., where it would require resources and effort as well as take time to achieve), then account must be taken of the consequences of this possibility (i.e., the opportunity costs associated the chance that research finding that the intervention is not cost effective but being unable to immediately implement these findings and withdraw its use).
3. If an intervention offers longer-term benefits which will ultimately justify initial treatment costs (e.g., any effect on mortality risk) its approval or reimbursement is likely to commit initial losses of NHB compensated by later expected gains. In these circumstances its approval or reimbursement commits irrecoverable opportunity costs for each patient treated. If the uncertainty about it costs-effectiveness might be resolved in the future (e.g., due to commissioned research reporting) then if may be better to withhold approval or reimbursement until the research findings are available even if the research could be conducted while the technology is in widespread use. This is

more likely to be the case when a decision to delay initiation of treatment is possible and associated with more limited health impacts (e.g., in chronic and stable conditions).

4. There is a common and quite natural aversion to iatrogenic effects, i.e., health lost through adopting an intervention not in widespread use tends to be regarded as of greater concern than the same health lost through continuing to use existing interventions that are less effective than others available. However, it should be noted that the consequences for patients are symmetrical and this 'aversion' also depends entirely on which intervention happened to have diffused into common clinical practice first.

These considerations can inform an assessment of whether more health might be gained through efforts to implement the findings of existing research or by acquiring more evidence to inform which intervention is most cost effective. Although there are many circumstances where approval or reimbursement should not be simply based on the balance of evidence (i.e., expected cost effectiveness or expected NHB), it should be noted that these considerations are likely to differ between decisions and certainly do not lead to a single 'rule' based on notions of the statistical significance of the results of a particular study, a meta-analysis of existing studies, or the results of a cost-effectiveness analysis. They can be, and have been, dealt with explicitly and quantitatively within well conducted value of information analysis.

Uncertainty, Variability, and Individualized Care

It is important to make a clear distinction between uncertainty, variability, and heterogeneity. Uncertainty refers to the fact that we do not know what the expected effects will be of using an intervention in a particular population of patients (i.e., the NHB of an intervention on average). This remains the case even if all patients within this population have the same observed characteristics. Additional evidence can reduce uncertainty and provide a more precise estimate of the expected effects in the whole population or within subpopulations that might be defined based on different observed characteristics. Variability refers to the fact that individual responses to an intervention will differ within the population or even in a subpopulation of patients with the same observed characteristics. Therefore, this natural variation in responses cannot be reduced by acquiring additional evidence about the expected or average effect. Heterogeneity refers to those individual differences in response that can be associated with differences in observed characteristics, i.e., where the sources of natural variability can be identified and understood. As more becomes known about the sources of variability (as variability is turned into heterogeneity) the patient population can be partitioned into subpopulations or subgroups, each with a different estimate of the expected effect of the intervention and the uncertainty associated with it. Ultimately, as more sources of variability become known the subpopulations become individual patients, i.e., individualized care.

Overall patient outcomes can be improved by either acquiring additional evidence to resolve the uncertainty in the expected effects of an intervention, and/or by understanding the sources of variability and dividing the population into finer subgroups where the intervention will be expected to be cost effective in some but not in others. However, a greater understanding of heterogeneity also has an impact on the value of additional evidence. As more subgroups can be defined the precision of the estimates of effect is necessarily reduced (the same amount of evidence offers fewer observations in each subgroup). However, the uncertainty about which intervention is most cost effective may be reduced in some (e.g., where it is particularly effective or positively harmful), but increase in others. Therefore, the expected consequences of uncertainty per patient, or value of additional evidence per patient may be higher or lower in particular subgroups. The expected value of evidence across the whole population (the sum across all subgroups of the population) may rise or fall. However, in the limit as more sources of variability are observed the value of additional evidence will fall. Indeed, if all sources of variability could be observed then there would be no uncertainty at all.

Value of information analysis can be applied within each subgroup identified based on existing evidence. Conducting an analysis of the expected health benefits of additional evidence by subgroups is useful because it can indicate which types of patient need to be included in any future research design and others that could be excluded. Although the potential value of additional evidence about the whole population is simply the sum of values for each of its subpopulations, the value of acquiring evidence within only one subgroups depends on whether that evidence can inform decisions in others. For example, if subgroups are identified based on differing base line risks then evidence about the relative effect of an intervention in one subgroup might also inform relative effect in others so the value of research conducted in one of the subgroups should take account of the value it will generate in others. However, evidence about a subgroup specific baseline risk might not be relevant and offer value in others. In principle, these questions of exchangeability of evidence can be informed by how existing evidence and ought to be reflected in how it is synthesized and the uncertainties characterized.

Therefore there is potential value of research which might not resolve uncertainty but instead reveal the reasons for variability in outcome; informing which subgroups could benefit most from an intervention, or the choice of the physician patient dyad in selecting care given their symptoms, history and preferences (i.e., individualized care). This type of research may be very different from the type of evaluative research that reduces uncertainty about estimates of effect. For example, it might include: diagnostic procedures and technologies, pharmacogenetics; analysis of observational data and treatment selection as well as novel trial designs which can reveal something of the joint distribution of effects. Much methodological and applied work has been conducted in this rapidly developing area. There is an opportunity to explore ways of estimating the potential value of such research (the expected benefits of heterogeneity) based only on existing evidence. This would provide a very useful complement to estimates of EVPI and EVPPI. It would allow policy makers to consider whether HCS resources should be invested in: providing early access to new technologies; ensuring the findings

of existing (or commissioned) research are (or will be) implemented; conducting research to provide additional evidence about particular sources of uncertainty in some (or all) subgroups; or conducting research which can lead to a better understanding of variability in effects. Of course some combination of these policy choices may well offer the greatest impact on overall health outcomes.

Value of Information and Cost-Effectiveness Analysis

The discussion of value of information analysis has been founded on a HCS which faces some constraints on the growth of health-care expenditure so additional HCS costs displace other care that would have otherwise generated improvements in health. In the UK recent estimates of the rate at which health-care cost displace health elsewhere (the cost-effectiveness threshold) are now available. However, in all HCS new technologies impose costs (or offer benefits), which fall outside the health care and displace private consumption rather than health. If some consumption value of health is specified then these other effects can also be expressed as their health equivalent and included in the expression for NHB. Impacts on health, HCS resources, and consumption can also be expressed in terms of the equivalent net private consumption effects or the equivalent HCS resources (these monetary values will only be the same if the estimate of the threshold is the same as some consumption value of health). Therefore the methods of analysis outlined above are not restricted to cost-effectiveness analysis applied in HCS which have administrative budget constraints and/or where decision making bodies disregard effects outside the HCS. It is just as relevant to an appropriately conducted cost-benefit analysis (one which accounts for the shadow price of any constraints on health-care expenditure).

Equally the principles of value of information analysis can be usefully applied even in circumstances where decision making bodies are unwilling or unable to explicitly include any form of economic analysis in their decision making process. For example, a quantitative assessment of the expected health (rather than net health) benefits of additional evidence is possible by applying value of information analysis to the results of standard methods of systematic review and meta-analysis. Insofar as there are additional costs associated with more effective interventions this will tend to overestimate the expected *NHB* of additional evidence. Also the endpoints included in the meta-analysis of previous trials may not capture all valuable aspects of health outcome. For example, although mortality following acute myocardial infarction maybe the appropriate primary outcome in the evaluation of early thrombolysis, it is not necessarily the only relevant outcome. Stroke and its consequences are also very relevant as well as length of survival and the type of health experienced in the additional years of life associated with mortality effects.

Specifying a minimum clinical difference required to change clinical practice is one way to incorporate concerns about potential adverse events and other consequences of recommending a more effective intervention, including the additional costs, albeit implicitly. This concept of an effect size has been central to the design of clinical research and determines the sample size in most clinical trials. The effect size does not represent what is expected to be found by the research, but the difference in outcomes that would need to be detected for the results to be regarded as clinically significant and have an impact on clinical practice. The same concept can be used to report estimates of the expected heath benefits of additional evidence for a range of minimum clinical differences (MCD) in outcomes. The value of additional evidence and the need for further research depends on the clinical difference in key aspects of outcome that would be need to be demonstrated before clinical practice 'should' or is likely to change. There are a number of circumstances where a larger MCD might be required. For example: (1) where the quantitative analysis is restricted to the primary endpoint reported in existing clinical trials but there other important aspects of outcome that are not captured in this endpoint (e.g., adverse events or quality of life impacts that have not been accounted for in the meta-analysis); (2) when there is an impact on HCS costs, out of pocket expenses for patients or the wider economy; and (3) it maybe that larger clinical difference in effectiveness would need to be demonstrated before research would have an impact on practice and the findings of proposed research would be widely implemented.

Requiring that further research must demonstrate larger differences in effect will tend to reduce its expected potential benefits because large differences are less likely to be found than smaller ones. Specifying an MCD through some form of deliberative process would implicitly account for the other unquantified aspects of outcome, HCS costs and other non-health effects. Of course decision makers would need to consider whether proposed research is still a priority at an MCD that is regarded as sufficient to account for these other effects. Importantly, whatever the policy context, the principles and established methods of value of information analysis are relevant to a wide range of different types of HCS and decision making contexts and should not be regarded as being restricted to situations where probabilistic decision analytic models to estimate cost effectiveness based on QALYs as a measure of health are available and routinely used within the decision making process.

See also: Analysing Heterogeneity to Support Decision Making. Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties. Economic Evaluation, Uncertainty in. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Statistical Issues in Economic Evaluations. Synthesizing Clinical Evidence for Economic Evaluation

Further Reading

Ades, A. E., Lu, G. and Claxton, K. (2004). Expected value of sample information in medical decision modelling. *Medical Decision Making* **24**(2), 228–702.

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making* **27**(2), 112–127.

Briggs, A., Claxton, K. and Sculpher, M. J. (2006). *Decision analytic modelling for health economic evaluation*. Oxford: Oxford University Press.

Claxton, K. (1999). The irrelevance of inference: A decision making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* **17**(3), 341–364.

Claxton, K., Griffin, S., Hendrik, K. and McKenna, C. (2013). Expected health benefits of additional evidence: Principles, methods and applications. CHE Research Paper 83. University of York, York.

Claxton, K., Palmer, S., Longworth, L., et al. (2012). Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. *Health Technology* **16**(46), doi:10.3310/hta16460.

Colbourn, T., Asseburg, C., Bojke, L., et al. (2007). Preventive strategies for group B streptococcal and other bacterial infections in early infancy: Cost effectiveness and value of information analyses. *British Medical Journal* **335**, 655–662.

Eckermann, S. and Willan, A. R. (2008). The option value of delay in health technology assessment. *Medical Decision Making* **28**(3), 300–305.

Griffin, S., Claxton, K., Palmer, S. and Sculpher, M. (2011). Dangerous omissions: The consequences of ignoring decision uncertainty. *Health Economics* **20**, 212–224, doi:10.1002/hec.1586.

Griffin, S., Claxton, K. and Welton, N. (2010). Exploring the research decision space: The expected value of sequential research designs. *Medical Decision Making* **30**, 155–162, doi:10.1177/0272989 × 09344746.

Hoomans, T., Fenwick, E., Palmer, S. and Claxton, K. (2009). Value of information and value of implementation: Application of a framework to inform resource allocation decisions in metastatic hormone-refractory prostate cancer. *Value in Health* **12**, 315–324, doi:10.1111/j.1524-4733.2008.00431.x.

McKenna, C. and Claxton, K. (2011). Addressing adoption and research design decisions simultaneously: The role of value of sample information analysis. *Medical Decision Making*, doi:10.1177/0272989 × 1139992.

McKenna, C., Claxton, K., Chalabi, Z. and Epstein, D. (2010). Budgetary policies and available actions: A generalisation of decision rules for allocation and research decisions. *Journal of Health Economics* **29**, 170–181.

Instrumental Variables: Informing Policy

MC Auld, University of Victoria, Victoria, BC, Canada
PV Grootendorst, University of Toronto, Toronto, ON, Canada

Introduction

Health economists frequently face the challenge of estimating causal relationships in the absence of controlled experiments. For example, a long-standing issue in economics and in other disciplines is unraveling the observed relationship between education and health. Countless studies have documented a positive correlation between these outcomes, but fewer have successfully addressed the causal impact of education and health. In principle, randomized controlled trials (RCTs) could be used, but it is difficult to experimentally manipulate levels of education. Instrumental variables (IV) methods can be used when the real world provides some quasiexperimental variation in education. In this article, the use and the limitations of the IV approach are discussed. The authors illustrate how IV approach works, review its relationship with the experimental approach, identify the properties of good natural experiments, and discuss the statistical properties of the IV estimator when the natural experiment is less than ideal.

The Instrumental Variables Estimator

An Intuitive Explanation for the Univariate Model

Consider the statistical properties of the linear IV estimator. For the sake of simplicity, the univariate case is presented, and the constant is suppressed by assuming that all variables are expressed as deviations from their respective sample means. Suppose that the effect of a broadly defined 'treatment,' x, on an outcome y is to be estimated. Data on y and x are collected for a random sample of n observations; y_i and x_i denote the values of these variables for the ith observation. The treatment affects the outcome according to a linear regression of the form

$$y_i = \beta x_i + u_i \qquad [1]$$

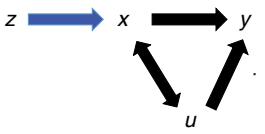
where β is an unknown parameter to be estimated and u_i is an unobserved error term, interpreted as all causes of y_i other than x_i. Here, β is interpreted as the causal effect of x on y, and x and u are possibly correlated. The variables u and x will be correlated if there are variables unobserved to the researcher which cause both x and y ('omitted variables' in econometrics, or 'unobserved confounders' in some other disciplines) or if y 'reverse' causes x. The researcher may attempt to address omitted variables by using standard multivariate regression specifications and adding more independent variables to the model, but commonly, as in the education and health example above, even very rich datasets will exclude information on countless personality, cognitive, background, and contextual variables that may affect both the outcome and the intensity of treatment. Moreover, controlling for additional variables does not help resolve the 'reverse' causation problem. Methods other than IV are sometimes available – such as regression discontinuity designs, or certain longitudinal data approaches – but attention here is limited to IV.

When a regressor is correlated with the error term u, it is said to be endogenous; if not it is said to be exogenous. If ordinary least squares (OLS) is used to estimate the parameters of this equation, then the OLS estimator of β, denoted $\hat{\beta}$, will be biased and inconsistent if x is endogenous. It can be shown that

$$\mathrm{E}(\hat{\beta}|x) = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \beta + \frac{\mathrm{Cov}(x,u)}{\mathrm{Var}(x)} \qquad [2]$$

where E is the expectation operator, $\mathrm{Cov}(x, u)$ is the covariance between x and u, β is the true value of the causal effect that is to be estimated, and $\mathrm{Var}(x)$ is the variance of x. That is, the distribution of the OLS estimator is centered on the causal effect of interest plus a term which depends on the extent to which unobserved causes of the outcome (u) vary with the treatment (x). Here, y may move with x even if x has no causal effect on y, either because y 'reverse' causes x, or because x and u share common causes, leading to biased and inconsistent OLS estimates of the causal effect of interest.

The method of IV can solve the problem in some circumstances. Suppose that z is a variable which has the property that z affects y only because z affects x, which in turn affects y, as illustrated in the diagram



If z affects y only through its effect on x, then correlation between the instrument z and the outcome of interest y implies that x causes y. Under this assumption, the effect of a one unit change in z on y is the product of the effect of z on x and the effect of x on y. The observed association between z and y reveals only the product of these two effects. However, the effect of x on y can be isolated by dividing the observed association between z and y by the observed association of z and x.

The derivation of the IV estimator can be shown more formally (using the method of indirect least squares) as

$$y = \beta x(z,u) + u \qquad [3]$$

expressing the treatment x as a function of the instrument z and the unobserved causes of y, u. Note that the key condition that z only affects y because z affects x is imposed. Differentiate with respect to z to find

$$\frac{dy}{dz} = \beta \frac{dx}{dz} \qquad [4]$$

as $du/dz=0$ by assumption. Rearrange to find

$$\beta = \frac{dy/dz}{dx/dz} \quad [5]$$

which tells one that the causal effect of interest is the ratio of the effect of z on y to the effect of z on x. If those effects are estimated using linear regressions, then

$$\beta = \frac{\mathrm{Cov}(y,z)/\mathrm{Var}(z)}{\mathrm{Cov}(x,z)/\mathrm{Var}(z)} = \frac{\mathrm{Cov}(y,z)}{\mathrm{Cov}(x,z)} \quad [6]$$

Replacing the population moments in the expression above with sample moments calculated from the data yields the linear IV estimator for this model, denoted $\hat{\beta}_{\mathrm{IV}}$,

$$\hat{\beta}_{\mathrm{IV}} = \frac{\sum_i z_i y_i}{\sum_i z_i x_i} \quad [7]$$

Note that, in contrast to the OLS estimator, the IV estimator depends in no way on the correlation between y and x, which is confounded by the common cause u and therefore does not tell us anything useful about the causal effect of x on y. Note also that, unlike the OLS estimator, the denominator of the expression above is a covariance rather than a variance, and it is therefore not bound away from zero. It is clearly required that $\mathrm{Cov}(x, z)$ be different from zero. The problems this issue causes are dealt with below in the discussion on 'weak' instruments, which arise when $\mathrm{Cov}(x, z)$ is not zero but is small.

General Linear Model and Two-Stage Least-Squares Interpretation

Now consider the general linear problem of estimating causal effects when there are k covariates, an arbitrary number k_1 of the covariates are endogenous (correlated with the error term u), and the remainder $k_2=k-k_1$ covariates are exogenous. Let X_{1i} denote the k_1-vector of observations on endogenous regressors for the ith sampled unit and X_{2i} the vector of k_2-vector of observations on the exogenous regressors, so that the model to be estimated can be expressed as

$$y_i = X_i\beta + u_i = X_{1i}\beta_1 + X_{2i}\beta_2 + u_i \quad [8]$$

It is possible to show that the parameters β_1 and β_2 can be estimated if there are $l \geq k_1$ variables, which are correlated (in a sense defined formally below) with the endogenous regressors X_1 but have no direct effect on y after conditioning on X_2, that is, these variables only affect y because, conditional on X_2, they affect the endogenous regressors X_1. If there are fewer than k_1 such variables, the model is said to be underidentified, and the model is not identified. If there are exactly $l=k_1$ such variables, the model is said to be exactly identified, and if there are $l>k_1$ such variables the model is overidentified.

Let $Z_i=(Z_{1i}, X_{2i})$ denote the $(l+k_2)$-vector of observations for all exogenous variables for the ith unit. Here, Z_{1i} is the vector of observations on l variables which only affect y because they affect X_1 – these variables do not appear in the equation that is being estimated (eqn [8]), so they are called the excluded instruments. The vector X_{2i} of observations on exogenous variables in eqn [8] can 'act as their own instruments.' The multivariate version of the estimator defined in eqn [6] is

$$\tilde{\beta}_{\mathrm{IV}} = (X'P_Z X)^{-1}X'P_Z y \quad [9]$$

where $P_Z=Z(Z'Z)^{-1}Z'$. It is possible to show that $\tilde{\beta}_{\mathrm{IV}}$ may be calculated by executing the following steps:

1. Separately for each of the endogenous regressors in X_1, regress the endogenous regressor on the complete set of exogenous variables Z. Save the set of predicted values, $\hat{X}_1$.
2. Regress y on $\hat{X}_1$ and X_2 using OLS.

The estimated coefficients in step 2 are numerically identical to $\tilde{\beta}$ defined in eqn [9]. For this reason, the linear IV estimator is sometimes referred to as the 'two-stage least squares' (2SLS or TSLS) estimator.

Statistical Properties of the IV Estimator

In this section, the sampling properties of the IV estimator are briefly described. Formally, the assumptions that the excluded instruments z_1 only (after conditioning on X_2) affect the outcome y through their effect on the endogenous regressors X_1 can be expressed as

$$\mathrm{plim}_{n\to\infty}\frac{1}{n}Z'u = 0 \quad [10]$$

where plim is the probability limit operator as the sample size n tends to infinity. The condition that the excluded instruments must be correlated with the endogenous regressors can be expressed as

$$\mathrm{plim}_{n\to\infty}\frac{1}{n}X'Z \text{ exists and has full rank } k \quad [11]$$

Under some further regularity conditions, which is omitted, it is possible to show that

$$\mathrm{plim}_{n\to\infty}\tilde{\beta}_{\mathrm{IV}} = \beta \quad [12]$$

that is, the IV estimator is consistent under these assumptions. If the sample size is allowed to grow arbitrarily large, the difference between the estimates and the causal effects of interest becomes arbitrarily small. Further, the estimator is asymptotically normal, permitting conventional inference with standard test statistics (such as the z-ratios and F-statistics). The covariance matrix can be estimated as $s^2(X'P_zX)^{-1}$ if the errors u_i are homoskedastic and serially uncorrelated, where s^2 is a consistent estimate of the variance of u; covariance estimators consistent in the presence of arbitrary heteroskedasticity and serial correlation are also readily available. Finally, the IV estimator is asymptotically efficient in the class of linear estimators.

Note that the IV estimator generally has no desirable small sample properties. It is possible to show that in exactly identified models (models with exactly as many excluded instruments as endogenous regressors),

$$E(\tilde{\beta}_{\mathrm{IV}}) \to \infty \quad [13]$$

that is, the estimator has no moments, its distribution has such 'fat tails' that the integral defining the expected value of the estimator does not converge. In practice, this means that

not uncommonly that one gets 'wild' estimates many standard deviations away from the causal effect of interest. Recall that there are k_1 endogenous regressors and l excluded instruments, and that l is required to be at least as large as k_1. The difference $(l–k_1)$ is the number of overidentifying restrictions. It is possible to show that the number of existing moments of $\tilde{\beta}$ is equal to the number of overidentifying restrictions. For example, if there is one endogenous regressor and one excluded instrument, the model is exactly identified and $\tilde{\beta}$ does not even have a mean. If one more excluded instrument is added, there is one overidentifying restriction and $\tilde{\beta}$ has a mean but not a variance nor any higher order moment, and so on.

The IV estimator is generally biased even when at least one overidentifying restriction exists. As the degree of overidentification rises, the bias of the IV estimator rises and approaches the bias of the OLS estimator as the number of overidentifying restrictions approaches the sample size. At the same time, it is possible to show that the dispersion of the IV estimator falls with the number of overidentifying restrictions.

Generally, researchers face a trade-off: The OLS estimator in the presence of endogenous regressors is inconsistent, but is less dispersed than the IV estimator. Which estimator is preferred depends on the trade-off the researcher is willing to make between bias and dispersion. Adding more excluded instruments (and thus increasing the number of overidentifying restrictions) decreases the dispersion of the IV estimator, but increases its bias.

Examples of Instrumental Variables in Health Research

In this section, some examples of applied IV estimation drawn from the health economics literature are discussed. RCTs are considered as a special case of IV models, and build to more complex models for, first, imperfect RCTs and then uncontrolled experiments.

Example 1: RCT with Perfect Compliance

As a trivial example of IV, consider interpreting standard analysis of an RCT with perfect compliance as an IV estimator. Suppose that y is the outcome of interest, x is a binary variable denoting treatment status such that $x_i=1$ if subject i is given the new therapy and $x_i=0$ if given the standard therapy. The researcher randomly draws a binary variable from a process independent of y (a figurative coin flip); z denotes the outcomes of this process. The researcher then assigns treatment statuses: $x_i=z_i$. In this scenario, z is determined independently of u, and z is perfectly correlated with x; z thus satisfies the conditions for an IV given above. In this special case, z completely determines x (subjects comply perfectly with their assigned treatment), so that x cannot be correlated with u. As x is exogenous in this case, the IV estimator is the same as the OLS estimator.

Example 2: RCT with Imperfect Compliance

Now consider a common problem with RCTs: suppose some subjects who are assigned to receive the standard therapy nevertheless take the new therapy; others assigned to receive the new therapy actually take the standard therapy. Generally, the difference in sample means across the treatment and control groups reflects both the causal effect of treatment and nonrandom selection into treatment, so it cannot be used to estimate the treatment effect. Assuming that assignment, z, affects the treatment decisions, X, of at least some people, treatment is not randomized because of the noncompliers, but it is *quasirandomized* in the sense that some of the variation in treatment status is a result of the coin toss. In the case with no other covariates, it is possible to show that the IV estimator defined in eqn [6] takes the form

$$\hat{\beta}_{\mathrm{IV}} = \frac{\overline{y}_{z=1} - \overline{y}_{z=0}}{\overline{x}_{z=1} - \overline{x}_{z=0}} \qquad [14]$$

where $\overline{y}_{z=i}$ denotes the sample mean of the outcome y in the subpopulation for which the assigned treatment status was i. The numerator is the difference in the average outcome between those assigned to the new therapy and those assigned to receive the standard therapy, regardless of the realized treatment status. This is the key object in 'intention to treat' analysis common in the medical literature. The denominator is the difference in the proportion who receive the new therapy across those assigned to new therapy and those assigned to the standard therapy. Note that the denominator is equal to one if compliance is perfect.

Example 3: The Causes of the Cholera Outbreaks in Victorian Era London

Even if one cannot run an RCT, the real world sometimes provides a mechanism that comes close to the experimental ideal. Perhaps the earliest IV application was that by John Snow, an epidemiologist who was interested in the causes of the cholera outbreaks that afflicted residents of London, England in the 1800s. Snow's hypothesis, which was not widely accepted at the time, was that cholera is a waterborne pathogen. In particular, Snow suspected that cholera was transmitted via contaminated drinking water. He noticed that one supplier of London's drinking water provided water contaminated by raw sewage, whereas another supplier provided relatively clean water. The reason was that these suppliers sourced their water from different points along the Thames River, one downstream of the city's sewer discharge and one upstream. Hence, the first condition for a good IV was satisfied: the identity of water supplier (z) resulted in marked variation in the quality of water consumed by households (x). Moreover, the source of water supply appeared to be independent of u, the other sources of the incidence of cholera. This was important because the quality of the water piped to households, though an important determinant of the quality of water consumed by households (x), was not the only determinant. The level of hygiene and cleanliness also played a role and this varied with household socioeconomic status. However, Snow observed that both the suppliers served a wide crosssection of Londoners, rich and poor alike. Thus Snow's instrument z was independent of u, the other determinants of y. A comparison of the rates of cholera of households that were supplied by the two water providers provided convincing evidence in support of Snow's hypothesis.

Example 4: Efficacy of Healthcare Treatments without Experimental Randomization

Several studies have compared the effectiveness of different types of healthcare used to treat particular health conditions. Conventional approaches must contend with the possibility that more severely compromised patients may be steered to one treatment over another. IV methods present a way forward when there is a mechanism that causes exogenous variation in the treatment received.

Some analysts have used the 'differential distance' to travel to obtain a particular therapy to treat a given health condition. Differential distance is the distance from the patient's residence to the nearest healthcare facility providing the treatment of interest minus the distance from the patient's residence to the nearest facility that provides any form of care to treat the condition. The idea is that, particularly for urgent problems such as acute myocardial infarction, the patient receives treatment from the nearest facility, regardless of the illness severity. If the nearest facility happens to provide the treatment of interest (i.e., zero differential distance) then the patient is more likely to receive it. The longer the differential distance, the less likely the patient will receive the treatment of interest. Differential distance is an invalid instrument if particularly ill patients relocate to be close to facilities that provide the treatment of interest.

Other analysts have exploited the marked geographic or interprovider variations in medical practice patterns that appear to be unrelated to medical need or patient preferences. These variations were first noted by Glover; he highlighted the striking geographic differences in the rate of tonsillectomy among British school districts. The literature, however, is most closely associated with the small-area variations research of Jack Wennberg. Brookhart, Rassen, and Schneeweiss review the ways in which analysts have used these variations to implement IV estimation of comparative treatment effectiveness. They note that to successfully implement IV, the practice variations must be independent of u, the unmodeled factors that affect patient health outcomes. These include the background characteristics of the patients themselves. It cannot be the case, for instance, that patients with particularly high values of u gravitate toward providers who tend to use the treatment under study. Moreover, practice style must affect health outcomes only through its influence on the treatment under study. Thus, providers who preferentially use one treatment must be of comparable quality and skill to those who preferentially use another treatment.

A third source of exogenous variation is changes over time in the availability of treatments. For instance, a new drug may become approved for use, or, conversely, a drug may be withdrawn from the market for safety reasons. Access to a treatment might also be temporarily impeded. For example, Evans and Lien use the disruption in the availability of public transit due to a bus strike to assess the impact of the use of prenatal care on birth outcomes. They focused on individuals for whom the disruption in bus service would impede access to prenatal care: pregnant black inner-city women. Analyses of this sort require a comparison of outcomes between two periods of time. To implement IV, the expected value of u must be the same in both periods. As Brookhart and colleagues note, to ensure that this condition holds, IVs based on calendar time are most reasonable in situations where a dramatic change in treatments occurs over a relatively short period of time.

Example 5: Effect of Education on Health

Return to the motivating example in the opening paragraph: To estimate the causal effect of an additional year of education on some measure of health status. Correlations or partial correlations between health and education do not reveal this causal effect because many personal and contextual characteristics (such as intelligence, conscientiousness, and family wealth) affect both health and education and are unobservable to the researcher, and because poor health while young may 'reverse' cause poor educational outcomes. That is, the effect of education on health is hard to estimate because of confounding on unobservables and because of 'reverse' causation. Neither conventional regression models such as OLS or logit nor matching estimators recover the causal effect of interest, and controlled experimentation on educational outcomes is restricted by both cost and ethical concerns.

In an influential study, UCLA economist Adriana Lleras-Muney employed an IV strategy to address this problem. She estimated regressions in which mortality is the health outcome of interest. Using large samples from the US census, she matched cohorts to the number of years of compulsory schooling specific to each combination of state government and year. Years of compulsory schooling acts IV: It is plausible that the only reason a change in years of compulsory schooling affects health is because (for some students) changes in years of compulsory schooling affects realized years of schooling. Intuitively, Lleras-Muney asks, "Is an adult who was required by law to take more schooling healthier, on average, than a statistically identical adult required to take less schooling?" Her estimates suggest that an additional year of schooling causes as much as a 1.7 year increase in life expectancy at the age of 35 years.

Problems with Instrumental Variables Estimation

In theory, it is easy to write down conditions (10) and (11) and derive that an estimator satisfying these conditions can recover causal effects from observational data. In practice, finding variables that satisfy those conditions can be very difficult or impossible. Worse, it turns out that even small deviations from those conditions can yield estimators with extremely poor properties.

The most difficult problem to overcome is instruments which are themselves endogenous, that is, correlated with the error term in the equation of interest, violating condition (10). It is possible to show that the IV estimator is inconsistent when the instruments are endogenous. Intuitively, if our condition that the only reason y varies with z is because z causes x fails, then observing that z and y move together is not evidence that x causes y.

For most problems finding variables that only affect the outcome of interest because they affect the endogenous

regressors is challenging. Consider, for example, one of the key problems in the social determinants of health literature: estimating the causal effect of personal income on health. A variable is required which affects health solely through its effect on income. It is unlikely that any personal characteristic satisfies that condition: personal characteristics such as education, smoking status, or cognitive ability all affect income, but all potentially affect health conditional on income, so none are valid instruments. Regional characteristics such as the unemployment rate may affect income, but may also affect health through other channels, such as provision of local public goods or through sorting of people across states. Researchers therefore need to be creative in finding valid instruments: one study, for instance, uses lottery winnings as an exogenous source of income to assess the effect of income on the health of lottery players. In other applications, valid instruments may simply not be available.

It may seem that variables which are almost, but not quite, exogenous may yield reasonable estimates, provided that there is a large sample and can thus rely on the consistency property of the IV estimator. In particular, from the formula for the probability limit of the univariate IV estimator presented above, it is possible to show

$$\text{plim}_{n\to\infty}\hat{\beta}_{IV} = \frac{\text{Cov}(y,z)}{\text{Cov}(z,x)} = \beta + \frac{\text{Cov}(z,u)}{\text{Cov}(z,x)} \qquad [15]$$

As long as Cov(z, u) is close to zero, then the ratio of Cov(z, u) to Cov(z, x) should itself be close to zero. This intuition is correct provided that Cov(z, x) is sufficiently large. If, however, there is only weak correlation between z and x then even small violations of exogeneity lead to very poorly behaved estimates. The reason is that Cov(z, u) is divided by a number close to zero, which has the effect of amplifying Cov(z, u). The result is that the IV estimator $\hat{\beta}_{\text{IV}}$ can be centered on a value wildly different from the true value of β, even as the sample size grows arbitrarily large. A low level of correlation between the instruments and treatment is known as the 'weak instrument problem.'

What is more, even if the instruments are exogenous, if the instruments are weak the IV estimator will tend to be badly biased in finite samples and, perhaps worse, the usual estimator of the covariance matrix, and test statistics based on that matrix, will be biased, leading to severe size and power distortions. The bias stems from the fact that the IV estimator is the ratio of two estimators – the numerator being the estimator of the effect of z on y and the denominator the estimator of the effect of z on x. In large samples, these estimators converge to their population quantities. In finite samples, however, sampling error in the two estimators can cause the ratio to behave erratically. The weaker the instruments, the greater is the sampling error.

In short, instruments with poor properties – either endogenous or weak – may be 'cures worse than the disease.' The good news is that in overidentified models it is possible to construct test statistics against the null that the instruments are exogenous, and it is always possible to test the strength of the instruments.

Heterogeneous Causal Effects

Over the past two decades the IV literature has focused on the following issue: If different entities or 'units' (people, firms, hospitals, etc.) experience different causal effects as a result of the same treatment, how are we to interpret IV estimates? It turns out that when treatment effects are heterogeneous, identification of causal effects using IV can be challenging.

Consider a slight modification to eqn [1],

$$y_i = \beta_i x_i + u_i \qquad [16]$$

which differs from eqn [1] only in that the slope coefficient β_i may vary arbitrarily across units. In the interest of simplicity, again suppose x_i is a binary indicator of whether unit i received treatment.

In this model, it is incoherent to refer to 'the' causal effect of x on y, as each unit generally experiences a different causal effect. Estimation of counterfactual outcomes in this model is also more complicated than in model (1). When treatment effects are constant, the outcomes of untreated units can be used to infer the counterfactual outcomes of those that were treated (and vice versa). This is not generally possible when causal effects vary across i. Therefore it is not possible to estimate the effect of treatment for any given unit. Researchers instead attempt to estimate features of the distribution of the causal effect, β_i, such as the population average treatment effect, $E(\beta_i)$, or the average treatment effect for those who actually received the treatment, $E(\beta_i \mid x_i = 1)$.

Without loss of generality, write $\beta_i = \overline{\beta} + \varepsilon_i$, where $\overline{\beta}$ is the population mean effect and ε_i is a zero-mean idiosyncratic effect specific to unit i. Substituting into eqn [16],

$$y_i = \overline{\beta} x_i + [x_i \varepsilon_i + u_i] \qquad [17]$$

Notice that the error term contains two components: unobserved causes of the outcome specific to unit i, μ_i, and the interaction between treatment status and unit i's return to treatment. If both μ_i and ε_i are uncorrelated with x_i, OLS estimation is consistent for the average treatment effect, $\overline{\beta}$. However, even when μ_i is uncorrelated with x_i, correlation between ε_i and treatment status creates an endogeneity problem and OLS does not recover the average treatment effect. In this case, 'essential heterogeneity' is said to exist. Essential heterogeneity commonly occurs in observational studies of treatment efficacy when individuals with the most to gain from taking a particular treatment are more likely to receive that treatment. Essential heterogeneity can also exist in RCTs with imperfect compliance. This occurs if subjects are able to: (1) determine the treatment to which they have been assigned, (2) predict better than chance which treatment will benefit them most, and (3) if advantageous, switch therapies. Condition (1) occurs if subjects are not blinded or if they are blinded, subjects can infer treatment status from side effects, or other physiological clues. The extent to which condition (3) holds depends on the context. Subjects assigned to the new therapy who wish to use the standard therapy can presumably obtain the standard therapy outside the trial. Conversely, subjects assigned to the standard therapy who wish to use the new therapy might be able to obtain the new therapy from friends enrolled in the trial.

Estimation using IV is complicated by essential heterogeneity. The instrument must be correlated with treatment status: It must move some people into or out of treatment. Even if all of the conditions defined in section The Instrumental Variables Estimator hold, the properties of the IV estimator depend on which people get moved into or out of treatment when treatment effects vary across people. Consider again example 2 in section The Instrumental Variables Estimator above, an RCT with imperfect compliance. Under a condition called monotonicity, which requires that there be no 'defiers' – people who only receive treatment if they are assigned not to receive treatment or vice versa – it is possible to show that the IV estimator converges to the average causal effect of treatment of compliers, that is, subjects who use the treatment that they were assigned to. This is called the 'local average treatment effect' arising from this treatment.

Intuitively, some people will always take the new treatment and others will always take the standard treatment, regardless of the assignment. The experiment does not change these people's behavior and therefore the experiment generates no information about the causal effects of treatment for these people. The IV estimator depends solely on the outcomes of subjects whose treatment status was experimentally manipulated; the estimator tells us the average effect only for that (unobservable) subpopulation. If the instrument takes many values instead of just two, it is possible to show that (under monotonicity) the IV estimator converges to a difficult-to-interpret weighted average of local treatment effects, in which units for which treatment status is most responsive to variation in the instruments receive the highest weights.

In addition to complicating the interpretation of conventional IV estimates, heterogeneous causal effects complicate specification testing. Most tests of the assumption that the instruments are exogenous are based on stability of the estimates as different sets of instruments are used to construct the estimator. Under homogeneous responses, all of these estimates converge to the causal effect. When effects are heterogeneous, different instruments recover different weighted averages of local effects, and will differ even if the classical conditions (10) and (11) hold, so rejection of the null can no longer be interpreted as evidence that the instruments are endogenous.

Example: Reinterpreting an Estimate of the Effect of Education on Health

Consider again example 5, above, of research using IV on the effect of education on health. Earlier, Lleras-Muney's estimates were interpreted as suggesting that an additional year of education causes an increase in life expectancy of 1.7 years at the age of 35 years. Lleras-Muney's estimates are based on variation in compulsory schooling laws, so she interprets her IV estimates as: among the subpopulation who only receive additional education if and only if they are forced to do so by law, an additional year of education increases life expectancy by 1.7 years at the age of 35 years. This subpopulation may experience substantially different health returns to education than other people who choose to go on to receive more than the legally mandated minimum schooling. Thus, Lleras-Muney's local average effect may not reflect the health returns to education for other groups. However, Lleras-Muney's estimates may be more relevant than results from a hypothetical RCT randomizing education if policy questions hinge on effects experienced by people whose educational outcomes are affected by changes in compulsory schooling laws, as the RCT would recover population average effects rather than effects for the subpopulation affected by policy changes.

See also: Instrumental Variables: Methods

Further Reading

Brookhart, M. A., Rassen, J. A. and Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety* **19**(6), 537–554.

Brooks, J., Irwin, C., Hunsicker, L., et al. (2006). Effect of dialysis center profit-status on patient survival: A comparison of risk adjustment and instrumental variable approaches. *Health Services Research* **41**, 2267–2289.

Evans, W. and Lien, D. (2005). The benefits of prenatal care: Evidence from the PAT bus strike. *Journal of Econometrics* **125**, 207–239.

Glover, J. (1938). The incidence of tonsillectomy in school children. *Proceedings of the Royal Society of Medicine* **31**, 1219–1236.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* **88**, 389–432.

Jones, A. (2009). Panel data methods and applications to health economics. In Mills, T. and Patterson, K. (eds.) *Palgrave handbook of econometrics*, vol. 2, pp. 557–631. London: Palgrave MacMillan.

Jones, A. and Rice, N. (2011). Econometric evaluation of health policies. In Glied, S. and Smith, P. (eds.) *Oxford handbook of health economics*, vol. 1, pp. 890–923. Oxford: Oxford University Press.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *Review of Economic Studies* **72**(1), 189–221.

McClellan, M., McNeil, B. and Newhouse, J. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**, 859–866.

McConnell, K., Newgard, C., Mullins, R., Arthur, M. and Hedges, J. (2005). Mortality benefit of transfer to level I versus level II trauma centers for headinjured patients. *Health Services Research* **40**, 435–457.

Pop-Eleches, C. (2006). The impact of an abortion ban on socioeconomic outcomes of children: Evidence from Romania. *Journal of Political Economy* **114**, 744–773.

Stock, J. H., Wright, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* **20**(4), 518–529.

Tan, H., Norton, E., Ye, Z., et al. (2012). Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. *Journal of the American Medical Association* **307**, 1629–1635.

Wennberg, J. (2008). Commentary: A debt of gratitude to J. Alison Glover. *International Journal of Epidemiology* **37**, 26–29.

Xian, Y., Holloway, R., Chan, P., et al. (2011). Association between stroke center hospitalization for acute ischemic stroke and mortality. *Journal of the American Medical Association* **305**, 373–380.

Instrumental Variables: Methods

JV Terza, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

Introduction

Most empirical research in health economics is conducted with the goal of providing causal evidence of the effect of a particular variable (the causal variable – X) on an outcome of interest (Y). Such analyses are typically conducted in the context of explaining past behavior, testing an economic theory, or to evaluating a past or prospective policy. Common to all such applied contexts is the need to infer the effect of a counterfactual *ceteris paribus* exogenous change in X on Y, using statistical results obtained from survey data in which observed differences in X are neither *ceteris paribus* nor exogenous. In such nonexperimental sampling circumstances, statistical methods that essentially measure observed differences in Y per observed differences in X typically miss the mark because they fail to control for unobserved variables that are correlated in sampling with both X and Y. Such unobserved confounding variables, which vary in sampling with both X and Y, obfuscate the true causal effect (TCE) as it would have manifested if the value of X were exogenously perturbed *ceteris paribus*. Consider, for instance, attempting to obtain inference regarding the effect of cigarette smoking during pregnancy on infant birth weight using survey data. Suppose there exists an unobserved variable, say 'health mindedness' that causes pregnant women to both refrain from smoking and engage in other healthy prenatal behaviors. In such a scenario it is possible that observed smoking levels could be negatively associated with birth weight even though a *ceteris paribus* exogenous change in smoking (as might be brought about through policy intervention) would have no causal effect on birth weight. The present article discusses available regression methods designed not only to control for observable confounding influences but also to account for the presence of unobservables that would otherwise thwart causal inference.

The remainder of the article is organized as follows. The next section offers a more formal discussion of estimation bias due to unobserved confounding. In Section Instrumental Variables Methods, we consider a commonly implemented remedy for such bias – the use of instrumental variables (IV). Therein extant IV methods for both linear and nonlinear models are reviewed. The article concludes with a summary and some recommendations.

Unobserved Confounder Bias

At issue here is the presence of confounding variables which serve to mask the TCE of X on Y. The author begins by defining a confounder as a variable that is correlated with both Y and X. Confounders may be observable or unobservable (denoted C_o and C_u, respectively, – in the present discussion both are assumed to be scalars (i.e., not vectors)). In modeling Y, if the presence of C_u cannot be legitimately ruled out, then X is said to be endogenous. Observations on C_o can be obtained from the survey data, so its influence can be controlled in estimation of the TCE. C_u, however, cannot be directly controlled and, if left unaccounted for, will likely cause bias in statistical inference regarding the TCE. This happens because estimation methods that ignore the presence of C_u will spuriously attribute to X observed differences in Y that are, in fact, due to C_u. The author refer to such bias as unobserved confounder bias (henceforth C_u–bias) (sometimes called endogeneity bias, hidden selection bias, or omitted variables bias). One can formally characterize C_u–bias in a useful way. For simplicity of exposition, the author casts the true causal relationship between X and Y as linear and write

$$Y = X\beta + C_o\beta_o + C_u\beta_u + e \qquad [1]$$

where β is the parameter that captures the TCE, β_o and β_u are parametric coefficients for the confounders, and e is the random error term (without loss of generality, it can be assumed that the Y intercept is 0). In the naive approach to the estimation of the TCE (ignoring the presence of C_u), the ordinary least squares (OLS) method is applied to

$$Y = Xb + C_o b_o + \varepsilon \qquad [2]$$

where the b's are parameters and ε is the random error term. The parameter b is taken to represent the TCE. It can be shown that OLS will produce an unbiased estimate of b (here and henceforth, when the author refers to unbiasedness it is done so in the context of large samples). It is also easy to show, however, that

$$b = \beta + b_{XC_u}\beta_u \qquad [3]$$

where b_{XC_u} is a measure of the correlation between C_u and X. As is clear from eqn [3], C_u – bias in OLS estimation is $b_{XC_u}\beta_u$, which has two salient components: the correlation between the unobserved confounder and the causal variable of interest and the correlation between the unobserved confounder and the outcome. Equation [3] is helpful because it can be used to diagnose potential C_u – bias. Consider the smoking (X) and birth weight (Y) example discussed in the Section Introduction, in which C_u is health mindedness. In this case one would expect that b_{XC_u} would be negative and that β_u would be positive. The net effect of which would be negative C_u – bias in the estimation of the TCE via OLS.

Clearly an approach to estimation is needed that, unlike OLS, does not ignore the presence and potential bias of C_u. One such approach exploits sample variation in a particular type of variable (a so-called IV) to eliminate bias due to correlation between C_u and X (C_u – bias as characterized in eqn [3]). This is the subject of the following section.

Instrumental Variables Methods

As eqn [3] demonstrates, if the correlation link between the causal variable and the unobservable confounder were somehow broken, concomitant estimation bias would be eliminated. If the researcher could exert control over the sampled values of X, then such disjunction of C_u and X could be accomplished by random assignment of X values to the individual sample members. Under such randomization, b_{XC_u} would be equal to zero, by eqn [3] b would be equal to β, and conventional estimation methods like OLS, which ignore the presence of C_u, would be unbiased. Unfortunately, in applied health economics and health services research, as in other social sciences, explicit randomization (experimentation) is often prohibitively costly or ethically infeasible. A form of pseudorandomization is, however, possible in the context of survey (nonexperimental) data. If, for instance, a variable that is observed as one of the survey items is highly correlated with X but correlated with neither Y nor C_u (except through its correlation with X), then the sample variation (across observations) in the value of that variable can be viewed as providing variation in X that is not correlated with C_u – a kind of pseudorandomization for X. Such a variable is typically called an IV. In the context of our smoking birth weight example, cigarette tax is an arguably valid IV in that it should be highly correlated with cigarette consumption but not directly correlated with birth weight.

IV estimation methods all require observable confounder (C_o) control – typically within a regression framework akin to eqn [1]. Most often, however, the linear regression model in eqn [1] is not realistic in that it precludes cases in which the relationship between Y and the right-hand side variables (X, C_o, and C_u) is nonlinear – for example, when Y is limited in range (e.g., nonnegative and binary outcomes); and/or when such characteristics of the outcome induce interactions among the causal variable and confounders. In the following, the presence of a valid IV (call it W) in the relevant survey is assumed and IV estimation methods for both linear and nonlinear contexts are considered.

Instrumental Variables Estimation in Linear Models

By way of motivating the conventional linear IV estimator in the context of eqn [1], the author examines the underpinnings of the OLS estimator of the TCE for the case in which $\beta_u=0$ (i.e., the case in which there is no unobservable confounder). When $\beta_u=0$, eqn [1] becomes

$$Y = X\beta + C_o\beta_o + e \qquad [4]$$

and the formulation of the OLS estimator of β (and β_o), which involves data on observable variables only (viz., X and C_o), can be derived from the fact that X and C_o are not correlated with the error term e. A similar tack cannot, however, be taken when $\beta_u \neq 0$. In this case, eqn [1] can be rewritten as

$$Y = X\beta + C_o\beta_o + e^* \qquad [5]$$

where $e^*=C_u\beta_u+e$, and although C_o and e^* are arguably uncorrelated, the correlation between X and e^* is clearly nonzero because X and C_u are, by the definition of the term confounder, correlated. As a consequence of the undeniable correlation between X and C_u, the aforementioned derivation of the OLS estimator cannot be replicated for eqn [5]. This approach is not, however, entirely futile if an IV (W) is available in the data. By definition, the IV W is uncorrelated with both C_u and e. W is, therefore, not correlated with e^* so, analogous to the derivation of the OLS estimator based on eqn [4], it can be used to formulate an unbiased estimator of β and β_o (the so-called IV estimator). The IV estimator is available in all of the most widely used statistical and econometric software packages (e.g., Stata and SAS).

There are two relatively more intuitive two-stage versions of the IV estimator. Both of these approaches implement an auxiliary regression of the form

$$X = C_o\alpha_o + W\alpha_w + C_u \qquad [6]$$

where the α's are parameters. In the first stage of each of these methods, OLS is applied to eqn [6] to obtain estimates of parameters ($\hat{\alpha}_o$ and $\hat{\alpha}_w$) and the regression predictor of X ($\hat{X} = C_o\hat{\alpha}_o + W\hat{\alpha}_w$). One of these methods, called two-stage least squares (2SLS) has as its second stage the OLS estimation of β and β_o via eqn [5] with $\hat{X}$ substituted for X. The other approach, called two-stage residual inclusion (2SRI) calls for OLS estimation of

$$Y = X\beta + C_o\beta_o + \hat{C}_u\beta_u + e \qquad [7]$$

where $\hat{C}_u = X - (C_o\hat{\alpha}_o + W\hat{\alpha}_w)$ – i.e., the residual from first-stage OLS estimation of eqn [6].

When true causal model is eqn [1] both 2SLS and 2SRI produce estimates of the TCE (β) and β_o that are identical to those obtained via the IV estimator.

Instrumental Variables Estimation in Nonlinear Models

Although the linear IV estimator (or its equivalent versions 2SLS or 2SRI) is intuitive and simple to apply due to its availability, the linear true causal model (as specified in eqn [1]) on which it is based does not conform to most empirical contexts in health economics. In most applied settings, the range of the outcome is limited in a way that makes a nonlinear specification of the true causal model more sensible. For example, the researcher is often interested in estimating the causal effect of a policy variable (X) on whether or not an individual will engage in a specified health-related behavior. In this case, the outcome of interest is binary so that a nonlinear specification of the true causal model would likely be more appropriate. In the smoking birth weight example discussed in the Section Introduction, the outcome of interest (birth weight) is nonnegative and an exponential regression specification of the true causal model is more in line with this feature of the data than is the linear specification in eqn [1]. Another common example of inherent nonlinearity in health economics and health services research, is in the modeling of healthcare expenditure or utilization (E/U). It is typical to observe a large proportion of zero values for the E/U outcome. In this and similar empirical contexts, the two-part model (2PM) has been widely implemented. The 2PM allows the process governing observation at zero (e.g., whether or not the individual uses the healthcare service) to systematically differ

from that which determines nonzero observations (e.g., the amount the individual uses (or spends on) the service conditional on at least some use). The former can be described as the hurdle component of the model, and the latter is often called the levels part of the model. Both of these components are nonlinear – binary response model for the hurdle; nonnegative regression for E/U levels given some utilization.

To accommodate these and other cases, the generic nonlinear version of the true causal model in eqn [1] is written as

$$Y = \mu(X, C_o, C_u;\ \theta) + e \qquad [8]$$

where $\mu(X,\ C_o,\ C_u;\ \theta)$ is known except for the parameter vector θ. It is very often assumed that $\mu(X,\ C_o,\ C_u;\ \theta) = M(X\beta + C_o\beta_o + C_u\beta_u)$, where $M(\ )$ is a known function and $\theta = (\beta\ \beta_o\ \beta_u)$. In this linear index form the true causal models corresponding to binary and nonnegative outcomes are commonly written, respectively, as

$$Y = F(X\beta + C_o\beta_o + C_u\beta_u) + e\ (Y = \{0,\ 1\}) \qquad [9]$$

and

$$Y = \exp(X\beta + C_o\beta_o + C_u\beta_u) + e\ (Y \geq 0) \qquad [10]$$

where $F(\ )$ is a function whose range is the unit interval. It is noted here that for the generic nonlinear model characterized by eqn [8] the TCE is not embodied in any particular parameter (e.g., β) as in the linear models defined by eqn [1]. Instead, the TCE will be a nonlinear function of all parameters (θ) and all of the right-hand side variables (X, C_o, C_u) of the model. Moreover, the exact form of the TCE in nonlinear settings will differ depending on the researcher's policy relevant analytic objective(s). These issues will not, however, be discussed here. In the present discussion, focus is on estimation of the vector of parameters θ.

In the remainder of this section, various approaches to the estimation of θ in nonlinear models of the generic form given in eqn [8] are examined. The author begins by examining the feasibility and appropriateness of the generalized method of moments (GMM) estimator – the nonlinear analog to IV estimation in the linear model. Next, the nonlinear counterparts to the linear 2SLS and 2SRI are examined. Nonlinear 2SRI (N2SRI) is a member of a class estimators called control function estimators. Other control function estimators that are specifically designed for cases involving binary causal variables are discussed. This section concludes with a description of cases in which the maximum-likelihood method can be applied.

The generalized method of moments

To estimate of the parameters of nonlinear causal models like eqn [8], one may seek to apply the GMM as an extension of the linear IV approach, detailed in Section Instrumental Variables Estimation in Linear Models. Recall that the derivation in that section relied on two facts:

1. Equation [1] could be rewritten as eqn [5] – a linear regression representation involving observable variables only and an additive error term.
2. The IV W is correlated with neither C_u (the unobservable confounder) nor e (the random error term in eqn [1]).

Unfortunately, there is only one case (that we know of) in which such a derivation is feasible in the context of eqn [8] – the exponential regression version of the model given in eqn [10]. This model is discussed later. In (all?) other cases, it is the nonadditive involvement of C_u in eqn [8] that makes the derivation of a GMM-type estimator infeasible. The generic nonlinear form of $\mu(\ )$ precludes reformulation of the model as the sum of a nonlinear parametric component in the observable right-hand side data (X and C_o) with an additive error term. Some have suggested the use of an approximation to eqn [8] in which C_u is artificially cast in an additive role in the respecification of the model. For example, following this approach, models like eqn [9] would be rewritten as:

$$Y = F(X\alpha + C_o\alpha_o) + C_u\alpha_u + e^{\dagger}\ \{(Y = \{0,\ 1\}) \qquad [11]$$

In which case, the IV condition that W is correlated with neither C_u nor $e^{\dagger}$ would be sufficient to establish the appropriate GMM estimator. Clearly, however, eqns [9] and [11] are not equal; and the argument in favor of eqn [11] as a good approximation to eqn [9] is, at best, strained. Moreover, TCE estimation methods that incorporate GMM results obtained from such additive approximations are clearly biased. The extent of this bias has yet to be investigated.

As mentioned earlier, the only nonlinear context (of which one is aware) in which conditions like (1) and (2) are sufficient for derivation of an unbiased (in large samples) GMM estimator is the linear-index exponential case given in eqn [10]. Not only does this GMM estimator yield unbiased estimates of β and β_o but also unlike the additive approximations discussed earlier and exemplified in eqn [11], the exponential GMM results can be used to obtain unbiased estimates of the various policy relevant versions of the TCE.

Two-stage control function methods

In the Section The generalized method of moments, it is noted that extending the linear IV method to the generic nonlinear model in eqn [8] (i.e., the GMM estimator) is not generally feasible. Therefore, aside from the exponential case, we need a desirable (unbiased) feasible alternate to GMM. In search for such an alternative one turns to the discussion of the linear model in Section Instrumental Variables Estimation in Linear Models wherein the 2SLS and 2SRI estimators for β and β_o in eqn [1] are detailed. These estimators yield results identical to those produced by the linear IV method. Consider the feasible nonlinear analogs to linear 2SLS and 2SRI estimation. In the generic nonlinear context eqn [8] is supplemented with the following nonlinear analog to eqn [6]

$$X = r(C_o, W; \alpha) + C_u \qquad [12]$$

In 2SLS and 2SRI, the parameters of eqn [12] (α) are first estimated using an appropriate nonlinear regression estimator (e.g., nonlinear least squares (NLS)) and the following predictor of X is computed

$$\hat{X} = r(C_o, W;\ \hat{\alpha}) \qquad [13]$$

where $\hat{\alpha}$ denotes the parameter estimates. In the second stage of the nonlinear analog to 2SLS, an appropriate nonlinear regression estimator (e.g., NLS) would be applied to eqn [8]

with the predictor $\hat{X}$ substituted for X (this has also been called the two-stage predictor substitution (2SPS) estimator). In the second stage of the nonlinear analog to 2SRI, instead of substituting the predictor for X in eqn [8], C_u is replaced by the residual from eqn [13] ($\hat{C}_u = X - r(C_o, W; \hat{\alpha})$) and an appropriate nonlinear regression estimator (e.g., NLS) is applied to the following version of eqn [8]

$$Y = \mu(X, C_o, \hat{C}_u; \theta) + e^{2SRI} \quad [14]$$

where e^{2SRI} is the relevant regression error term. Unlike the linear case, the 2SPS and 2SRI estimators are not identical. Note that the actual value of X is used in eqn [14]. The 2SRI estimator is generally unbiased but the 2SPS estimator is not.

The 2SRI estimator is member of a general class of models called control function methods in which a specified function of the IV (W) (and some parameters) is used to 'control' for unobserved confounder bias. In the special (but very common) case in which X is binary, an alternative control function method is available. In this alternative control function framework eqns [8] and [12] are respectively replaced by

$$Y = \mu(X, C_o, C_u^*; \theta) + e \quad [15]$$

and

$$X = I(C_o\alpha_o + W\alpha_w + C_u^* > 0) \quad [16]$$

where $I(A)$ is equal to 1 if condition A holds and 0 otherwise, and the probability distribution of C_u^* is known. For example, if C_u^* is assumed to be logistically distributed, eqn [16] defines a conventional logit model. Similarly if C_u^* is normal eqn [16] is tantamount to a probit model. Given the known distribution of C_u^*, it can be 'integrated out' of eqn [15] and the resultant regression form can be used as the basis for nonlinear estimation (e.g., NLS) estimation of θ. When eqn [15] is linear and C_u^* is normally distributed, this control function method coincides with the classical Heckman-type dummy endogenous variable model estimator. Note that both 2SRI and this nonlinear extension of the Heckman approach are feasible and unbiased when X is binary (assuming, of course, that the respective sets of underlying assumptions hold).

Maximum-likelihood methods

When Y is a binary probit outcome and C_u^* is normally distributed, the control function approach described in the Section Two-stage control function methods leads to the bivariate probit model. In this case, the parameters of the model can be estimated using the maximum-likelihood method. Maximum-likelihood methods are also available for the special case in which the auxiliary regression is linear (akin to eqn [6]) and the outcome regression is a normal-based limited dependent variable model (e.g., probit or Tobit). These methods require joint normality of the random error terms in the outcome and auxiliary regressions.

Common factor models have also been suggested for the case in which X is qualitative. In these models, conditional on an unobserved 'common factor' (and the other conditioning variables), Y and X are assumed to be independently distributed. Moreover, these independent distributions and the distribution of the common factor are assumed to be of known form. The maximum-likelihood method can be used to obtain estimates of the parameters in this framework.

Summary

The most widely applied remedy for endogeneity in a causal modeling framework is the conventional linear IV (LIV) estimator described in Section Instrumental Variables Estimation in Linear Models. The popularity of LIV can be attributed to its off-the-shelf software availability, and to its intuitive appeal when cast as a two-stage method – 2SLS or 2SRI. The most attractive feature of LIV is that it need not be estimated in two stages and therefore does not require the specification of an auxiliary regression like eqn [6]. Very often, however, in applied health economics and health services research, endogeneity must be confronted in inherently nonlinear empirical contexts. For example, binary response outcomes, limited dependent variables, and two-part models with endogenous causal regressors abound in these fields. One might think that the GMM, which is the most direct approach to extending the LIV estimator to the nonlinear case, would provide a solution to the unobserved confounding problem in nonlinear models. Unfortunately, except for exponential regression models, the GMM is not feasible as a means of dealing with endogeneity in nonlinear settings.

The easiest to implement approach for such cases is the extension of the linear 2SRI estimator to nonlinear models. The primary drawback to the use of N2SRI is that it requires the specification and estimation of an auxiliary regression as defined in eqn [12]. The main advantages of N2SRI are that it can be applied in any nonlinear regression context and will produce unbiased estimates of the regression parameters (and, therefore, the relevant TCE) under general conditions.

There are alternatives to N2SRI for some specific cases. When the outcome is binary, the nonlinear extension to Heckman-type control functions can be used. These methods, although feasible, are not as simple to apply as N2SRI. A similar criticism holds for the maximum-likelihood common factor models.

When the outcome is limited in range (e.g., probit and Tobit) and the auxiliary regression is linear, maximum-likelihood methods can be applied. These methods, though packaged in Stata and therefore easy to apply, require the relatively strong assumption of joint normality between the outcome and the causal variable. N2SRI imposes no such joint distribution assumptions. Moreover, it is often difficult to justify the linearity of the auxiliary regression and the implied normality of the causal variable. It is typical, that the causal variable will itself be limited in range (e.g., binary and nonnegative), making both linearity and normality implausible.

Simulation-based performance comparisons of the models discussed in this article have yet to be conducted.

See also: Instrumental Variables: Informing Policy. Modeling Cost and Expenditure for Healthcare. Models for Count Data. Models for

Discrete/Ordered Outcomes and Choice Models. Sample Selection Bias in Health Econometric Models

Further Reading

Blundell, R. W. and Smith, R. J. (1989). Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economics and Statistics* **56**, 37–58.

Blundell, R. W. and Smith, R. J. (1993). Simultaneous microeconometric models with censored or qualitative dependent variables. In Maddala, G. S., Rao, C. R. and Vinod, H. D. (eds.) *Handbook of statistics*, vol. 2, pp. 1117–1143. Amsterdam: North Holland Publishers.

Deb, P. and Trivedi, P. K. (2006). Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization. *Econometrics Journal* **9**, 307–331.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46**, 931–959.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics* **79**, 586–593.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* **39**, 347–366.

Smith, R. J. and Blundell, R. W. (1986). An exogeneity test for a simultaneous equation Tobit model with an application to labor supply. *Econometrica* **54**, 679–685.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* **84**, 129–154.

Terza, J. V. (2006). Estimation of policy effects using parametric nonlinear models: A contextual critique of the generalized method of moments. *Health Services and Outcomes Research Methodology* **6**, 177–198.

Terza, J. V. (2009). Parametric nonlinear regression with endogenous switching. *Econometric Reviews* **28**, 555–580.

Terza, J. V., Basu, A. and Rathouz, P. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* **27**, 531–543.

Interactions Between Public and Private Providers

C Goulão, Toulouse School of Economics (GREMAQ, INRA), Toulouse, France
J Perelman, Universidade Nova de Lisboa (UNL), Lisbon, Portugal

Glossary

Asymmetry of information A situation in which the parties to a transaction have different amounts or kinds of information, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances, or people seeking insurance have more reliable expectations of their risk exposure than insurance companies.

Coinsurance Coinsurance is the practice whereby the insured person shares a fraction of an insured loss with the insurer. For example, the insurance policy may require the insured person to pay 10 per cent of the expenses of medical care, with the insurer paying 90 per cent. The sum paid by the insured person is known as a copayment, so if the expenses are US$1000 and the coinsurance rate is 10 per cent, the copayment is US$100.

Copayment An arrangement, whereby an insured person pays a particular percentage of any bills for health services received, the insurer paying the remainder.

Deductible An insurance arrangement, under which the insured person pays a fixed sum, when healthcare is used in any year and the insurer pays all other expenses (usually with further copayments). Thus, if the deductible is US$100 and the coinsurance rate 10 per cent, should the event involve an expense of US$1000, the insured person pays US$190 (US$100 plus US$90 copayment).

Dual practice A combination of public and private practice by doctors, sometimes even within the same hospital.

Gatekeeping The process by which a professional, usually a general practitioner, select patients and guides them into secondary care. In many countries patients, other than emergency cases, cannot consult a specialist without being referred by a general practitioner.

Horizontal equity Horizontal equity is treating equally those who are equal in some morally relevant sense. Commonly met horizontal equity principles include 'equal treatment for equal need' and 'equal treatment for equal deservingness'. When applied to insurance, the notion that two individuals facing the same risks should have access to the same coverage at the same premium.

Infant mortality rate Deaths in one year of infants under one year of age divided by number of live births in that year, all multiplied by 1000.

Life expectancy The statistically expected remaining years of life for a representative person (usually in a specific jurisdiction and by subgroup – male, female, by ethnicity, etc.) at a given age (say, at birth, or having already reached 65), assuming that age-specific mortality remains constant.

Moonlighting Same as dual practice.

Moral hazard Moral hazard can occur when the insurer has imperfect information on the likely behavior of insured individuals. There are two main types. Ex ante moral hazard refers to the effect that being insured has on safety behavior, generally increasing the probability of the event insured against occurring. Ex post moral refers to the possibility that insured individuals will behave in such a way after an insured event has occurred that will increase the claim cost to insurers, partly because the user price of care is lower through insurance and demand may therefore rise. It is also often related to insurance fraud.

Potential years of life lost (PYLL) A measure of the burden of disease, preventable premature mortality, or potential benefit from an effective intervention to improve health. Its calculation involves summing deaths occurring at each age and multiplying them by the number of remaining years up to a limiting age, which is often 70 years.

Propitious selection Propitious selection is a phenomenon in insurance which compares people with different levels of risk aversion. Those with higher levels are more likely both to buy insurance and to exercise care. Those with low levels, or who are actually risk seeking, will tend to do neither.

RAND experiment The largest social science empirical trial of health policy options ever conducted. The aim was to examine the effect of health insurance on health care costs, utilization, and outcomes. More than 1974–79 families were randomly assigned to different insurance plans with a variety of limits and coinsurance rates. Its principal investigator summarized the main results that "For most people enrolled in the RAND experiment, who were typical of Americans covered by employment-based insurance, the variation in use across the plans appeared to have minimal to no effects on health status. By contrast, for those who were both poor and sick – people who might be found among those covered by Medicaid or lacking insurance – the reduction in use was harmful, on average."

Risk aversion The most common definition in economics is the extent to which a sure and certain outcome is preferred to a risky alternative with the same expected value. For example, risk-averse individuals may prefer to have US$45 for sure than face a gamble in which they may win US$100 or nothing, each with a 50% chance.

 doi:10.1016/B978-0-12-375678-7.01317-1

Introduction

The existence of duplicate private health insurance (DPHI), which is observed in many countries with a National Health Service (NHS), is paradoxical at first sight. NHSs are usually characterized by universal coverage of every resident, large and comprehensive benefit packages, very low copayments or free care at the point of delivery, progressive tax financing, and are strongly guided by principles of equity in access to health care. Additionally, residents cannot opt out of the NHS, meaning that they are not given the option of not contributing to the NHSs' financing and relying exclusively on other forms of health care. Why then would people be willing to pay for private health insurance (PHI) covering roughly the same services as the NHS? This is even more surprising because NHSs have been generally performing quite well over the last decades. This paradox is a major issue in health economics, which health economists have been trying to understand theoretically and to document through empirical work. This article presents these findings.

Before going further, it is important to define the concept of DPHI, often called also double coverage or substitutive PHI. Under DPHI, private insurers offer coverage for health care already available under public delivery systems. Note that DPHI differs from supplementary PHI (SPHI). Under SPHI, patients access additional health services not covered by the public scheme such as luxury care, elective care, long-term care, dental care, pharmaceuticals, rehabilitation, alternative or complementary medicine, or superior hotel and amenity hospital services. Also worthy of remark is that DPHI is distinct from complementary PHI, which complements the coverage of publicly insured services by covering all or part of the residual costs not otherwise reimbursed (e.g., copayments). It is worth noting that there is no full consensus as regards this terminology, as some authors use the concepts of SPHI and DPHI indifferently. DPHI should also be distinguished from 'parallel private health insurance' where individuals are covered by one among several parallel insurance systems. These roughly insure for the same health care but an individual is entitled to only one of the insurance systems' benefits. For example, in the US, an individual in need of health care as a consequence of a workplace accident is covered by the Workers' Compensation Board and not by Medicare.

Finally, note that an NHS is a necessary but not sufficient condition to observe the emergence of DPHI. According to a large review of health systems in Organization for Economic Cooperation and Development (OECD) countries, DPHI exists to different extents in the following countries with an NHS (percentages in parentheses indicate the percentage of people enjoying double coverage): Australia (43.5%), Ireland (51.2%), Italy (15.6%), New Zealand (32.8%), Portugal (17.9%), Spain (10.3%), and the UK (11.1%) (Paris *et al.*, 2010). At the same time, double coverage is absent in other NHS-type health systems like Denmark, Norway, and Sweden.

In Section 'Stylized Facts and Preliminary Insights,' some stylized facts that allow a preliminary overview about health systems' performance in the presence of double coverage are presented. Then, the main theoretical concepts that are indispensable to analyze this question are presented in Section 'Theoretical Concerns: Uncertainty and Information.' In the Section 'Empirical Evidence of Uncertainty and Informational Problems: Who Buys Duplicate Private Health Insurance?', the main results from empirical analyses that have tested several aspects of double coverage, in particular, who is more likely to purchase duplicate private insurance and why is displayed. Finally, Section 'Political and Financial Sustainability of a DHPI Health Sector' focuses on the political and financial sustainability of a system with duplicate private insurance.

Stylized Facts and Preliminary Insights

DPHI coverage is usually advocated for at least three reasons:

- It promotes population health.
- It limits public and global health expenditures.
- It increases population choice and health system 'responsiveness,' a term defined below in this section.

Roughly speaking, DPHI emerges because the NHS alone will fail to reach these aims. However, is there really a failure of NHS that justifies the emergence of DPHI? In this section, the preliminary evidence as regards these three objectives, for three NHSs, namely Portugal, Spain, and the UK, is provided. These countries are suitable cases for investigation because their public system has long lived without a significant DPHI – actually, the weight of DPHI on total health expenditures only became relatively significant ($>4\%$) after 2000.

It is of course very difficult to attribute good health outcomes to a health system, because population health depends on many factors. Nevertheless, at first sight, these three public schemes certainly do not do worse than the OECD average. The three indicators commonly used to assess health systems' performance are considered here: infant mortality, potential years of life lost (PYLL), and life expectancy at 65. In the 1960s, before the creation of the Portuguese and Spanish national services, Portugal had almost double that of Spanish infant deaths (43/1000) and four times those of the UK (22/1000) (**Figure 1**). Yet, since the late 1980s, Spain has reached UK levels, well below the OECD average, and Portugal has fallen below the OECD average since 1990. For the last 35 years both the UK and Spain have been constantly below the OECD female average as regards PYLL, whereas Portugal, starting from very high levels, has been approaching the OECD average (**Figure 2**). Finally, as regards life expectancy at 65 years, Spanish women live longer than any other, whereas UK women have a similar life expectancy as the OECD average. Portugal, however, has approached the OECD average in the last 20 years having started from significantly lower levels (**Figure 3**). It seems thus that health concerns might not be the major explanation for the development of DPHI.

Double coverage is commonly defended as a means to restrain total and public health expenditure. Consistently, Portugal, Spain, and the UK have for long been allocating a lower-than-the-average share of their gross domestic product (GDP) to the health sector. However, the pace of growth in these countries has been quite similar to that observed elsewhere. Values have even got further above the OECD average since the mid-1990s in Portugal and in very recent years in Spain and the UK, coinciding with the development of DPHI

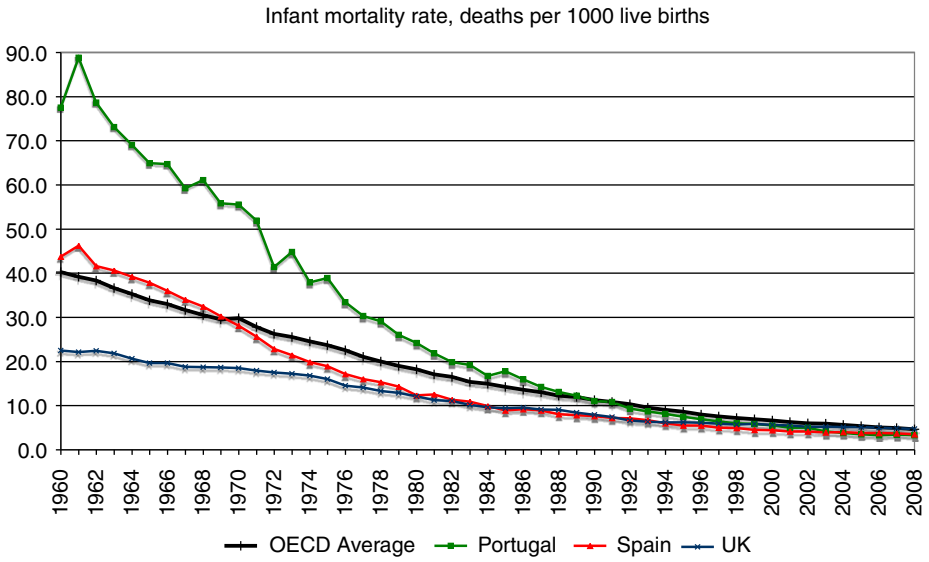


Figure 1 Infant mortality rates. *Source*: OECD Health Data (2011).

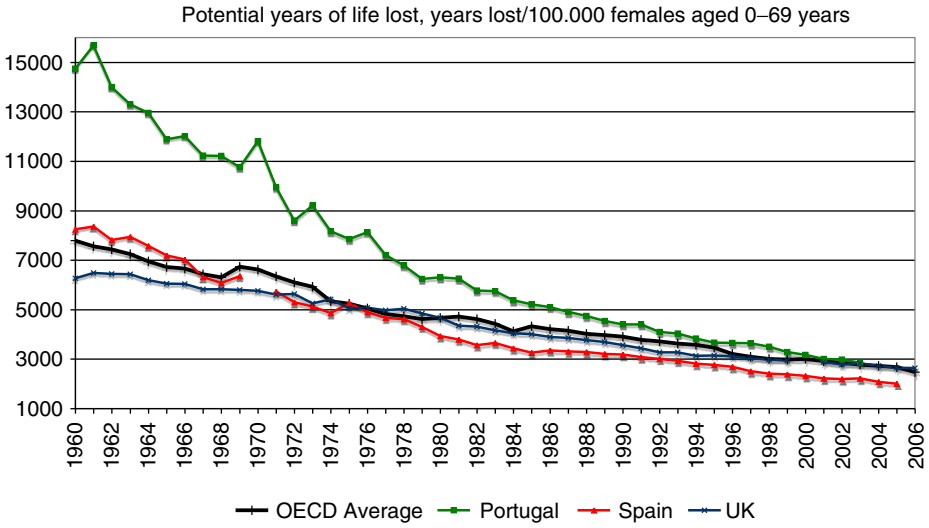


Figure 2 PYLL. *Source*: OECD Health Data (2011).

(Figure 4). Note also that the share of public expenditures is similar in those countries as compared to the average, and has not decreased with the development of DPHI (Figure 5). Instead, since early 2000 this indicator has been constant for Spain and Portugal and has even increased for the UK. Only a deeper analysis would allow one to draw more definitive conclusions. However, at first sight, public health expenditures were not particularly high before DPHI nor has DPHI been very effective in restraining public and general health care expenditures.

Finally, another argument relates to the supposed public service's inability to respond to specific aspects of demand, the so-called lack of responsiveness. The NHS is usually strongly guided by the principle of horizontal equity ('equal treatment for equal needs'); hence it provides comprehensive but needs-based uniform health care, which limits the possibility for patients to express their preferences, even if they are ready to pay for them (this rigidity may sometimes create unexpected and morally conflicting situations, see Box 1 'If you want to choose, go private'). Additionally, principles of rationality and efficiency have prompted these three countries to adopt measures such as gatekeeping, and access to GP and hospital mainly according to the area of residence, which further limit patients' choice. Finally, waiting lists are used to restrain health care use considerably, as a means to ration demand in the absence of significant copayments. In 2010, Portugal had 161 621 patients waiting a median time of 3.3 months for elective surgery (there were however, 248 404 waiting a median of 8.6 months in 2005); in Spain 374 000 patients were waiting an average 1.9 months in 2009; and the UK has

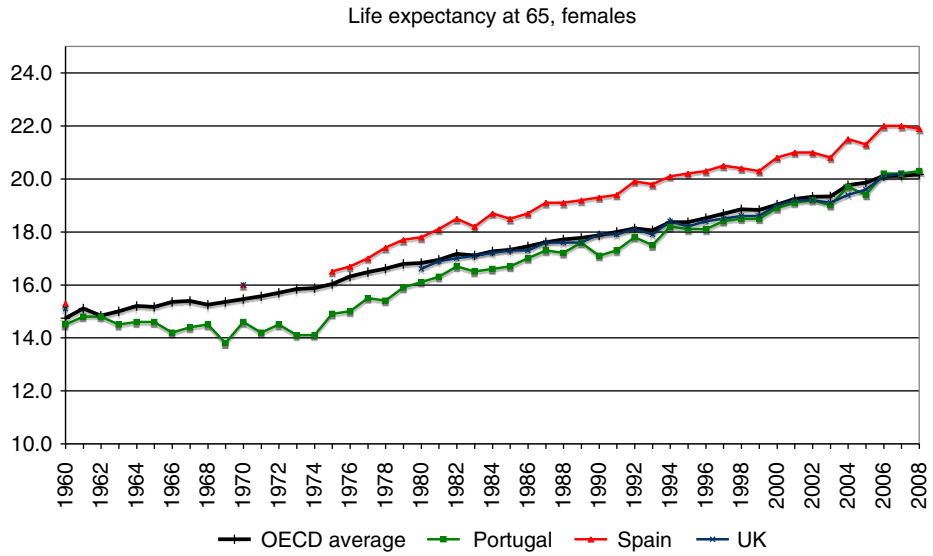


Figure 3 Life expectancy at age 65, females. *Source*: OECD Health Data (2011).

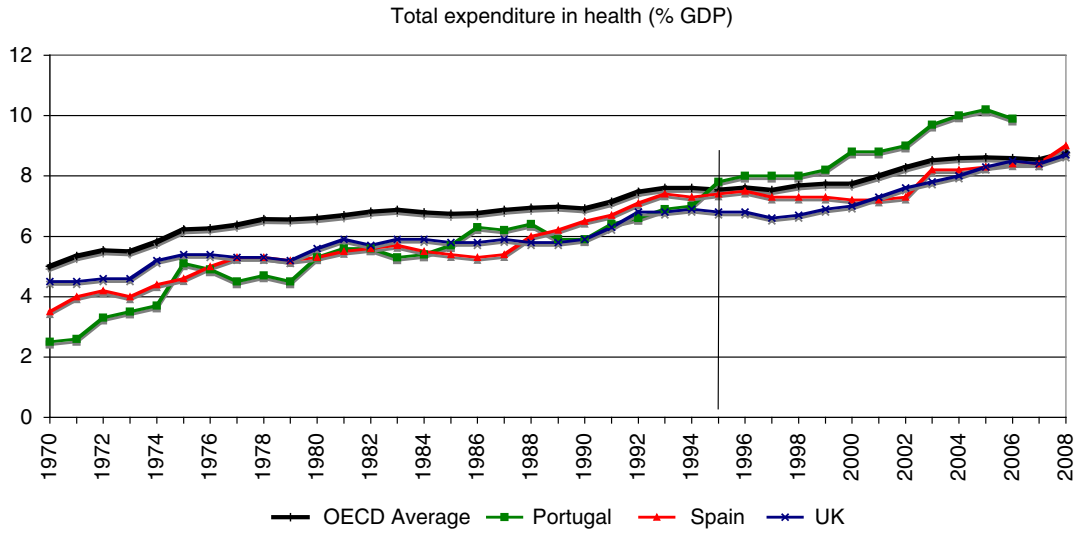


Figure 4 Total expenditure in health as a share of GDP. *Source*: OECD Health Data (2011).

decreased 900 000 patients waiting more than a median of 20 weeks in the 1980s to 620 000 patients waiting a median of less than 5 weeks in 2010. There are thus elements related to the rigidities of NHS-type systems that may favor double coverage. Note that recent decreases in waiting times have been in part obtained through contracts with private practices, so that the potential benefits of DPHI may play some role in these results.

Theoretical Concerns: Uncertainty and Information

To understand why DPHI coverage emerges, who buys it, and with which consequences, it is crucial to understand some economic concepts related to insurance in general and health insurance in particular. To start with, it is important to realize that the health care sector is affected by uncertainty in mainly two dimensions. First, there is unpredictability with respect to an individual health status and future health care needs. Second, there is uncertainty regarding the precise effects of a given health care procedure on a particular patient.

Regarding health status, some individuals are at a higher risk of developing diseases than others. Such a risk is the result of a combination of an individual's genetics, aging, behavior, and environmental context. Neither the individual, nor the physicians, nor the insurers know with exactitude the status of the patient's health. What is more, they may not even share exactly the same information but instead have 'asymmetric information' regarding the individual's health status. Indeed, prior to any screening test or medical intervention, it is common to

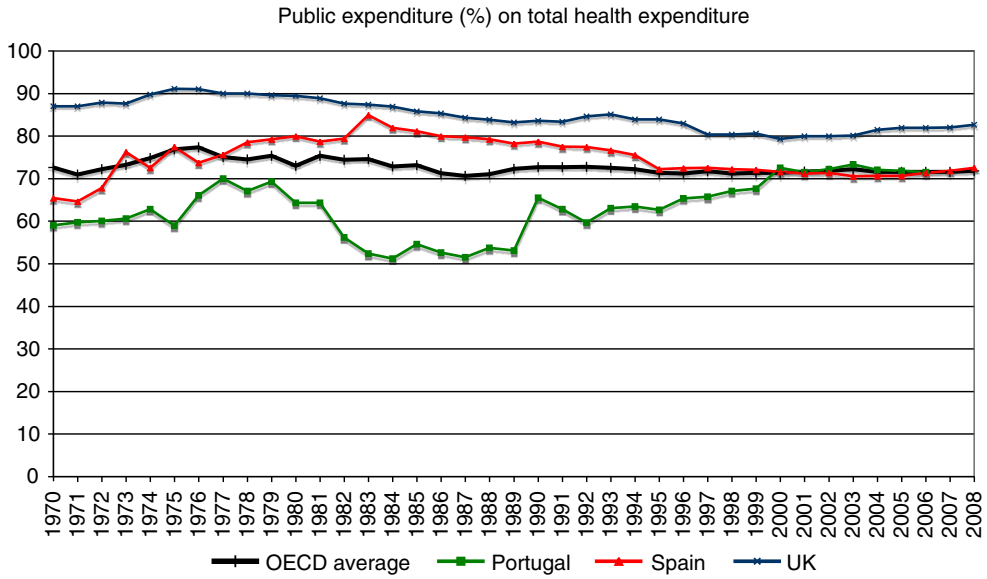


Figure 5 Share of public expenditure on total health expenditures. *Source*: OECD Health Data (2011).

Box 1 If you want to choose, go private

In 2008, Lynda O'Boyle, who suffered from bowel cancer, was not authorized to receive in her NHS hospital in England an expensive drug treatment, uncovered by the NHS, which she was however willing to pay. If she were willing to buy the treatment, she would have to opt out the NHS.

This case posed serious questions about the relationship between public and private systems, and about the trade-off between efficiency and equity. On the one hand, it reveals the rigidity of the NHS and its often criticized lack of responsiveness, emphasizing a major argument in favor of the DPHI: The freedom to choose about one's treatment and life. However, the denial of the drug was grounded on equity concerns, postulating that two equally needed patients treated in two NHS 'adjacent beds' should be treated equally. The use of an expensive drug, although paid by the patient herself, would lead to inequality in treatment, presumably increased by the necessity for doctors to administrate, monitor, and cure the adverse effects of the drug.

This case prompted England to modify drugs' cost-effectiveness threshold and to allow patients to pay privately for treatment without losing their entitlement to the NHS. It reveals the paradox of the DPHI: The equity concerns of the NHS, through the rigidity it creates, sustain the emergence and development of DPHI, which possibly provokes a much higher inequity than that one wanted to avoid.

assume that the individual is more informed about his chances of developing a disease or condition than anyone else. After all, the individual is more aware of one's family's health condition, one's own lifestyle, and one's environment than one's doctor or insurer may be. When facing the doctor, the individual may have all the incentives to disclose information – beyond everything, one wants to be treated – this is most often not the case when facing the insurer – ultimately one wants to be paid for all medical care expenses and would prefer to deny any responsibility of the events. The insurance companies' reaction to this asymmetric information depends very much on the structure of the insurance market but generally issues like 'adverse' and 'propitious selection,' 'moral hazard,' and 'insurance denial' emerge. Below, the Sections Adverse Selection, Risk Selection, Propitious Selection and Moral Hazard discuss the extent of these effects in a duplicate health insurance market.

However, the effects of a particular health care procedure on a patient are not certain. Additionally, physicians are better able to assess the effects of medical care than the patient or insurance companies, and it may not always be of their interest to reveal such information. For example, they may increase the number of consultations aiming at greater profits. Consequently, physician's 'induced demand' may arise. Section 'Supplier-Induced Demand and Dual practice' deals with induced demand in the context of duplicate insurance market.

Adverse Selection

Suppose some individuals in the economy have a high probability of disease and others a low one, known by the individual but unknown to the insurer. Adverse selection (also referred to as "screening" by some authors) may be a mean for insurance companies to force individuals to reveal their risk type. Indeed suppose they offer two types of contracts: One fully insuring individuals at a price reflecting high-risk individuals' probability of becoming sick; another offering incomplete insurance coverage at a price reflecting low-risk individuals' risk. The first contract is too expensive for low-risk individuals and therefore only the high-risk ones would be willing to buy it. The second would offer limited coverage at a price reflecting low-risk individuals' probability. It would therefore not be attractive to high-risk individuals who have a lot to lose for not being completely insured. Still, low-risk individuals would be willing to buy it. Thus, by limiting the coverage of one such contract the insurance company is able to force individuals to self-select by buying the contract intended for their type and hence distinguish low- from high-risk individuals. In the end, the 'bad' type, i.e., the high-risk individuals

end up being fully insured whereas the 'good' type, i.e., the low-risk individual, is prevented from being fully insured.

Note that insurance companies cannot break even by offering an insurance contract to all individuals at an average price that captures the average risk. Actually, low-risk individuals would find such a contract too expensive and therefore only high-risk individuals would end up buying it. Consequently, such a contract would not be financially viable.

The problem of adverse selection is that low-risk individuals are not fully insured by insurance companies even if they would be willing to pay for insurance. In the NHS, adverse selection is not an issue because uniform health care is provided to all resident population, irrespectively of their risk. Also, the effects of adverse selection in the insurance market are lessened when insurance takes the form of group policies. In this case a uniform coverage is offered to all individuals belonging to the group. This is the case of 50% of the duplicate insurance in Portugal, 20% in Spain, and 15% in the UK. Therefore, other things being equal, it would be expected adverse selection as being stronger in the UK and Spain than in Portugal.

It could be argued that adverse selection is not such an important issue in the context of a duplicate insurance market, as opposed to a complementary or supplementary one. In the context of a complementary/supplementary insurance market, insurance covers services not covered in the public system. Therefore, someone without insurance in this market is not insured for those services not covered by the NHS. In contrast, in a duplicate health insurance market, not being insured in the private market would not be so costly because individuals are ensured care in the NHS. Still, as stressed in the Section 'Stylized Facts and Preliminary Insights,' one of the reasons why individuals buy DPHI is because they are deterred by the NHS waiting lists to prompt health care. An NHS with waiting lists is thus offering incomplete health care provision just as a less-than-full coverage insurance contract is. Consequently, all individuals face incomplete health care provision at the NHS and, additionally, low-risk ones face incomplete insurance coverage at the private market due to adverse selection. Therefore, also under DPHI there are individuals never fully insured even if they are willing to pay for duplicate insurance.

Empirical evidence of adverse selection is tricky because its effects may be confounded with others (see the Section Empirical Evidence of Uncertainty and Informational Problems: Who Buys Duplicate Private Health Insurance?). Still, if adverse selection alone would be present in the private market, it would be expected to find empirical evidence that some individuals (high-risk ones) are fully insured at expensive prices and others (low-risk ones) only partially insured at lower prices. To assess adverse selection's unfavorable effects and its relative importance, it is important to identify which health care services are most affected by waiting lists or not provided at all publicly, and to understand which individuals suffer more from adverse selection.

Risk Selection

Another situation that should not be wrongly identified with adverse selection is when insurers select insurees or deny insurance on the basis of individual observable characteristics correlated to risk. For example, PHI is usually denied to individuals 65 years and older because being older is on average associated with higher medical expenses. Still, some of these individuals may be at higher risk of disease than others, which is then not observable. Hence even if insurance would not been denied, adverse selection would arise among the older.

In a duplicate insurance system individuals who are denied insurance coverage are nevertheless entitled to health care at the NHS. They may have to face long waiting times but in principle prompt care is guaranteed for emergencies and urgent situations. Yet, if they wish to buy duplicate insurance they are not able to do it. The situation is however better than in the case in which individuals have to pay the full price of health care if not insured, as it happens when the insurance market is complementary to public provision or in the absence of universal coverage.

Two arguments are commonly used to explain why risk selection exists and who might be denied insurance. First, according to theory, insurers either provide contracts based on true risks or provide different contracts to separate low- and high-risks. However, these are costly procedures which the insurers may not be willing to assume. Second, the empirical literature shows that the variance of health care costs increases with the mean, so that expenditures for high-risk groups are less predictable. Hence, insurers may prefer to provide contracts based on broad categories (age and sex) and then reject high-risk groups directly or indirectly.

In most cases selection is clearly stated in insurance contracts, for example, through an age criterion or through exclusion of benefits from preexisting conditions. Another more subtle form of selection derives from most contracts being short-term (usually 1 year), which enables insurers to modify conditions (or even avoiding renewal, even if this is usually forbidden) once a serious disease has been diagnosed. Contrasting the characteristics of the population receiving health care at the NHS with those relying on DPHI as well can also provide an indication of what constitutes a source of insurance denial. If insurance denial is a fact then there must be empirical evidence that NHS users share some characteristics that DPHI users do not. Still, results should be read with caution because asymmetries may be also due to differences in preferences regarding insurance or other issues.

Propitious Selection

'Propitious selection' in the insurance market occurs when low-risk individuals buy more insurance than high-risk ones. One possible explanation relates to risk aversion, that is, people with a higher risk concern would tend to be more cautious, hence more likely to purchase insurance. As they are more cautious, they also adopt more preventive behavior and are less prone to health hazards. Note that this is not a problem *per se*: It is just a consequence driven from the fact that individuals who have a stronger preference for being insured buy more insurance.

Propitious selection may also arise because high-risk individuals underestimate their risk, prompting them to purchase less insurance. A higher willingness to pay among wealthier persons could also explain propitious selection if wealth and health risk are negatively correlated, as it is usually observed.

The empirical testing of propitious selection is not trivial. On the one hand, we would expect to find evidence that those buying DPHI or higher coverage contracts are less prone to health accidents, and hence use less curative health care. On the other hand, if propitious selection is driven by preventive behavior, those who buy health insurance would use relatively more diagnostic tests and preventive health care. Hence empirical analysis requires reliable information on individuals' health conditions, behavioral, and environmental risks, which are generally difficult to obtain.

Moral Hazard

Moral hazard occurs when an individual facing risk changes one's behavior depending on whether or not one is insured. For example, dental care insurance may lead individuals to be less cautious about their mouth hygiene, which may be reflected in a higher probability of caries (*ex ante* moral hazard). Or, in a case a tooth is removed individuals may decide toward a dental implant only in case they are insured (*ex post* moral hazard).

To induce individuals to exert some effort in the limitation of damages (*ex ante* moral hazard) or to restrain medical care use (*ex post* moral hazard), insurance contracts typically impose the individual part of the incurred cost by making use of deductibles and/or coinsurance rates. This means that the consequence of moral hazard is partial insurance (incomplete coverage), just as in the case of adverse selection.

Because moral hazard consists of a reaction to insurance, it is present under an NHS just as in the private market, for the same level of insurance coverage. Still, the two sectors deal with moral hazard in very different ways. The NHS deals with it by rationing health care, for example, through waiting lists, gatekeeping, and limiting individuals' choices. Actually delayed access to health care is a sort of limited insurance coverage and can thus give incentives to prevention and consequent limitation of damages (*ex ante* moral hazard) or restrain individual use of medical care (*ex post* moral hazard). Similarly, in the private market, individuals are typically not fully insured (due to deductibles and coinsurance rates) and the same mechanism applies. The extent to which each sector is affected by moral hazard depends therefore on the importance of incomplete coverage of each sector.

Curiously, a duplicate insurance system may deal well with moral hazard. Indeed, if on the one hand individuals are twice insured, on the other hand, NHS health care and private insurance are mutually exclusive. In other words, as one owns an insurance policy for a given health event, if a patient goes to the NHS, the insurance coverage is not claimed and vice versa. In principle, the patient faces two sectors with incomplete coverage that deal better or worse with moral hazard. In contrast, complementary insurance destroys any incentives to promote prevention or deter unneeded care that may exist in the public sector because the individual is usually fully insured (because privately the individual is insured for the out-of-pocket payment).

To conclude concerning the presence of moral hazard: it is essential to test whether insurance contracts offering more coverage are associated with greater use of health care. Note that even though very different in their causes adverse selection and moral hazard lead exactly to the same observed market effect: Insurance contracts with less coverage are associated to individuals using less health care. As is discussed in the Section Empirical Evidence of Uncertainty and Informational Problems: Who Buys Duplicate Private Health Insurance?, it is not always easy to distinguish the two effects.

Supplier-Induced Demand and Dual Practice

We now turn to the implications of uncertainty regarding the effect of health care on patients. The physician is obviously the most informed and can use this information for his own benefit by increasing health care acts beyond what is adequate and necessary. Obviously physician behavior depends very much on the incentives faced. As a point of fact, supplier-induced demand (SID) is to be expected in the private market where physicians are usually paid by fee for service. Yet, insurance companies have been trying to redesign physicians' incentives to restrain such practice.

SID should be common to any health care system relying partially or fully in the private market. In this respect, there is no reason to think that a duplicate insurance system is more prone to SID than other systems are. After all, the determinant is the size of the private market and the incentives imposed by insurance companies. Still, in a duplicate insurance system an additional effect comes into action because physicians are often allowed dual practice, i.e., they provide health care both at the NHS and at the private market. There is general awareness that physicians deviate patients toward their private practice where they benefit from additional rents for the health care provided and can induce demand for private benefit. Additionally, SID can easily be transformed in a common and cultural practice because the same physicians act publicly and privately.

Finally, it is important to note that also in an NHS, SID may exist. In an NHS, physicians are usually paid on the basis of a fixed salary, but they may induce health care consumption due to the practice of defensive medicine with a view to avoiding malpractice liability.

It is a challenge to identify empirical evidence of SID. A strategy of identification would be to contrast health care provided across physicians for the same health condition, except that SID can be the norm. For example, in some countries, patients are given another appointment once the results of diagnostic tests are known whereas in others, results and accordingly prescription are given by postal mail or telephone (except for abnormal cases). Also, physicians may be members of a specific culture of medical practice that pushes them toward excess health care. Nonetheless, a duplicate health insurance system allows for the contrast between private and NHS medical practice where differences would be (partially) explained by SID.

Empirical Evidence of Uncertainty and Informational Problems: Who Buys Duplicate Private Health Insurance?

Testing for empirical evidence of uncertainty and informational problems in insurance markets is not easy because

Table 1 Empirical strategies and challenges in predicting uncertainty and informational problems

Analyzed concept	*Empirical strategies and challenges*
High-risk individuals are relatively less insured:	
Risk selection	● Not present in the NHS, probably present at the DPHI. ● 'Strategy': Identify the characteristics of individuals, such as age or declared diseases that prevent them from buying DPHI. ● 'Challenge': Differentiate risk selection from revealed preferences and propitious selection. Some characteristics may be correlated with risk-loving behavior, for example.
Propitious selection	● Not present in the NHS, probably present at the DPHI. ● 'Strategy': First, identify negative correlation between risk and insurance coverage. Second, identify reason of propitious selection: (1) Risk-aversion/preventive; (2) underestimation of risk by high-risk individuals; (3) higher willingness to pay of wealthier (and healthier) individuals. If (1) then should be found that those buying DPHI have less health hazards but tend to use more preventive medical care or adopt less risky health habits. If (2) then should be found more health hazards than *ex ante* predicted relatively more for high-risk individuals then for low-risk ones. If (3) then should be found that wealthier individuals should buy relatively more DPHI and have less health hazards. ● 'Challenge': Proxy for preferences; differentiate propitious selection from risk selection.
High-risk individuals are relatively more insured*:*	
Adverse selection	● Not present in the NHS, probably present at the DPHI. ● 'Strategy': Conclude whether in a DPHI context higher risk individuals buy insurance contracts with higher coverage and relatively more expensive than those bought by low-risk individuals. ● 'Challenge': Find a good proxy for risk; differentiate adverse selection from moral hazard.
Moral hazard	● Probably present both at the NHS and DPHI resulting in overconsumption of health care. ● 'Strategy': Identify change in preventive and curative health care use due to insurance. ● 'Challenge': Differentiate from adverse selection.
Supplier-induced demand	● Probably present at the NHS (defensive medicine) and DPHI (rent seeking) resulting in overconsumption of health care. ● 'Strategy': Identify different medical practices across competition contexts. ● 'Challenge': Differentiate SID from cultural medical practice.

several forces can be confounded. Yet, it is important to precisely identify which problems are present because each raises different equity and efficiency concerns and leads to different policy recommendations.

Table 1 summarizes the empirical prediction of each of the effects discussed in the Section Theoretical Concerns: Uncertainty and Information and helps in following the upcoming discussion. One should start to understand whether insurance coverage depends on the type of risk and then follow by inferring which mechanism is at the foundation of such outcome. If it is observed that high-risk individuals buy less DPHI than low-risk individuals it can be due to either risk selection or propitious selection. The empirical challenge consists precisely in identifying which of the two effects is in place because although risk selection and some causes of propitious selection may call for government intervention, that is less the case if propitious selection is due to more risk-averse individuals tending to buy more insurance – after all, individuals just act according to their own preferences.

If however, it is observed that high-risk individuals are relatively more insured, adverse selection may be at play and instead, it is low-risk individuals who are denied insurance. Still the measure of risk may be inconclusive. Indeed, in practice, it may just be observed that higher coverage insurance contracts are associated with more health care expenditures. This can be due to not only adverse selection but also due to moral hazard. In other terms, people with private insurance have higher expenditures either because high-risk individuals are more likely to purchase high-coverage contracts (adverse selection) or because people with higher coverage have lower incentives to parsimonious health care use (moral hazard).

A somewhat orthogonal problem to the risk issue is to what extent health care use is induced beyond what is adequate because physicians target higher profits. If this is the case there is SID. Incentives to induce demand may be related, for example, to generous fee for service reimbursement schemes under DPHI, or to very low copayments that ease the inducement process. Hence evidence on moral hazard may indeed be overestimated if the inducement of demand effect is not controlled for. The empirical strategy consists in identifying different medical practices across competition contexts but it is obviously a challenge to distinguish SID from cultural medical practice. Additionally, different medical practices may as well be explained by distinctive regional administrations or governances.

Although disentangling empirically the informational problems is challenging, some research has led to interesting results in the DPHI context. Confirming the seminal results of the Rand experiment, some UK studies have consistently confirmed a decrease in drug consumption which follows an increase in the copayment supporting the evidence of *ex post* moral hazard. Olivella and Vera-Hernández (2013) use data of the British Household Panel Survey for the period 1996–2007 to test empirical evidence of asymmetric information and distinguish the different effects at play. They contrast health care use of individuals having bought PHI with that of those who obtained PHI from their employer as a fringe benefit, using three measures of health care use: (1) hospitalization in a fully privately funded hospital, (2) hospitalization in publicly funded hospital, and (3) GP visits. Their reasoning is as follows. People with individual PHI are those who explicitly decided to buy it, and are called

'deciders'; the other group includes people who obtained PHI form their employer, i.e., they did not decide to buy it and are called the 'nondeciders.' Both deciders' and nondeciders' insurance contracts are equally affected by moral hazard and thus differences across the two groups cannot be due to moral hazard. Instead, if deciders use more health care services than the nondeciders, then adverse selection prevails (deciders use more care because they are in worse health, hence high risks are more likely to buy PHI). In contrast, if deciders use less health care services, propitious selection or risk selection prevails (deciders use less care because they enjoy a better health, hence low risks are more likely to buy PHI). The authors find that individuals having decided to buy a PHI use more health care irrespectively of the measure used, concluding on the existence of adverse selection.

In contrast with this latter finding, Doiron *et al.* (2008) suggest evidence of propitious selection in DPHI, using Australian data. They find that healthier individuals purchase relatively more DPHI. In this case, the difficulty is then to distinguish this effect from risk selection by private insurers, which would lead to a similar result. To do so, they observe that people engaging in risk-taking behaviors (smoking, drinking, and lack of exercise) demand less private insurance coverage. Thus the authors put forward that relatively more risk-averse individuals are more likely to buy DPHI. Additionally, the assumption that insurers deny insurance to high-risk individuals seems partly discarded by the higher insurance coverage among people with long-term conditions. These two studies, in different contexts, thus show opposite results. If an earlier literature is considerd, findings are more sustaining that DPHI is associated to a healthier condition, although these studies do not try to explain the correlation.

The empirical evidence of SID has long been and remains a subject of controversy among health economists. Recent natural experiments however show that substantial variations in copayments produce effects that are similar to those observed in the Rand experiment, whose design made the occurrence of SID very unlikely. Hence observed *ex post* moral hazard is certainly a more plausible explanation than SID for higher health care use under DPHI.

Other empirical issues related to DPHI deserve also to be mentioned, even if less related to the theoretical problems presented in the Section Theoretical Concerns: Uncertainty and Information. To begin with, the decision to buy DPHI cannot be analyzed without considering what happens in the NHS. The demand for private insurance depends on the perceived quality in both public and private sectors. One of the most popular indicators of NHS quality is waiting lists and waiting times, which are easy to obtain and to which people are usually highly sensitive. There is evidence that long waiting lists, expected waiting times, or more generally the perceived quality gap between the NHS and private provision are determinants for people to insure privately. These findings confirm empirically that responsiveness is a relevant factor for the emergence of DPHI.

Finally, most studies confirm the strong relationship between private insurance and high socioeconomic status, in particular, income and education. Supplemental or DPHI is without doubt a normal good, which is more purchased by richer people. Income is hence obviously one of the main determinants of the demand for PHI. This last finding poses crucial questions from a welfare viewpoint because then DPHI may contribute to inequity in health. If DPHI only allows for luxurious services unrelated to quality of clinical procedures (better amenities, faster care, etc.), this may not be such a relevant problem. However, if double coverage allows access to better care that unsatisfactory NHS cannot offer, this is a serious social concern. The higher use of physician services under double coverage would sustain the latter assumption. This conclusion would be reinforced if private insurers, through higher financial capacities, are able to attract better doctors under a dual practice regime. Unfortunately, to our best knowledge, no study has assessed the impact of DPHI on quality of care (probably because quality is quite difficult to measure). The higher health care use under DPHI also questions the efficiency of duplication in a context of scarce resources.

Political and Financial Sustainability of a DHPI Health Sector

So far, we have examined individual decisions with regard to purchasing private insurance and consumption of health care services in a context of double coverage. In this section, the potential impact of double coverage is discussed at an aggregate level, at the health system, or country level. What could explain the consensus in favor of double coverage? Does the existence of DPHI threaten the political sustainability of the NHS? What is the impact of double coverage on health care expenditures? Does it impose a higher financial burden on the NHS, or does it alleviate this burden?

First, political sustainability is considered. Suppose, as confirmed in the empirical literature, that richer people are more likely to purchase private insurance. These people will thus be paying higher taxes (assuming nonregressive taxation, as it is usually the case) without enjoying one of its major benefits, namely public health care provision through the NHS. Hence they may vote against the existence of an NHS, or at least against paying high taxes to finance it. As low-income people may also want to avoid large contributions and thus prefer lower level of health care provision, in the end the support for high public provision will decrease. However, three factors are likely to modify this finding of public under-provision:

- Opinion polls show the existence of a health-specific altruism and concern for equity in health, hence even richer people may support public provision of health care to the poor.
- Poor people are in most countries exempt from taxes and copayments, so that they favor a higher level of publicly-provided care.
- There is no perfect correlation between wealth and health, hence rich people experiencing poor health may also favor public care because its price will be lower than that of private insurance, even in a context of progressive taxation.

To conclude, the majority will vote for a lower level of public provision but would not choose zero public care. In a nutshell, this theory justifies the preference for a system with double coverage, although with a lower public provision

than the one in the absence of a private sector. To our best knowledge, political questions around double coverage have never received empirical validation. The only evidence so far is that people tend to favor increased public spending after it has decreased and the reverse after it has increased (Tuohy *et al.*, 2004). This result may emphasize the consensus for a target level of public expenditures compensated by private ones.

As regards financial consequences and economic sustainability, it is often put forward that DPHI may alleviate the burden on the public sector through providing care to a share of the population. This was one of the major motivations for the governments that favored the emergence of DPHI. However, the impact on total health expenditures depends on many factors. First, it depends on whether private providers are able to offer care at a lower cost than the one they would have experienced in the public sector, which is not that clear. Additionally, private insurers generally reimburse physicians through fee for service whereas salary payments have traditionally characterized NHS-type systems. Therefore physicians are more likely to tend to induce demand in the private sector. Second, double coverage is usually accompanied by dual practice, that is, physicians combining public and private practice, whose effects on health care expenditures are difficult to assess. However, dual practice enables doctors to earn additional revenue in the private sector allowing public institutions to pay lower wages, whereas attracting good doctors. Physicians in the public sector may also provide better care to build a reputation for their private activity, but perhaps also overtreat or induce demand. Physicians may also divert resources and patients from the public sector to their private practices. They may also 'import' more resource-consuming practice style of the private sector to their public activity. Finally, it is to be noted that evidence suggests a higher health care use for people with double coverage, in Portugal, Spain, Italy, Ireland, and the UK. In this regard, see **Box 2** – Opting-out as a solution for the duplicate insurance problems?

Box 2 Opting-out as a solution for the duplicate insurance problems?

An often argued solution for the adverse effects of DPHI is the opting-out system. Under opting out, patients buying private insurance are not entitled to NHS care or have to pay the full price for using it.

In Portugal, the first experience of opting out concerned the employees of Portugal Telecom (PT), the largest telecommunication company of the country. PT employees and retirees were enrolled in the PT-ACS (PT-Health Care Services Association) health insurance scheme. In 1998, the State paid PT-ACS a per capita value for its ensured individuals and PT-ACS became fully responsible for their health coverage. For services not provided by private facilities, patients were still entitled to use NHS health care but PT-ACS had to pay its full price.

This agreement however came to an end in 2008 and opting out failed as a solution to alleviate the NHS burden. Indeed the per capita value previously agreed came to be insufficient to cover health care expenditures. In particular, PT-ACS faced growing expenditures due to an increasingly larger share of retirees. Also, private facilities in Portugal did not provide care for a complete range of services so that PT-ACS insured individuals had very often to resort to the NHS.

In a nutshell, public–private interactions pass through a series of complex mechanisms whose final consequences are difficult to assess. A correlation between public and private health care expenditures is generally observed, but it can hardly be concluded that the former is driven by the latter given that both are influenced by the same determinants. What is granted is that health care expenditures in countries with double coverage have increased at a path that is common to most OECD countries, and its efficiency in achieving good population health is comparable too. Some studies report however that an increase in the private share of total health care expenditures is associated with a subsequent decline in public health spending as a proportion of total public expenditure. This would tend to sustain the hypothesis of the private sector alleviating the burden of the public sector in detriment of other theoretical assumptions. Yet studies consistently show that double coverage is associated with a higher use of health care services.

See also: Aging: Health at Advanced Ages. Alcohol. Economic Evaluation, Uncertainty in. Education and Health. Illegal Drug Use, Health Effects of. Intergenerational Effects on Health – *In Utero* and Early Life. Macroeconomy and Health. Markets in Health Care. Medical Malpractice, Defensive Medicine, and Physician Supply. Moral Hazard. Nutrition, Economics of. Peer Effects in Health Behaviors. Physician-Induced Demand. Risk Selection and Risk Adjustment. Sex Work and Risky Sex in Developing Countries. Smoking, Economics of. Supplementary Private Health Insurance in National Health Insurance Systems. Supplementary Private Insurance in National Systems and the USA. Waiting Times

References

Doiron, D., Jones, G. and Savage, E. (2008). Healthy, wealthy and insured? The role of self-assessed health in the demand for private health insurance. *Health Economics* **17**(3), 317–334.

Olivella, P. and Vera-Hernández, M. (2013). Testing for asymmetric information in private health insurance. *The Economic Journal* **123**, 96–130.

Paris, V., Devaux, M. and Wei, L. (2010). Health systems institutional characteristics: A survey of 29 countries. OECD Health Working Papers no. 50, OECD publishing.

Tuohy, C. H., Flood, C. M. and Stabile, M. (2004). How does private finance affect public health care systems? Marshaling the evidence from OECD nations. *Journal of Health Politics, Policy and Law* **29**(3), 359–396.

Further Reading

On Health Data

Barros, P. P. and de Almeida Simões, J. (2007). Portugal: Health system review. *Health Systems in Transition* **9**(5), 1–140.

Boyle, S. (2011). United Kingdom (England): Health system review. *Health Systems in Transition* **13**(1), 1–486.

García-Armesto, S., Abadía-Taira, M. B., Dúran, A., Hernández-Quevedo, C. and Bernal-Delgado, E. (2010). Spain: Health system review. *Health Systems in Transition* **12**(4), 1–295.

OECD (2010). OECD Health Data 2010 – Version: October 2010. Available at: http://www.oecd.org/document/30/0,3746,en_2649_37407_12968734_1_1_1_37407,00.html (accessed 21.07.11).

On Uncertainty and Information

Barros, P. P., Machado, M. and Sanz de Galdeano, A. (2008). Moral hazard and the demand for health services: A matching estimator approach. *Journal of Health Economics* **27**(4), 1006–1025.

Besley, T., Hall, J. and Preston, I. (1999). The demand for private health insurance: Do waiting lists matter? *Journal of Public Economics* **72**(2), 155–181.

Chiappori, P. -A. and Salanié, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy* **108**, 56–78.

Cullis, J. G., Jones, P. R. and Propper, C. (2000). *Waiting lists and medical treatment. Handbook of health economics.* Ch. 23. The Netherlands: Elsevier.

Jones, A. M., Koolman, X. and Van Doorslaer, E. (2006). The impact of having supplementary private health insurance on the uses of specialists. *Annals of Economics and Statistics* **83/84**, 251–275.

Intergenerational Effects on Health – *In Utero* and Early Life

H Royer, University of California-Santa Barbara, Santa Barbara, CA, USA, and National Bureau of Economic Research, Cambridge, MA, USA
A Witman, University of California-Santa Barbara, Santa Barbara, CA, USA

Glossary

Appearance, pulse, grimace, activity, respiration (APGAR) score A measure of infant health using a five component assessment completed immediately after birth.
Cross-sectional data Data observed for many subjects at one particular point in time.
Dynamic complementarity The notion that skills produced in one period increase the productivity of investment in later periods.
Ectopic pregnancy Implantation of the embryo outside of the uterus, usually resulting in a nonviable pregnancy and great health risk to the mother.
Elasticity of substitution A summary of how well one factor can be substituted for another in production.
Fixed effect model A type of statistical model intended to deal with omitted variables bias by using a dataset collected on multiple entities (e.g., individuals) over multiple time periods. To isolate the effect of some factor(s), the statistical model uses variation within entities over time.
Omitted variables bias Over- or underestimation of the impact of one variable on another that is caused by leaving out one or more important variables from the estimated model.
PROGRESA Income transfer program for the poor in Mexico.
Randomized-controlled trial A scientific experiment conducted to test the effect of an intervention by randomly assigning participants to a treatment and control group. Differences between the treatment and control group participants are interpreted as the causal effect of the intervention.
Regression discontinuity design A statistical technique used to address issues of omitted variables bias where assignment to treatment changes discretely based on some characteristic or set of characteristics. To estimate the effect of the treatment, groups of treated and untreated individuals with similar characteristics are compared. For example, college scholarships are frequently based on grade point averages. To estimate the effect of a scholarship, outcomes for individuals who had grade point averages that made them just eligible for the scholarship are then contrasted with the outcomes for individuals who had grade point averages that made them just ineligible for the scholarship.
Self-productivity The notion that skills acquired in one period persist into other periods and that higher levels of skills in one period can increase the amount of skills acquired in later periods.
Quasi-experiment or natural experiment A study where some entities (e.g., individuals) are exposed to an intervention outside of the researcher's control.

Introduction

Today an understanding of health is not complete without considering the role of *in utero* and intergenerational influences. The recent popularity of the fetal origins hypothesis, asserting that early life influences through the fetal environment (e.g., nutritional deprivation) have latent long-run effects on health, has nudged economists to think about the production of health beginning much earlier in life. This hypothesis, complemented with economic theories of parental investment characterizes how important early-life factors can be in explaining adult health.

The goal of this article is to give the reader a framework and an understanding of the strength of *in utero* and intergenerational influences. To provide a concrete structure to interpret these effects, the authors first outline a multiperiod investment model akin to the work of James Heckman that translates early life circumstances to health in adulthood. With such a mathematical model, one can be very specific about the roles of the *in utero* and intergenerational environments. Before discussing the various inputs into the health production function, the authors consider possible measures of health *in utero* and at birth. The bulk of the article then discusses how various inputs (e.g., maternal nutrition, sickness, maternal age, maternal education, family income, employment, maternal health behaviors, and environmental exposure) impact health using the Heckman investment model as a guide.

Theoretical Framework

Economic Framework for the Link between *In Utero* and Early Life Conditions and Later Health

Under traditional economic models of health, there is little room for early life and *in utero* events to impact later health. More recently developed models, however, demonstrate the importance of investment and events in early life for health production. Models such as Heckman (2007) provide the mathematical structure to understand how early parental

investment and initial endowments (e.g., health at birth) may affect adult health. These models invoke two important features: self-productivity and dynamic complementarity. Self-productivity embodies the notion that skills acquired in one period persist into other periods, and that higher levels of skills in one period can increase the amount of skills acquired in a later period. Dynamic complementarities embrace the idea that capabilities acquired in one period augment the productivity of investment in the future.

Consider a simplified model with two periods of childhood investment and a constant elasticity of substitution production function as Heckman (2007) does:

$$\boldsymbol{h} = f(b, \theta, [\gamma I_1^{\varphi} + (1-\gamma) I_2^{\varphi}]^{\frac{1}{\varphi}}) \qquad [1]$$

where $\boldsymbol{h}$ is a vector of adult capabilities in period 3 (adulthood), b are parental capabilities, θ is the initial endowment, I_t is investment in period t, $\frac{1}{1-\varphi}$ is the elasticity of substitution of inputs across periods, and γ represents the net effect of I_1 on $\boldsymbol{h}$ (the 'capability multiplier'). One can think of $\boldsymbol{h}$ as including health and other capabilities such as education in adulthood. Investments can include nutrition and medical care in childhood. Parental characteristics such as income and education influence the choice of inputs (i.e., I_1 and I_2), either by shifting tastes, acting as a constraint on the ability to purchase inputs, or changing θ.

Self-productivity implies that $\frac{\partial h}{\partial I_t} > 0$ for $t = 1,2$; that is, investment made in prior periods raises adult capabilities. Vaccinations would be one such example; the polio vaccine taken as a child nearly assures that as adult, an individual will not impeded by polio. Under dynamic complementarities, the function $\frac{\partial^2 h}{\partial \theta \partial I_t} > 0$, meaning that the effect of investment is an increasing function of capabilities. As an example of dynamic complementarities, consider an early childhood investment made in period 1 (e.g., Head Start). This investment may augment childhood capabilities in period 2, which then will make formal schooling following Head Start more productive.

Under some simplifying assumptions, this general model can generate some useful insights about the possible role of early and *in utero* investments. First, the larger the capability multiplier (i.e., γ) the higher the optimal ratio of early to late investment. Second, if early and late investments are perfect substitutes, disadvantage in period 1 can always be overcome with later investment. As the degree of substitution approaches ∞, optimal investment in period 1 is equal to optimal investment in period 2. Third, there is a tradeoff between investing in period 1 and investing in period 2. Owing to discounting, investment in period 2 is cheaper than investment of the same amount in period 1. This consideration pushes investment to period 2, but the productivity of investment in period 1 (i.e., the size of γ) encourages investment in period 1.

This model can explain why early life investments, even if they are small in magnitude, can have effects on more long-run outcomes. Moreover, although this article discusses the importance of early and *in utero* conditions collectively, it may be important to distinguish between these even further. Specifically, *in utero* investments, because of the extended period allowed for dynamic complementarities, may be more important than early childhood investments.

Fetal Origins Hypothesis – An Epidemiological Explanation for the Possible Connection between *In Utero* Conditions and Later Outcomes

The biological foundation for linking *in utero* conditions to later life outcomes is the fetal origins hypothesis. This hypothesis, championed by British physician David Barker, asserts that nutrient deprivation at the beginning of life can raise adult chronic disease risk. Looking across areas in England, Barker noted that infant mortality rates were correlated with later mortality rates of the same cohorts. The biological underpinnings of the fetal origins hypothesis suggest that nutrition during pregnancy affects fetal development. If a fetus is deprived of nutrients *in utero*, available nutrients are diverted for neurological development while the development of nonneurological systems are sacrificed. This tradeoff manifests itself later in life in the form of higher hypertension risk and increased insulin sensitivity.

Although the hypothesis has gained some acceptance, it is still highly disputed – partially because solid empirical support is difficult to come by. For one, the data demands of testing such a hypothesis require data on both early life conditions and later outcomes. This is an arduous demand given that the collection of high quality data in many countries is only a recent phenomenon. Probably the most foreboding critique of this hypothesis is that it was originally based on observational data and thus, is susceptible to typical omitted variables bias issues. However, it should be noted that animal studies (excluding humans) where the fetal environment is more easily manipulable generally show strong support of the fetal origins hypothesis.

The toolkit of economists is well-suited to addressing these two shortcomings of the public health and medical literature on this topic. Economists have used clever quasi-experimental strategies (many of which will be discussed later in this article) to identify causal relationships between early life conditions and later outcomes. A subset of these natural experiments include the 1918 and 1957 flu epidemics, maternal fasting during Ramadan, variation in malaria prevalence either due to seasonal variation or eradication campaigns, and the implementation of the federal Food Stamps program.

The application of the fetal origins hypothesis in the economics literature is broad. For instance, it is the most common explanation for the association between birth weight, a measure of *in utero* nutrition, and educational attainment, adult economic outcomes, and adult health outcomes. Some might argue that this is an incorrect interpretation of the fetal origins hypothesis because the hypothesis is specifically about how *in utero* circumstances have a latent impact which is only expressed in late adulthood, not in early adulthood.

It should be noted that although the fetal origins hypothesis provides a biological basis for the relationship between the *in utero* environment and subsequent outcomes, estimates of the relationship between early life circumstances and later outcomes will combine both the biological effect and the effects of any ensuing investment decisions. Several studies have been interested in whether investment responses are compensatory or reinforcing, but due to the difficulty measuring intermediate inputs the literature has not reached a consensus regarding which type of investment behavior is more predominant.

Measuring *In Utero* Health and Later Health

Of critical importance in the *in utero* and early life health literature is the measurement of health. Measurement of *in utero* health without intervention is nearly impossible. However, via blood samples, measurements of the maternal environment can be made (e.g., cortisol levels indicating stress). But such data are not part of standard datasets commonly used by economists.

As an alternative measure of *in utero* health, researchers frequently use measures of health at the time of birth. These include birth weight, appearance, pulse, grimace, activity, respiration (APGAR) score, length of gestation, and infant mortality. Most of these measures are likely a reflection of the effects of the *in utero* environment rather than the circumstances after birth. Although shifts in many of these outcomes (e.g., birth weight) may not be so meaningful, economists frequently are interested in the tails of the distribution of these outcomes. Low birth weight (<2500 g), very low birth weight (<1500 g), and premature birth (<37 weeks of gestation) are focal outcomes.

In the past 10 years, health economists have debated whether birth weight is an adequate measure of *in utero* health. Although the measurement of birth weight is easy, by itself, birth weight is not necessarily reflective of any health issues. Historically, interest in this measure by researchers is mainly predicated on the strong birth weight and infant mortality correlation. But such correlation does not imply causation. As an innovative approach to control for possibly confounding factors, researchers have compared birth weight differences between twins and have related those differences to within-twin-pair differences in infant mortality. A weaker birth weight and infant mortality relationship emerges from this approach. Nevertheless, the importance of birth weight as a leading health indicator has been reaffirmed with the many recent studies mapping a connection between birth weight and longer run outcomes such as educational attainment, wages, and rates of disability as adults.

Measuring early childhood health is equally difficult as measuring *in utero* health. Easily obtained health measures such as childhood mortality are rare, making it challenging to find effects of interventions on mortality. The most common chronic conditions in childhood are asthma, hay fever, and bronchitis, but they inflict less than 15% of children in a particular year. Aggregating these conditions to derive a single index measure of health is challenging because it is unclear how to combine these outcomes sensibly. For example, an outcome of the number of chronic conditions a child has would give equal weight to epilepsy as it does to bronchitis.

The Intergenerational Transmission of Health

The model outlined by eqn [1] allows for an intergenerational transmission of health via several different mechanisms. First, parental attributes can affect a child's health directly through changes in b, parental capabilities. Second, intergenerational relationships can arise because of genetics, θ in the model. Third, parental capabilities will likely affect investments represented by I_1 and I_2. Distinguishing between these three types of mechanisms is not possible empirically.

Arguably the best measures of the intergenerational correlation in health are those relating to birth weight. The correlation in birth weight across generations is typically smaller in the USA than the intergenerational correlation in wages. In a study using matched children–mother data from California, the likelihood that children were low birth weight increased by 50% if their mothers were low birth weight. These intergenerational relationships are slightly stronger among low socioeconomic status (SES) mothers.

Data on sibling mothers can help to understand how much of the intergenerational transmission in birth weight is genetic versus behavioral. Traditionally this is done by assuming a data-generating process where a child's birth weight is assumed to be an additively separable linear function of mother's birth weight and a mother's family fixed effect. The fixed effect is intended to capture genetic factors that mothers who are siblings share in common, but it also captures anything else the sibling mothers share. This assumed relationship is rather restrictive as it does not allow for a gene and environment interaction. Interestingly, based on nontwin mother sibling comparisons, family background characteristics do not explain the intergenerational correlation in birth weight. But some argue that these siblings are not nearly enough alike. Thus, other studies focus on twin sibling comparisons. Unlike in the case of sibling mothers, some of the intergenerational birth weight relation is explained by family background. The effect of mother's birth weight on child's birth weight in models that control for time-invariant features of the mother's family is approximately half the size of that from models that do not, suggesting a strong possible role for genetics.

This article continues by investigating maternal factors (e.g., income, nutrition) and other influences (e.g., environment, health care) that may explain these intergenerational correlations in health.

Factors Affecting *In Utero* and Later Health

Maternal Sickness and Stress

A natural empirical test of the fetal origins hypothesis (or the effect of the *in utero* environment more generally) is to examine influences on the maternal environment during pregnancy. These influences include maternal sickness, maternal stress, and maternal nutrition. The authors reserve discussion of maternal nutrition until later as the literature is more expansive on that topic. In general, it is difficult to isolate the pure effect of these factors because it is nearly impossible to conceive of a quasi-experiment that only manipulates sickness or stress. For example, terrorist attacks such as 11 September have been used to understand the effect of maternal stress, but one might imagine that these attacks could also have economic effects.

Of the maternal influences, maternal sickness is considered to be one of the most important. The 1918 flu epidemic provided a unique opportunity to examine the effect of prenatal flu exposure on long-run outcomes. This flu spread

rapidly and suddenly; 25 million people in the USA contracted the virus. Cohorts *in utero* at the time of the flu exhibited diminished health and economic outcomes as adults (i.e., higher disability rates, lower education attainment, and reduced wages). For the more recent Asian flu pandemic of 1957, it is possible to follow the effects of the flu across the lifecycle. In particular, unlike for the 1918 flu, one can test whether flu exposure is related to reduced birth weight, one of the underpinnings of the fetal origins hypothesis. Overall, the flu does not impact birth outcomes. However, these effects are quite heterogenous. The children born to smoking mothers or shorter mothers exhibit lower birth weights as a result of the flu. Effects on cognitive outcomes are present overall, not confined to a particular subgroup. Exposure to malaria *in utero* and during early childhood also has important consequences for long-run outcomes. Although today malaria is an issue in developing countries, in the early 20th century rates of malaria in the American South were comparable to those in developing world today. Exposed cohorts have lower educational attainment and higher rates of poverty.

Relative to maternal sickness, understanding the effect of maternal stress is more challenging. Measurement of maternal stress is typically indirect because measurement of stress is difficult. As a result, studies of the maternal stress often focus on events that are presumed to affect maternal stress. Terrorist attacks such as 11 September and armed conflict in Israel are two such examples. For these events, because they are more recent, evidence on the long-run impacts is limited. However, the stress-provoking events have substantial short-run effects on the incidence of low birth weight and prematurity. As an alternative to this case study approach, some research has measured maternal stress through cortisol levels directly. Sibling comparisons – effectively comparing maternal cortisol levels across births to the same mother and relating these within-family differences to differences in long-run outcomes are used. These cortisol differences have consequences for cognitive, educational, and health outcomes.

Overall, this literature evaluates the effect of negative shocks to the maternal environment. As such, these research findings may be less interesting for policymakers who are interested in deciding which policies are best to improve the fetal environment. Indeed more research is needed on positive shocks.

Maternal Characteristics

Maternal attributes such as education and age can impact early life health either directly or indirectly through the choice of familial inputs or endowments. For example, a mother's education may affect her knowledge regarding the health impacts of maternal smoking. However, in the presence of assortative mating, her education may influence the education of the mate she chooses.

There is a recent growing interest in the impact of maternal education within economics. This is in part due to an expanding focus on the nonwage effects of human capital. Moreover, maternal education is one of the strongest predictors of infant health. Based on USA data, an extra year of schooling reduces the rate of low birth weight by 10%. These effects are surprisingly linear, implying that the effect of a year of high school education is roughly equal to the effect of a year of college education.

Of course, these correlations do not necessarily imply that there is a causal relationship between maternal education and infant health. Omitted variables bias is a concern, particularly because maternal education is positively related to other attributes such as family background that might improve infant health.

The recent economics literature has made great strides in identifying the causal effect of maternal education. Two of the more frequently exploited quasi-experiments are the construction of new schools and the expansion of compulsory schooling. In the USA, the expansion of higher education through the building of new universities and colleges between 1940 and 1990 led to reductions in the rates of prematurity and low birth weight. Outside of the USA, the construction of new schools in areas without schools has resulted in similar improvements in infant health. When interpreting these estimates, however, one should think about these two settings as possibly identifying different effects of education in the case that there are nonlinear effects of maternal education.

Compulsory schooling reforms in the twentieth century led cohorts born close to one another to have different educational requirements. These compulsory schooling laws dictate when individuals can legally drop out of school. In countries where many individuals drop out at the minimum schooling age and the compulsory schooling laws are enforced, increases in the compulsory schooling age are useful instruments for maternal education. In the USA, the size of the population affected by compulsory schooling reforms is rather small. In contrast, in Britain, at least historically, most individuals drop out of school at the minimum schooling age. Thus, one can use regression discontinuity techniques where contrasts are made between individuals proximate in date of birth who might be otherwise identical except for their level of schooling. The British compulsory reforms generally point to no effects of maternal education on infant health.

The discussed quasi-experiments increase education by extending the end of schooling. Alternatively, an increase in educational attainment could be achieved by reducing the age at school entry. Increases in schooling via augmenting either the beginning or end of schooling could potentially estimate different effects of education. As for the latter, there could be a mechanical effect of extra schooling. Being in school longer may act as an incarceration effect, reducing rates of sexual activity and thus, result in delayed fertility.

This conceptual difference may be an explanation for the difference between the conclusions reached from using school entry policies and other studies. School entry policies impact the start of schooling. Despite their differences in acquired schooling, comparisons of individuals born before and after school entry dates (i.e., the date by which a child must have reached age 5 to enter school) show no evidence of effects of maternal education on infant health.

One difficulty often neglected in this literature is that an instrument for maternal education may affect both fertility and infant health. In the case that there are fertility effects of education, the measured effect of maternal education on infant health suffers from a selection problem.

Similar to that of maternal education, the effects of maternal age could be direct or indirect. Women at either end of the childbearing age spectrum experience worse infant health outcomes. Support of the biological effects of maternal age has been confirmed with animal studies, but maternal age may also influence the choice of prenatal and postnatal inputs. Women who give birth at earlier ages may not have the income or access to adequate medical care that older mothers do. Thus, the fact that maternal inputs vary with maternal age obfuscates the causal effect of maternal age. Specifically, women who give birth at younger ages are of lower SES than women who give birth at older ages. Thus, the adverse impacts of giving birth at a younger age may be overstated in the cross-section although the opposite is true for older ages.

The main empirical evidence of the effects of maternal age comes from sibling-based comparisons. That is, one can compare the outcomes of children born to the same mother. Such an approach effectively controls for fixed differences (e.g., SES which may be fixed) across mothers. However, to the extent that maternal age is correlated with other attributes that vary across a woman's lifecycle, these sibling contrasts will not capture solely the effect of maternal age. The sibling estimates do confirm the expected direction of biases – the effects of young maternal age are not as adverse as one would expect from correlations and the effects of advanced maternal age are worse than what the cross-sectional correlations imply.

Income

There is a well-documented, positive correlation between income and child health. Income is not a direct input into health production, thus the impact of parental income on child health must operate through either budgetary constraints or by shifting parental preferences. Higher-income parents can afford to purchase more food, health care, and safer environments for their children. Parental tastes for child health inputs may also vary by income, as evidenced by income gradients in smoking, drinking, and prenatal care.

The effect of income can operate through many channels and economists have distinguished between the effects of transitory and permanent income because each type may have a distinct impact on health outcomes. A temporary income shock (e.g., drought, famine, variation in rainfall) can have an immediate, one-time effect that lasts into adulthood, particularly if the shock occurs during gestation or just after birth. Permanent family income has a direct correlation with child health, with the impact of permanent income on health growing as children age into adulthood.

Disparities in health across socioeconomic groups are evident at birth. Low income children have a higher incidence of low birth weight, poorer reported health status, and higher rates of chronic conditions in childhood; however, there is little evidence that the impact of being low birth weight varies by SES. Researchers have documented an income–health gradient that steepens over time, indicating that the disparities in health between high and low income children grow with age. The hypothesized mechanism behind the steepening of the gradient is the prevalence of shocks experienced by low income children. Although a health shock does not differentially impact low income children, the higher frequency of shocks experienced by low income children causes the gap in health status to widen with age.

Temporary income shocks near the time of birth produce detectable effects on health in only some studies. Negative income shocks, such as the phylloxera infestation that destroyed 40% of French vineyards between 1863 and 1890 and the Dust Bowl phenomenon in the American Midwest during the 1930s have been found to have minimal effects on health in adulthood. Individuals born in a phylloxera-affected region were shorter than their unaffected peers; however, other measures of population health were unchanged. Health in old age was also unaffected for individuals born in the Dust Bowl era. Positive income shocks as measured by rainfall improved the adult health, height, and completed education of females in Indonesia who were less than 1 year old during the increase in rainfall. No results were found for men or for rainfall shocks while the child was *in utero*, suggesting that improved outcomes for women during high rainfall years may be related to gender bias in nutritional intake during infancy.

Means-tested government transfer programs provide an exogenous, measurable income shock to eligible families and have been shown to improve child health. Mexico's randomized-controlled experiment of PROGRESA provides cash transfers to households that comply with required behaviors including prenatal care, medical checkups, meeting nutritional guidelines, and attending educational meetings. Although it is not possible to separate the impact of the income transfer from the other features of the program, children born into the program have lower rates of illness than control families, are less likely to be anemic, and are slightly taller than control children. Furthermore, the impact of the program increased the longer the family received PROGRESA transfers. In the USA, it is unclear whether cash transfers to families participating in the Aid to Families with Dependent Children Program increased infant birth weight, whereas maternal participation in the Food Stamp Program (comparable to an income shock) increased the birth weight of infants at the low end of the birth weight distribution.

Macroeconomic conditions at the time of birth are related to both health at birth and long-run health and the relationship appears to have changed over time. Research using data on individuals born in the Netherlands between 1812 and 1912 finds that babies born in boom years have lower mortality rates later in life and live longer than babies born in recession years. More recent data suggest that the relationship between macroeconomic conditions and child health may have reversed. In the USA, a higher unemployment rate is associated with improvements in birth outcomes such as incidence of low birth weight and postneonatal mortality. During times of high unemployment, maternal health behaviors (smoking and drinking) improve and different types of women select into motherhood, which may explain the improved birth outcomes. Although aggregate birth outcomes improve during times of high unemployment, the impact of a job displacement for an individual family negatively impacts infant health. Comparing children in the same family, children born just after a parental job loss have lower birth weight than siblings born before the job loss.

Health Care

Prenatal care can improve infant health by identifying conditions that can harm health such as low weight gain and by providing health and nutrition information to the mother. Athough it is well documented by researchers that policy levers can improve rates of prenatal care utilization, it is still unclear whether increased prenatal care translates to better infant health. Examinations of Medicaid expansions yield mixed results, but other policy changes that increased care have resulted in improvements in birth outcomes. Access to prenatal care appears to improve birth outcomes for those most at risk for poor birth outcomes such as low-income women and minority women who would have otherwise had minimal or low-quality prenatal care. A primary mechanism through which prenatal care improves birth outcomes is to reduce maternal smoking, which is the leading cause of growth retardation for fetuses. Health care at the time of birth is associated with a decline in the neonatal mortality rate, likely a result of access to life-saving technology.

Public health insurance programs such as Medicaid in the USA and National Health Insurance (NHI) in Canada provide prenatal and delivery care with the goal of improving both infant and maternal health. Introduction of universal health insurance in Canada during the 1960s and 1970s reduced infant mortality by 4% and reduced low birth weight classification on average, with single mothers experiencing a substantial reduction in the incidence of low birth weight. In the 1980s and 1990s, Medicaid significantly expanded its eligibility threshold to include a larger share of low-income, pregnant women. The program expansion initiated cost-saving measures, changing the insurance structure from fee-for-service to managed care for some enrollees. Evaluations of the changes consistently show impacts on prenatal care utilization but yield differing results on birth outcomes, with some researchers concluding that the changes improved birth outcomes and others finding no effect. Physician incentives to provide care are influenced by the type of payment structure Medicaid uses. Of particular interest is the relative incentives of Caesarian versus vaginal deliveries. Reduced incentives to provide care have been shown to increase the probability of low birth weight, prematurity, and neonatal mortality; however, studies that examine increased incentives to provide care find no effect on infant health.

The 1964 Civil Rights Act mandated desegregation of hospitals and greatly improved the quality of prenatal care available to blacks, particularly in the southern USA where hospitals for non whites were of poor quality. Desegregation reduced postneonatal mortality rates with gains driven by reductions in preventable deaths from pneumonia and gastroenteritis. The health of infants at birth also improved, as evidenced by reduced incidence of low birth weight and improved APGAR scores for the cohort born after desegregation. The narrowing of the black–white test score gap in the 1980s can be traced back to improved health of black cohorts born after desegregation, indicating that access to care that improved birth outcomes translated to increased human capital development later in life.

Another way to identify whether increased care translates to better outcomes is to examine infants on either side of the 1500 g very low birth weight classification. Infants below 1500 g receive more intense care than infants just above the threshold, resulting in lower mortality rates for infants classified as very low birth weight. In line with the findings that improved care after desegregation increased the test scores of black children, very low birth weight infants just below 1500 g who received additional care outperform their peers with birth weights exceeding 1500 g.

Maternal Behaviors

Negative correlations between income and behaviors such as smoking, drinking, and drug use suggest that these habits may be a possible mechanism for transmission of health to infants. The decision to drink or smoke may be related to other maternal behaviors or characteristics that could affect infant health; therefore, an extensive set of control variables or a natural experiment that changes smoking behavior independent of maternal characteristics is necessary to isolate the impact of these behaviors' outcomes such as birth weight and infant mortality. Numerous studies have linked maternal drinking and smoking with reduced infant health and long-term human capital outcomes.

Alcohol

In a survey of Danish mothers who had recently given birth, women who reported drinking four or more drinks per week while pregnant were more likely to have a preterm delivery than women who reported drinking no alcohol. This finding may be a result of omitted variable if women who choose to drink during pregnancy are negatively selected on other attributes. Accordingly, there has been a shift to the use of quasi-experimental approaches to unraveling the alcohol and child outcome relation.

Variation in the legal drinking age across states and over time has been used to identify the causal effect of maternal drinking on infant health. A lower drinking age is associated with more alcohol consumption during pregnancy, an increase in premature births, and an increase in the probability of low birth weight. The reduction in health at birth can partially be attributed to changes in the composition of births, increasing the number of births without a father listed and suggesting that more unplanned pregnancies occur when drinking laws are less stringent.

Maternal alcohol consumption can have long-term effects on human capital development, as demonstrated by a policy experiment in Sweden. In 1967, grocery stores in certain regions were temporarily allowed to sell strong beer that was previously only available in government-run liquor stores. Children exposed the longest to the policy while *in utero* had lower completed education, lower earnings, and higher rates of welfare participation than children that were not exposed to the policy experiment.

Smoking

Smoking during pregnancy increases health risk for both the mother and infant in the form of complications such as miscarriage, membrane ruptures, ectopic pregnancy, pneumonia, and stillbirth. Women who smoke during pregnancy

have lower birth weight babies on average and are at a greater risk for having an infant classified as low birth weight. The seminal study of the impact of smoking on infant health is the randomized-controlled trial of Sexton and Hebel (1984), in which pregnant smokers were randomized into a treatment group receiving assistance quitting smoking and a control group receiving no intervention. Babies whose mothers were in the treatment group were on average 92 g heavier than control group babies.

The 1964 Surgeon General Report on Smoking and Health alerted the nation to the health hazards of smoking resulting in a reduction in smoking among pregnant women that was concentrated among higher-educated mothers. A study comparing birth outcomes of children before and after the release of the Surgeon General Report reveals that higher smoking rates are associated with lower birth weight. However, no effect of smoking was found on gestation, prematurity, or the likelihood of having a low birth weight baby. These results are similar to studies that use increases in cigarette excise taxes to estimate the impact of smoking on birth weight.

Nutrition

From famines in developing countries to supplemental nutrition programs in developed ones, studies consistently conclude that nutrition is a fundamental input into health production, impacting both short- and long-run health. Randomized-controlled trials that offer nutritional supplements to the treatment group have demonstrated that micronutrients play a key role in cognitive development. Assessing the direct impact of nutrition on health is difficult due to significant measurement error in the nutritional content of food items; therefore, most natural experiments examine how quantity of food relates to health outcomes. Research suggests that policies that improve the nutrition of pregnant women and infants will be effective at improving the health and human capital of the next generation.

The ideal setting for conducting research is the randomized-controlled trial, a technique that has been used in developing countries to study the impact of poor nutrition on cognitive development. In Jamaica, babies that were given nutritional supplements had higher mental development than the control group, indicating that lack of nutrition is a causal factor in stunted mental development. Children in Guatemala who received a nutritional supplement tested higher on knowledge, numeracy, reading, and vocabulary assessments than children given a placebo. The same children were followed up with as adults. Adults who were treated with the nutritional supplement as a child had higher reading comprehension, nonverbal and cognitive scores, and higher completed education (women only) than the control group.

The majority of economic research on nutrition in developing countries studies the impact of famines on health, education, and labor market outcomes. Famines are extreme events and estimating the impact of a famine can be confounded by selection because only survivors are observed. Furthermore, the health effects of a famine may not solely operate through nutritional deprivation – famines may affect other inputs to health and human capital such as disease-resistance and school attendance. The Chinese Famine of 1959–61 had a significant impact on children and babies *in utero* during the event. Children exposed *in utero* were shorter, lighter, and acquired fewer years of education than children born just before and after the famine. Exposure in early childhood had a detectable, yet smaller effect on long-term outcomes than *in utero* exposure. The famine also tilted the sex ratio in favor of girls, reduced the literacy rate, reduced employment, and reduced the marriage rate for children born during the time of the famine.

European famines during World War II had long-term impacts on health and human capital accumulation for individuals exposed early in life. Individuals who were *in utero* during the Dutch Famine experienced higher rates of chronic disease in adulthood. Children exposed to the Greek Famine during gestation and the first two years of life showed reduced educational attainment and literacy, with the largest impacts on children who were 0–12 months old during the famine. The impact of a famine can reach late into life – men exposed *in utero* to the Dutch Potato Famine of 1846–47 had a lower life expectancy at age 50 than cohorts born just after the famine.

Controlled nutritional deprivation for brief periods of time is associated with reduced physical and cognitive development, as evidenced by recent research into the outcomes of children *in utero* during Ramadan. Ramadan occurs for one lunar month per year and observance includes fasting between sunrise and sunset. In a study using data from the USA, Iraq, and Uganda, the authors document reduced birth weight, reduced gestation length, a decline in male births, reductions in educational attainment, and even increased rates of mental disabilities for children of Arab mothers *in utero* during Ramadan.

Even in developed countries, nutrition interventions can positively impact the birth outcomes of at-risk children as evidenced by analyses of the Supplemental Nutrition Program in the USA for women, infants, and children (WIC). WIC is aimed at low-income pregnant women and women with young children with the goal of improving the nutrition and health of this group. Consistently estimating the effect of WIC participation on infant health is difficult due to nonrandom selection into the program – unobserved maternal characteristics that affect infant health may be systematically different for mothers that choose to enter the program than for mothers who do not. Estimates that account for selection into the program yield a positive impact of WIC participation on birth outcomes such as incidence of low birth weight and gestation length. Infants at the low end of the socioeconomic and birth outcome distribution gained the most from WIC.

Environment

Environmental quality can be considered a direct input into health, with infant health responding to maternal exposure to pollution while *in utero* as well as post-birth. Isolating a causal relationship between pollution and health is challenging for many reasons. First, measurement error in pollution levels attenuates coefficients and makes a relationship difficult to detect. Second, there are numerous pollutants, many of which

are measured infrequently or not at all. Lastly, a number of confounding variables must be ruled out in order to interpret a relationship between environmental quality and health as causal. For example, families may sort into areas of varying pollution levels based on socioeconomic characteristics or business cycles may have an independent effect on both pollution levels and health. Furthermore, the relationship between health and pollution may be nonlinear, meaning that reductions in pollution below a given level may not improve health.

Researchers have exclusively relied on quasi-experimental designs such as policy changes or temporal variation in pollution levels to assess the impact of environment on infant health. The introduction of the Clean Air Act of 1970 reduced infant deaths in the most polluted counties. Similarly, infant mortality declined more in counties with greater reductions in total suspended particulates during the 1981–82 recession. The introduction of the EZPass toll system in the Northeastern USA reduced traffic and thus pollution levels near the freeway, subsequently increasing birth weight and reducing prematurity for newborns near the freeway. The Chernobyl fallout over Sweden did not detectably affect infant health; however, students that were *in utero* during the fallout experienced deficiencies in human capital as evidenced by lower test scores and high school graduation rates.

Conclusion

From both a theoretical and empirical perspective, there has been an increasing focus on the importance of *in utero* and early life conditions on later health and outcomes. Theoretical models emphasize the timing of investments. If investments are substitutable across periods, then disadvantage early in life can be overcome by later life investments. However, early investment is important if skills acquired during early periods can help beget skills in later periods. This article highlights several mechanisms through which transmission of health may occur – initial endowments, environmental influences, parental abilities, and investments. Researchers have relied heavily on quasi-experimental strategies such as policy changes, natural disasters, and sibling studies to identify a causal relationship between early life influences and health. This is an emerging and growing literature.

See also: Alcohol. Education and Health: Disentangling Causal Relationships from Associations. Education and Health in Developing Economies. Education and Health. Fetal Origins of Lifetime Health. Macroeconomy and Health. Nutrition, Economics of. Pollution and Health. Smoking, Economics of

References

Heckman, J. (2007). The technology and neuroscience of capacity formation. *Proceedings of the National Academy of Sciences (PNAS)* **104**(33), 13250–13255.

Sexton, M. and Hebel, J. R. (1984). A clinical trial of change in maternal smoking and its effect on birth weight. *The Journal of the American Medical Association* **251**(7), 911–915.

Further Reading

Almond, D. and Currie, J. (2011). Human capital development before age five. In Ashenfelter, O. and Card, D. (eds.) *Handbook of labor economics, vol 4B*, pp. 1315–1486. Amsterdam: Elsevier.

Almond, D. and Currie, J. (2011). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives* **25**(3), 153–172.

Barker, D. (2004). The developmental origins of adult disease. *Journal of the American College of Nutrition* **23**(supplement 6), 588S–595S.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature* **47**(1), 87–122.

Currie J. (2011). Inequality at birth: Some causes and consequences. *National Bureau of Economic Research Working Paper No. 16798.* Cambridge, MA: National Bureau of Economic Research.

Relevant Website

http://www.thebarkertheory.org/
The Barker Foundation.

Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity

P Serneels, University of East Anglia, Norwich, Norfolk, UK

Introduction

This article discusses internal imbalances of health care in low- and middle-income countries. Throughout this article, 'internal' refers to within country, and the emphasis will lie on the differences between rural and urban areas. Much of the work in this area focuses on the imbalance in the quantity of health workers, but recent evidence indicates that imbalances in the quality of health care are as important. The paper discusses both, focusing throughout on the human resources aspect of health-care delivery.

What Geographical Imbalances?

With poverty reduction taking a prominent place on the international agenda in the early-1990s – later resulting in a consensus around the 2015 Millennium Development Goals (MDGs) – there has been an increased interest in rural service delivery. Many of the poor live in the countryside, where poverty is at its deepest, and the emphasis placed on coverage of cardinal interventions makes access to services in rural areas key to reach the MDGs. But although health outcomes are unfavorable in rural areas, there is also less care provided in those areas. This is sometimes referred to as 'the inverse care law.' These geographical imbalances have mostly been discussed in terms of shortages of health-care workers, which is perhaps best illustrated by the World Health Organization guideline recommending 2.28 health professionals – including doctors, nurses, and midwives – per 1000 inhabitants to allow the delivery of quality health services. Contemporary work is concerned with both the quality and quantity of services. Evidence indicates that low numbers are not the single constraint for the delivery of appropriate services. A narrow focus on the numbers of health personnel is therefore misguided. It also stands in the way of thinking critically about health care in remote areas, particularly in the context of rapid urbanization, as is the case in most developing countries, which may require more fundamental changes to rural health policies, as discussed in the section Encourage and Support Self-Help among Rural Populations. In what follows, the paper discusses the evidence on quantitative and qualitative imbalances in human resources for health (HRH).

Imbalances in the Number of Health Workers

Although a focus on the quantity of health-care providers is not enough, considering the figures does provide a starting point and reveals striking differences. Table 1 illustrates the within-country geographical imbalances across the world for the countries for which there are data available.

The contrasts are stark. On average, more than 80% of doctors work in urban areas, and the remaining 20% works in rural areas. The figures are more favorable for nurses, midwives, and medical assistants, of whom approximately 40% work in rural areas. The distribution is more skewed for dentists, pharmacists, and radiographers, of whom 18%, 12%, and 18%, respectively, work in rural areas. This implies that urban areas count, on average, 15 times more physicians, 6 times more nurses, and 3 times more midwives and medical assistants for the countries in the dataset. This ratio is higher for radiographers, dentists, and pharmacists, who are typically employed in hospitals or in the private sector. With more than 45% of people living in rural areas worldwide, the overall distribution is highly skewed in favor of urban areas.

A number of shortcomings to the data limit the inference that can be drawn from these figures. First, the data are available for only a relatively small sample of countries, with sub-Sahara African countries very well represented but other continents heavily underrepresented, as is clear from Table 2.

The data also suffer from a number of biases. Countries with a weak administration, ill-functioning government, or in conflict are largely missing from the data; they are also likely to have higher concentrations of health professionals in urban areas. The same applies to regions within countries: areas with weak governance are more likely to have missing data. A second bias stems from the lack of data on private sector health workers as the figures only reflect public sector health professionals. Both types of bias will lead to underestimation of health professionals in urban areas, and thus an underreporting of the problem.

Studies at the regional level paint a similar picture, confirming the general pattern and also highlighting divergences between regions. A recent study on sub-Sahara Africa, where the problem is deemed most striking due to the relative high proportion of the population living in rural areas, illustrates this. The results summarized in Figure 1 show the concentration of doctors in urban areas for 13 countries. Densities are considerably higher in urban areas, a pattern that is confirmed by other country-specific studies. In Cote d'Ivoire, for example, 70% of all doctors work in the southern, urban regions that harbor only 40% of the population, and similar disparities are seen in data from Zambia, Sudan, and Uganda.

In Asia, the case of Thailand has been well researched. Several studies provide updated estimates of the geographic distribution of the country, illustrating that Bangkok has four times more nurses per 10 000 people than the North East, the most rural region. A similar picture emerges for Bangladesh, where 30% of nurses are located in four metropolitan districts that represent 15% of the population. An early study confirms the problem of urban–rural imbalances for Indonesia. China provides another interesting example because the majority of its nurses (98%) and doctors (67%) have been educated only

Table 1 Types of health professionals in rural and urban areas worldwide

Health professionals	*Share of health professionals in rural areas (mean)*	*Ratio of urban to rural health professionals (mean)*
Physician	0.20	15.6
Nurse	0.39	6.3
Dentist	0.18	18.1
Pharmacist	0.12	11.8
Midwife	0.39	2.8
Radiographer	0.18	23.4
Medical Assistant	0.37	2.9

Note: Author's calculation from WHO Global Atlas data.

Table 2 Countries in cross-country dataset

Region	*Countries*	*Number of countries*
East Asia and Pacific	Myanmar, Timor-Leste	2
South Asia	India, Maldives, Pakistan, Sri Lanka	4
Middle East and North Africa	Algeria, Djibouti, Egypt, Iraq, Morocco, Oman, Tunisia, Yemen	8
Europe and Central Asia	Romania	1
Latin America and Caribbean	Brazil, Honduras	2
Sub-Sahara Africa	Burkina Faso, Benin, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Côte d'Ivoire, Democratic Republic of the Congo, Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Sierra Leone, Sudan, Swaziland, Togo, Uganda, Tanzania, Zambia	35

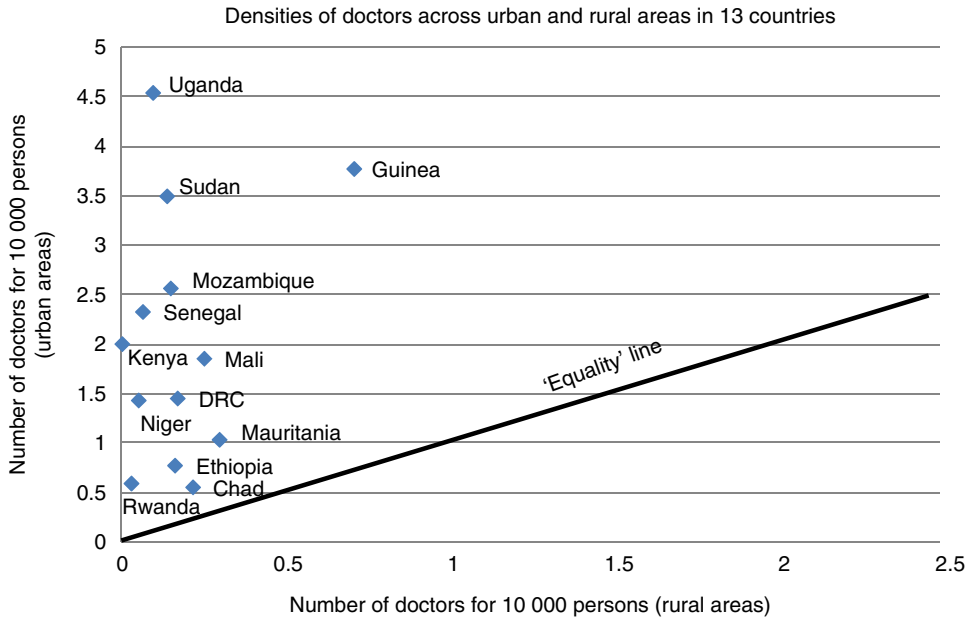


Figure 1 Density of doctors in urban and rural areas in 13 African countries. Data from Lemiere, C., Herbst, C. H., Dolea, C., Zurn, P. and Soucat, A. (2013). Rural-urban imbalances of health workers in sub-Saharan Africa. In Soucat, A., Scheffler, R. and Ghebreyesus, T. A. (eds.) *The labor market for health workers in Africa. A New Look at the Crisis*. Washington, DC: World Bank.

up to junior college or secondary education. This provides a unique setting in which the level of education for health professionals, often believed to be a major explanatory factor for reluctance to work in rural areas, is relatively low, and because this has resulted in having more doctors than nurses. Still, urban China has more than twice as many doctors and more than three times as many nurses per capita than rural China.

Evidence for Latin America also shows similar patterns, with health workers concentrated in the capitals and more affluent areas. In Argentina, which has one of the highest numbers of health workers per person, Buenos Aires counts seven times more doctors per capita than Formosa or Misiones. In Chile, approximately 60% of public sector health professional are concentrated in the region of Santiago, which hosts only 40% of the population. In Ecuador although the capital's province of Pichincha has 2 doctors per 1000 inhabitants, the more remote provinces of Galápagos and Orellana Esta have 0.56 and 0.43, respectively. In Guatemala, which has 0.9 doctors per 1000 inhabitants on average, 71% of health workers are concentrated in the metropolitan zone, whereas the more remote Quiche has a ratio of less than 0.1 doctors per 1000 inhabitants. A similar imbalance exists for all medical professions in this country. Nicaragua has 0.4 doctors per 1000 inhabitants on average, whereas its capital, Managua, has 1.1. In Peru, 53% of physicians, 40% of nurses, 44% of dentists, and 41% of technicians and health assistants were concentrated in the Lima Metropolitan Area, which represents close to one-third of the country's population. Stark differences in densities across professional groups are also reported for the Dominican Republic, with 0.41 nurses per doctor in Region 1 and 3.63 in Region III. A detailed analysis of Brazil also shows substantial inequalities in numbers of health workers.

Although this article focuses on low- and middle-income countries, it is useful to remember that the same problem exists in high-income countries. In USA, for instance, 9% of physicians work in rural areas, which represent 20% of the population. The figures for Canada are very similar (9% of physicians for 24% of the population). In France, the wealthy areas of Paris city and the South have considerably more doctors than the rest of the country. In Norway, the rural and remote Northern areas have historically been underserved, and there is long tradition of policy making to try to address this.

The pattern that emerges from the above descriptive statistics is clear, but also blunt. For a subset of countries with more detailed data, a more advanced analysis and decomposition of inequalities across and within subgroups is possible – for example, according to profession or gender. A 2008 study considers the Gini coefficient, Theil T, and Theil L measures for the distribution of health workers in China, and decomposes overall inequality in between and within province inequality. The findings indicate that underlying distributions can be very different across regions. Across the measures, within province inequality accounts for between 82% and 84% of intercounty inequality. A later study draws similar conclusions for sub-Sahara Africa, calculating the Concentration Index and Gini coefficient for doctors and nurses for nine sub-Sahara African countries. Although the Concentration Index for doctors varies between 0.25 and 0.48 (for Kenya and Senegal, respectively), that for nurses varies from 0.05 to 0.54 (for Kenya and Mauritania, respectively), and generally confirms that imbalances tend to be more severe for more educated health professionals. The results for both China and Africa illustrate how aggregate figures on numbers of health workers can be misleading and may provide a highly insufficient base for policy making.

Imbalances in the Quality of Care

With the development strategy of the past decades emphasizing increases in the supply of care, much of the debate has focused on gaps in the quantities of health-care providers. But there are also important imbalances in the quality of care that patients receive (often referred to as process quality), which is a function of the number of health workers, their performance, and the availability of complementary inputs. This chapter focus on performance differences stemming from human resources and abstracts from differences in the import of complementary inputs like the clinic's physical condition or the availability of drugs (often referred to as structural quality).

The significance of health worker performance – or better, underperformance – is perhaps best illustrated by the results of surprise visits to health facilities in six developing countries that found 35% of health workers to be absent on average. Although the study does not set out to compare between rural and urban areas, it finds that absenteeism is generally higher in poorer areas and among higher qualified health professionals (e.g., doctors). Results from qualitative studies in Ethiopia, Rwanda, and Ghana suggest that absenteeism is higher in rural areas mainly because of poor monitoring. Other work illustrates the importance of on-the-job performance. High health-care usage rates, combined with poor health outcomes often indicate problems with quality of care. A 2007 study provides direct evidence for underperformance of doctors in rural Tanzania. Measuring what doctors know (using a vignette) and comparing this with what they do (using direct observation as well as patient recall), the authors observe a substantial know–do gap. In other words, these doctors provide lower quality care than what they could provide. Qualitative studies in Ethiopia, Rwanda, and Ghana also find indications that health worker attitudes toward patients tend to be poorer in rural areas, whereas performance problems like corruption and embezzlement seem to be higher. This is supported by studies of corruption in the health sector in Tanzania. Further underlining the importance of taking quality of care into account, other studies find that households in Tanzania bypass low-quality facilities that are nearer and increase travel time to reach facilities with better care. Also relevant are study results on medical quality in urban and rural areas in five countries that find that households in poor areas not only have more access to private facilities that provide low-quality care, but are also more likely to receive low-quality care in any facility, particularly in the private sector. The inquiry also finds that indigenous patients that come from a poorer background receive less quality care in the private sector, and infers that this is due discrimination against those patients (rather than households choosing low-quality providers). A separate study shows how workload is not the reason for poor performance among health workers in Tanzania, observing that clinicians have ample amounts of idle time. The authors conclude that scaling up the number of health workers is unlikely to raise the quality of health care. Taken together, these study results provide strong evidence for quantitative and qualitative imbalances between urban and rural areas.

Implications of Imbalances for Health Outcomes

A number of studies have looked at the implications of quantitative imbalances. Evidence from cross-country regressions

suggests that the number of health workers has a strong relationship with health outcomes. Controlling for GNI, income poverty and female adult literacy, it has been found that HRH density is strongly related with especially maternal mortality, but also infant and under-five mortality. Decomposing the effect for doctors and nurses, there is a large association for the former and absence of such an association for the latter – except for maternal mortality where nurses do seem to play a role. Other work, looking at the disease burden, also tends to find a (negative) relationship. One study of the relationship between different health worker densities and DALY's (and DALY's disaggregated according to three different groups) finds a strong relationship with the number of doctors in particular, whereas the association for nurses and midwives is insignificant. A similar analysis argues that countries with fewer than 2.5 health workers per 1000 population are very unlikely to achieve a desirable 80% level of coverage for skilled birth attendance and measles immunization. A further inquiry, updating these studies by making use of an extended sample of 192 (instead of 177) countries, finds an aggregate relationship between health worker density and measles immunization and birth attendance, but no longer with infant and under-five mortality. It also observes a significant association for doctors but not for nurses, concluding that that the threshold should be 2.28 rather than 2.5 health workers per 1000 population.

Although the findings from these studies are indicative, they do not provide conclusive evidence for a causal relationship, as the analysis may suffer from omitted variable bias. The number of health workers may, for instance, be correlated with government expenditures on health, donor activity, the number of clinics, the availability of equipment and medicine, or the presence of conflict, none of which are included in the analysis. Other factors, such as skill mix, negative work environment, and weak knowledge base may also be important, and their omission may further bias the estimates upwards. Conversely, the lack of a relationship between nurses and health outcomes may also be due to unobserved factors – like absenteeism among health workers – which may bias the estimates downwards. Another potential problem is that the sample suffers from selectivity. Countries with good data tend to be better organized and may have surmounted other constraints that may matter more than the number of health workers. There is, finally, also a question as to how the limitations in data comparability across countries play a role. Different countries use distinct definitions, for instance for nurses, and this introduces both measurement error and unobserved heterogeneity. More sophisticated approaches are needed if one would like to test the robustness of these findings, as recognized by the authors of some of these studies.

To address some of the shortcomings associated with using aggregate cross-country data, more recent studies focus on within country variation using subnational data. One such analysis of China concludes that the density of doctors and nurses is significant in explaining differences in infant mortality across counties in China. A similar approach to data from Brazil finds that a 1% increase in health worker density is associated with a 0.12% increase in the coverage of antenatal care on average. The papers also illustrate that there is considerable variation in the level of coverage by municipality for a given number of health workers, thereby illustrating how the analysis suffers from similar shortcomings as the ones mentioned above. As a result, they do not provide conclusive evidence on the extent to which shortages of health workers cause poor health outcomes. Although identification of causality remains a challenge, it is necessary when informing policy making, especially when providing advice on target numbers of health workers.

An analysis of data from Ghana yield evidence on a causal relationship between the number of health workers and demand and usage of health care. Making use of exogenous policy changes in the late-1980s, it is found that increasing the number of doctors and nurses to three (representing a 50% increase from the mean) would lead to a 20% increase in the predicted probability of households choosing public health care. A recent study focusing on Indonesia also provides causal evidence that relates the number of health workers and quality of health care to health outcomes. Exploiting the fact that deployment of health staff in Indonesia is based on quantitative targets per facility although not related to quality or health outcome targets, it was found that increasing the number of MDs, nurses, and midwives increases adherence to clinical protocol, which in turn leads to improved child health (measured by length). The largest gains are made by increasing the number of MDs, followed by nurses, whereas increasing the number of midwives had no effect. As the study did not include the most remote areas in Indonesia, its estimates may well be conservative.

Recent evidence for Kenya shows how absenteeism causes poorer health outcomes. Using longitudinal data for rural health clinics, it is shown how women whose first clinic visit coincides with nurse attendance are approximately 60 percentage points more likely to be tested for HIV and 13% more likely to deliver in a hospital or health center, and how this in turn affects expected HIV status. The presence of other health workers may also increase the quality of care, as shown by one study interpreting Hawthorn effects in direct observation of doctor activity as evidence that performance increases when colleagues are present.

The above-mentioned literature confirms the causal relationship between quantities of health workers and quality of care on the one side and health outcomes on the other, but does not allow clear conclusions to be drawn on the relative importance of these factors. To identify pathways through which rural health outcomes can be improved, the next step is then to return to theory to better understand why quality of care and numbers of health workers are lower in rural areas and result in lower health outcomes.

Causes of Imbalances in Health Care: Theory and Evidence

From a theoretical perspective, there are a limited number of reasons why health outcomes in rural areas are lower. Abstracting from potential differences in disease burden, three factors play a role: poor infrastructure – including scarcity of clinics, lack of equipment, medicine, etc.; weak human resources which relates to the number of health workers, their presence and performance, as well as the combination of health worker types; and, finally, limited demand for health

care, which is related to households' information and health seeking behavior. There is currently limited understanding of the relative role of these factors, and where the binding constraints lie. This relationship can be presented in a more systematic way, by considering patient health outcomes (H) as a function of the infrastructure of the facilities in that area (K), the human resources in the facility (L), which entails number and different types of health workers (n), their presence (p) and performance (γ), as well as patient household characteristics in that area (hh). It is helpful to think of this relationship in terms of a production function where the inputs are imperfect substitutes, and write health outcomes as a product of these three factors, with their power reflecting their relative weight.

Health worker inputs (L) can thus be seen as the product of three factors: the number, presence, and performance of the respective health worker categories. A next step is to consider the determinants of these respective factors. This chapter focuses on two issues central to human resources: the quantity of health workers in rural areas (n) and their performance (γ). The paper refers to other work for in-depth discussion of absenteeism and issues not related to human resources, including infrastructure, availability of drugs, funding, and factors to do with demand.

Quantity of Health Workers (n)

Ultimately, the relatively low numbers of health professionals in rural areas is rooted in the choice of health workers themselves. Job choice is typically modeled as a process of matching between job attributes and preferences. Focusing on earnings and effort, in addition to other job attributes like social status, recent work adds motivation, which is especially relevant for professions where a personal mission is important, like in public service. Considering that health workers will choose to work in a rural area when they expect to derive more utility from a rural than an urban job, this framework predicts that, since earnings and amenities typically receive high weights, while differences in effort between rural and urban areas may be limited (even if weights to effort may be high), most health workers prefer an urban job. Only those with a mission that matches to working in rural areas, or those who attach a high value (weight) to living or working in a rural area, for instance because of proximity to family and friends, prefer a rural post. These predictions immediately illustrate the limited leverage that policy makers have at their disposal if they want to get more health workers into rural areas. Although in theory people can be compensated for unattractive job attributes, for most health workers, earnings will have to be very high to compensate for the disutility caused by poor amenities in rural areas.

This situation may be aggravated when taking a more dynamic perspective and consider health workers to be making a career rather than a job choice. Taking a lifetime perspective, the outcome is now determined by the discounted sum of utilities across different periods, allowing for health workers to change from rural to urban areas, with income in each period a function of human capital accumulated in the previous periods (h_{t-1}). If an individual expects that the accumulation of human capital is slower in rural posts, for example, due to lower opportunities for formal training or because the type of experience built is not rewarded in urban jobs, she may be even less likely to choose a rural post. In a more sophisticated approach, valuations of job attributes could be allowed to vary as staff gets older. The weight attached to amenities may change and health workers may stick higher values to jobs in urban areas at certain ages, for example, at marriage age because it offers access to a larger pool of potential marriage partners, at child-bearing age because it offers access to better child care, or when children reach school going age because of the proximity of better schools, etc.

The basic predictions of the above models are supported by empirical evidence. Results from comparative qualitative research illustrate why health workers in Ethiopia, Ghana, and Rwanda generally prefer jobs in urban areas. Although rural jobs may offer extra payment and benefits, these are usually insufficient to compensate for other disadvantages. Professional isolation, limited access to training, and poor working conditions characterized by limited access to equipment and infrastructure are seen as strong drawbacks. Urban postings also provide the possibility of working in a second job in the private sector, which is usually absent in rural areas. But the reasons why rural posts are unattractive go beyond job attributes, as factors like personal isolation, the general absence of infrastructure and amenities, including the low quality of housing and absence of good schools, also play an important role. Rural postings are associated with lower career perspectives as well, as they provide less access to training, limited access to equipment and modern technology, and thinner professional networks, among others. In some rural areas, salaries are often paid with delay. However, the lack of supervision from colleagues may give more freedom in rural posts.

A growing body of quantitative work analyses health worker willingness to work in rural areas, typically studying the role of wages and other job attributes. In the absence of incentive compatible study set ups, two types of methods have been applied contingent valuation and discrete choice methods. Each of these methods have their advantages and draw backs. While contingent valuation methods find the precise reservation wage to work in rural areas, discrete choice methods focus on trade-offs between different sets of attributes. A 1998 investigation of health worker willingness to work in remote areas in Indonesia uses the first method to find that modest cash incentives can make health workers more likely to work in moderately remote areas, but that it would be prohibitively expensive for staffing of very remote areas. Health workers who grew up in remote areas are found to require lower compensation to take up a remote position. Results from a cohort study with final year health students in Ethiopia also find that expected wages affect take up of a rural post. Here, in order to get 80% of health workers in rural areas (who harbor 80% of the population), salaries would need to increase with 83% for doctors and 57% for nurses, requiring an increase in annual health expenditures of 0.9%. The study also observes substantial heterogeneity, with health professionals who grew up in more remote areas, come from a less wealthy background, or are more motivated to help the poor being more willing to work in rural areas. Assessing what

other job attributes matter, the study finds that chances for promotion, access to professional training and access to schools for education of children turn out to be important. Other studies present very similar findings. A resurvey of the same Ethiopian health professionals 2 years later (when they had entered the labor market), finds that wages and other job attributes are only part of the story and that health worker characteristics like rural background and motivation play an important role, with the latter influenced by the type of school attended. An identical study with health students in Rwanda confirms these results with rural background and motivation to help the poor as important determinants of willingness to work in rural areas. Health workers who were participating in a local (church-based) bonding scheme were also more willing to work in remote areas. Another similar study in Ghana finds that doctors who grew up in a rural area, as well as those with higher motivation are more willing to work in rural areas.

Results from discrete choice studies provide further insights into the relative roles of job attributes. Focusing on doctors in Ethiopia, it has been found that doubling wages would increase the share of doctors willing to work in rural areas from 7% to more than 50%, whereas providing high-quality housing would increase it to 27% (the equivalent of a wage bonus of 46%). For nurses, doubling the salary would increase their number from 4% to 27%, whereas the nonwage attribute that is most effective in inducing take up of a rural post is the quality and availability of equipment and drugs, which would reach the same result as a salary increase of 57% for men and 69% for women. Focusing on Tanzania, another study finds that offering continuing education after a certain period of service, as well as increasing salaries and hardship allowances, would encourage health workers to work in rural areas. Decent housing and good infrastructure were also found to be important. Women were found to be less responsive to financial incentives and more concerned with factors that directly allow them to do a good job, whereas those with parents living in a remote rural area are generally less responsive to the proposed policies. When willingness to help others is a strong motivating factor, policies that improve conditions for assisting patients are effective. Analyses of similar discrete choice experiments with health students in Kenya, Thailand, and South Africa, underline that results can strongly differ between countries. Financial incentives are likely to have important effects, especially in poorer countries, but only if they are larger than a 10% salary increase as smaller raises were found to be ineffective in all three countries. Nonfinancial incentives are found to be important as well, especially access to training and career development opportunities. Improved housing and accelerated promotion were moderately effective. A study using propensity score matching also suggests that improving Clinical Officer's access to upgrade training would not improve their retention in rural areas. A study of the situation in Liberia and Vietnam has found that although in Liberia increased pay would be the single most-powerful incentive, long-term education was the primary factor in Vietnam, and considers the differences in cost effectiveness of implementing corresponding policies. A recent study of Uganda sets out to design packages able to get medical and nursing officers in rural and remote areas using discrete choice methods. The preferred package for medical officers is a 100% increase in salary (from a current base salary of 750 000 Uganda Shilling), improvements to health facility quality, a contractual commitment to the posting for 2 years, and full tution support for continued education at the end of the contractual commitment. For nursing officers, the most preferred package contains a 122% increase in salary, improvements to health facility quality, and improved support from health facility managers. These packages would get an estimated 82% of medical officers and 90% of nursing officers in remote areas. Other studies that do not focus directly on the rural–urban choice also shed light on the importance of job attributes. Evidence for Malawi showed that graduate nurses valued high pay, as well as the provision of housing and the opportunity to upgrade their qualifications quickly. In South Africa, earning more was most attractive, whereas better facility management and equipment were next. Nurses in rural areas were more concerned about facility management.

Both quantitative and qualitative evidence indicate that there is important heterogeneity in health workers' willingness to work in rural areas. Although the majority of health workers prefer not to work in rural areas, some do, in particular in provincial towns. Rural background in particular has been found to be strongly positively associated with willingness to work in rural areas in Indonesia, Ethiopia, Rwanda, and Thailand, among others. Higher-level health workers (e.g., doctors) are generally less willing to work in rural areas compared to lower-level ones (e.g., clinical officers or nurses), for example, in Ethiopia and Uganda. Female health workers are often less willing to work in rural areas, as shown by evidence from Congo and Ethiopia, often for security or marriage-related reasons. Younger health workers may also be more likely to take up a rural post as part of their training, although their willingness may fall rapidly when entering the labor force, as found for Ethiopia. A number of studies also observe heterogeneity in health worker motivation, with health workers who are more motivated to help the poor more likely to take up a rural post. Identical surveys among medical and nursing students in Ethiopia and Rwanda both find that helping the poor is an important explanatory factor for willingness to work in a rural job for a substantial minority of health workers. This result for intrinsic motivation is strikingly similar for the two countries, and indicates that some health workers prefer to work in rural areas because this provides for a better match between their own beliefs and the belief of the facility they work for. Recent work also finds evidence for mission matching in nonprofit organizations in Ethiopia. The higher motivation to work in rural areas in Ethiopia is also linked to the school where one was trained, with health workers trained at an NGO school more willing to work in a rural area. This suggests that either health workers get socialized into motivation, or that they self-select at an earlier stage and choose the school that matches their beliefs and motivations. Overall, this evidence underlines that certain types of health workers self-select into rural jobs.

Qualitative studies also suggests that other factors, like appreciation for a slower pace of life, may play a role, indicating that adverse selection may be important. A recent test of whether less skilled health workers – as measured by a medical knowledge test – are less likely to work in rural areas in Ethiopia finds no evidence for such adverse selection. Exploiting the

existence of a lottery for allocating doctors to jobs, however, another study finds that adverse selection may occur in a different way, with lottery participants, who are not able to use their first job as a signal of ability, having flat wage profiles and higher exit rates.

Although rigorous studies using revealed (rather than stated) preferences and identifying causal effects are currently absent in this area, the above evidence provides a base for an increased understanding of the labor market for health workers in low and middle-income countries. It also points already to three types of policies: those working on the demand side using wages and job attributes, those operating on the supply side focusing on certain profiles of health workers (such as those with a rural background), and policies considering matching of demand and supply and allocation of health workers to posts (or vice versa). The section Lessons for Policy Making and Ways Forward will discuss these policy options in more depth.

Performance of Health Workers (y)

The performance of health workers can best be understood in a classic principal agent framework, where it can be seen as a function of three factors: incentives (w), monitoring (s), and individual motivation (m). Like before, incentives are used in a broad sense, and monitoring includes both supervision and accountability to the local community; it can also include workplace, professional, and society norms regarding professional behavior.

Qualitative research suggests that performance problems of health workers may be more important in rural and remote areas taking the form of absenteeism, poor attitudes toward patients, engagement in corruption and embezzlement, or poor performance in general. Moral hazard seems mostly attributed to four factors: the perceived lack of compensation for personal and professional sacrifice; poor monitoring and enforcement; a culture of poor performance with weak norms; and lacking motivation. The public sector in general is associated with more corrupt practices, and in a number of places, a culture of corruption and free riding is deeply embedded in the public health sector.

Quantitative evidence on determinants of performance is scarcer. One study, using data for Tanzania, provides a good starting point, comparing performance of doctors in the public and private not-for-profit sectors. Like in much of the rest of Africa, these two sectors share a similar mission, often run similar health facilities, and many not-for-profit facilities also follow public sector salary scales. It was found that clinicians in the not-for-profit sector have almost exactly the same average competence as clinicians in the public sector, but their adherence to the prescribed script, an indicator of quality of care, is higher. Thus, although the not-for-profit sector hires clinicians with the same capacity as the public sector, clinicians in the not-for-profit sector perform better.

In other settings, it is more relevant to compare health workers in the public with those in the private, for-profit sector. This approach was used in a 2007 study of Delhi. It was found that, on the whole, private sector providers spend substantially more time and effort on patients. Public sector providers also do less than they know they could. This sector disparity further masks variations in the public sector, with public providers in smaller clinics and dispensaries performing substantially poorer than public providers in hospitals, who tend to do comparable to private practitioners.

Although these studies generate relevant insights into differences in performance, they do not provide guidance for ways forward. Differences in the observed average know–do gap between sectors, sometimes seen as a measure of motivation, can arise for a variety of reasons, as variations between sectors are many (including pay, type of contract, monitoring, work environment, funding available, etc). They may, in addition, arise from the different types of workers that they employ.

There is also useful evidence of how performance can change. Using the above-mentioned data for Tanzania and exploiting the presence of a Hawthorne effect, it has been shown that being observed leads to higher effort and that there exists a link between variation in doctor performance and ability and motivation. An exploration of ways forward shows that clinician performance can be improved by peer encouragement as well as token gifts. Unconditional encouragement, where doctors are asked to do more, seems at least as useful, and has at least the same long-run impact, as conditional encouragement, where doctors are incentivized to do more, although little is known about the long-term effects in either case. Assistant Medical Officers in particular are able to do much better without significant additional resources and have sufficient capacity but insufficient motivation.

Studies identifying a causal effect of incentives, monitoring, or motivation on health worker performance remain scarce, but two studies stand out, one on pay for performance, and one on community monitoring. The remainder of this section discusses each in turn. The first study reports the results of a quasi-experiment in Rwanda where part of the funding received by health facilities depended on their performance. Results from qualitative research have illustrated many performance challenges in Rwanda. Both health workers and users point toward serious problems with health workers' attitudes toward patients, which are often characterized by impolite and rude behavior. Absenteeism and shirking are common problems, and some public facilities have 'ghost doctors' who are on the pay roll but do not show up for work. Especially in urban areas, absenteeism seems mostly related to having a second job, usually in the health sector and often unofficial. In this context, linking part of the funding that facilities receive to their performance may have beneficial effects. Indeed, the study finds large and significant positive effects on deliveries and preventive care visits by young children and improved quality of prenatal care, but no effects on the number of prenatal care visits or on immunization rates. It was concluded that pay for performance had the greatest effect on services with the highest payment rates that required the least provider effort. Unfortunately, the study did not investigate how health worker behavior changed. Results from qualitative research shed some light, and suggests that performance pay decreased absenteeism as well as shirking and improved the work environment, to the general satisfaction of health workers.

The second study provides evidence from a randomized intervention on community-based monitoring of public

primary health-care providers in Uganda. Making use of local NGOs to encourage communities to hold their local health providers accountable, the study finds that 1 year after the intervention, child mortality had seen significant reductions and child weight had increased in treatment communities. Studying the underlying processes, evidence was found for increased monitoring from the community through existing and new channels (e.g., local councils, evaluations) and increased activity from health workers, including improved consultation of the community (suggestion boxes), better information provision (posters regarding family planning), better management of patient care (numbered waiting cards), and higher medical effort (immunizations). The study also finds a drop in absenteeism, an increased use of equipment, a reduction in patient waiting times. A follow-up paper analyses the heterogeneity in some of these treatment effects and argues that the local social context, for example, income inequality and ethnic fractionalization, plays an important role and negatively affects the community's drive to collective action, which in turn holds back improvements of service provision.

Lessons for Policy Making and Ways Forward

Although this field would benefit from more structured research that takes special care in identifying causality, the above studies have increased the understanding and suggest five possible ways forward to address geographical imbalances in both quality and quantity of care, focusing on human resources.

A first approach concentrates on the supply side of labor for health care, given demand. A second approach starts from the demand side, focusing on how facilities can acquire the human resources they want or need, given the available supply. A third approach looks at matching health workers and jobs and how health workers get allocated to jobs. Fourth, one can look at the coordination between public and private sectors. Finally, new directions can be explored to encourage self-help in rural populations. In what follows, each of these are discussed in turn. In a final section, the authors discuss how one can go about choosing between or combining these different alternatives.

Emphasizing Labor Supply

Most human resource policies in the health sector today focus on the supply side. Starting from a needs-based approach, they concentrate on how to attain the desired number of human resources in rural and remote areas. Although this approach may be justified in settings with extremely low numbers of health workers, it generally suffers from a number of problems. A key issue that remains unclear, for instance, is what the ideal number of health workers should be. As the precise causal relationship between the number of health workers and improved health outcomes is yet to be understood, current figures represent preliminary estimates. One underlying assumption is that existing human resources are fully used, whereas evidence indicates that this is not the case. This approach also abstracts from the quality of care, assuming that existing and new health workers all provide high-quality care, in contrast to existing evidence. It also tends to undervalue the potential role of modern technology (see section How to Choose the Appropriate Approach?). An exclusive focus on the quantities of health workers supplied thus misses several important points. As a result, policies grounded solely in this approach often disappoint and are not the silver bullet they are believed to be. A classic example is the training of more health workers, which is often presented as a promising strategy to address shortages in rural areas. However, if health workers are free to choose where they work, training more professionals does not necessarily lead to more health workers in rural areas.

An important step forward with supply-based policies lies, therefore, in recognizing the heterogeneity of health worker preferences. Although health workers generally prefer to work in urban areas, a clear picture is emerging of the type of health worker that is more willing and likely to take up a remote post. Having grown up in a rural area, giving more importance to helping the poor, being lower rather than higher educated (e.g., nurse vs. doctor), and possibly being young rather than middle aged (little is known about elderly), all play a role. Polices that target these workers may therefore be more effective. Evidence from Indonesia shows that much can be gained from stimulating nurses to take up rural jobs, rather than doctors, who are considerably less likely to work in rural areas. Not surprisingly, these types of policies are becoming more common. Thailand, for instance, focuses on recruiting health workers with a rural background. This can further be combined with more rural-oriented training and education, as is the case in a number of high-income countries, including the US and Norway, who have built specific institutions to provide training for rural health care.

Where increasing the number of health workers is appropriate, a detailed cost–benefit analysis is needed to assess what would be the best approach. In many cases, it may be more cost effective to increase health workers in existing facilities, rather than building new facilities. Indeed, one recent study illustrates how patients can bypass facilities to get to the ones with better services. A higher number of health workers per facility also increases monitoring and reduces professional isolation. Other evidence has also shown how increasing the number of health workers per facility improves outcomes.

Besides focusing on the number of health workers, one alternative solution is to improve the quality of care. Improved training is often raised as an important way forward. There are clear examples where the curriculum does not capture local disease realities, particularly the disease burden of the poor and those in rural areas, leaving ample room for improvement. Mozambique (nonphysician) surgeons for instance, until recently received no training in HIV/AIDS, even though it was the most common disease they treated. Over the past years, many countries have updated their curriculums, with the Malawi approach that tailors content of the training to meeting the community's most pressing needs, often as a model. Recent examples are South Africa's Walter Sisulu University, community-based programs at Jimma University in Ethiopia and University of Gezira in Sudan; as well as initiatives at Makere University in Uganda and the National University of Rwanda, who are making epidemiology-based curriculum revisions. Recent evidence indicates that, on the

whole, health workers know what to do; the failure lies in doing it. The role of training to improve this, for example, by ameliorating attitudes or shifting norms, remains unexplored.

An issue that is receiving increased attention both for quantity of human resources and quality of care is the role of intrinsic and altruistic motivations. A 2008 study finds that health workers in Ethiopia who attended a Catholic NGO school are more willing to work in a rural area. Similarly, evidence for Tanzania shows that motivations matter for performance. Neither study, however, can distinguish whether this is an issue of selection or socialization. Are health workers' motivations set at the time they enter the profession, or does their training and professional environment shape their motivation? In the latter case, there would be a role for training institutions shaping motivations to improve the quality of care (and possibly willingness to work in rural areas).

Demand-Side Policies

Policies focusing on the demand side try to attract more of the existing work force to and improve health care in rural facilities, taking labor supply as given. Like many other policies, human resources have seen a shift from a manpower and central planning approach to a market-based approach. In most countries, compulsory placement has been abolished. Demand-side policies that focus on human resources quantities usually start – implicitly or explicitly – from compensating differentials theory, which argues that undesired job attributes need to be offset by attractive job attributes. Although individual preferences play a role for the precise level that is considered acceptable for a job attribute, there seems agreement as to whether an attribute is desirable or not and a pretty clear picture is emerging as to what health workers want in their job. The studies reviewed in the section Causes of Imbalances in Health Care: Theory and Evidence indicate that raises in rural salaries do increase health workers' willingness to work in rural areas, but also that increasing salaries is not enough. In most low-income countries, the discrepancy in amenities between rural and urban areas may be too large to be compensated by salaries alone. Moreover, government budgets are tight and policy makers are nervous about creating precedents for salary increases among public servants, especially in highly unionized environments. Providing other benefits like housing, transportation, and especially access to training and promotion may bring some relief. Giving more certainty about future career opportunities might also help, as concerns about unsteady future postings, together with a fear for professional isolation and lack of access to training seem to prohibit health workers from planning their career and makes rural postings less attractive. It also seems to affect job satisfaction.

Alternative approaches on the demand side are to improve efficiency and to maximize the effort and quality of care provided by existing personnel. This may be done by adapting the type of contracts offered. Although evidence for the health sector remains scarce, it has been indicated that tying pay to performance can have major impacts. Studies outside the health sector also provide evidence. Qualitative research in Ghana, which implemented a pay for performance scheme, suggests that concerns that performance pay erodes intrinsic motivation and attract the 'wrong type' of health workers seem unwarranted. A deeper concern, namely, that linking individual pay and performance may skew health worker behavior along the dimensions by which performance is measured (which is imperfect and may be arbitrary) remains largely unaddressed. Recent approaches, like the one in Rwanda, try to address this by evaluating performance at the facility level and letting it determine the budget allocated to each facility (which the facility was free to use however it wanted).

Alternative changes in contract consist of increasing monitoring and accountability of health workers, both of which tend to be weak in remote areas. This can have two effects. First, it may affect the amount of effort and quality of care delivered. The Uganda research discussed in the section Performance of Health Workers shows how improved accountability and community monitoring can ameliorate quantity and quality of care. A second concern is whether current conditions induce adverse selection, attracting health workers with undesirable attitudes, for instance the less skilled, into rural posts. This has been investigated using test scores as an indicator for the potential quality of care, but no evidence has been found that nurses and doctors with lower test scores self-select into rural jobs. It may, of course, still be possible that health workers who are less willing to apply their skills self-select into rural posts.

Matching Health Workers and Jobs: Allocation Schemes

In most countries, the allocation of health workers to jobs happens on a voluntary basis, with health workers choosing freely what job to take. But alternative allocation mechanisms exist. One example is the use of a draft or lottery. The use of lotteries in public employment has mostly been abolished (although is still present in military draft), but remains operational in some countries, including Ethiopia where, until recently, a national lottery was used to allocate health workers to jobs. Although participation in the lottery was initially compulsory, this could no longer be enforced and an opt-out has been allowed since the early-2000s (though there is still an expectation to work for a fixed period in the public sector). Allocation by lottery has been shown to be inefficient, resulting in adverse selection with the best personnel opting out of the lottery.

Other types of compulsory placement, even if limited in time, often suffer from similar problems. An example is provided by 'bonding schemes,' where health workers are expected to work in a remote area for a fixed number of years, for example, 2 years, often as a way to repay their studies. Although most countries have moved away from coercive schemes, bonding schemes remain popular, and are usually organized by state or by private institutions, often religious organizations. They suffer from similar risks as other coercive schemes, including adverse selection, erosion of motivation, and low performance. Bonding approaches have been tried and tested but have not led to the success that was hoped for. This probably explains why most countries have moved away from this, or if not, have moved to a long-term contract where

compulsory rural service is limited in time and compensated by access to additional training.

Policies focusing on matching would benefit from a deeper understanding of health workers' job decision process in developing countries. A US-based analysis provides an excellent example. Having observed that the majority of health students in the US base their job choice on their internship experience, taking their first job at the facility where they did their internship, researchers developed a two-sided matching model to optimize the allocation. Although this approach may be technically demanding, much can be learned from this type of designed matching mechanism.

Combining Public and Private Sectors

The ultimate objective of health policies is to improve people's health outcomes. A growing literature highlights the complementary role of public, private not-for-profit and private for-profit sectors to reach this objective. Studies on rural–urban imbalances in human resources often abstract from private sector activity. Perhaps the inclination of policy makers to emphasize health-care delivery through the public sector plays a role. Another reason may be that private, for-profit facilities are mostly absent in rural areas. However, not-for-profit organizations tend to be active, and, in many settings, concentrate on rural areas. The design of human resource policies that give more weight to health worker choice also require taking private sector activity into account more explicitly. Making the private sector part of the analysis also encourages bold and creative thinking about new ways to bring private health care to rural areas, moving beyond the dichotomous view that public sector's main task is to correct the imbalances caused by the private sector (or lack thereof). Letting pharmacies and especially drugstores play a more important role is one example of how public–private cooperation can contribute. More analysis is needed that compares across sectors. Existing evidence for Tanzania shows that doctors in the private sector perform considerably better than their colleagues in the public sector. Health workers in the public and not-for-profit sectors had similar levels of knowledge, but the know–do gap was smaller among NGO workers. The know–do gap is found to be largest among public sector health workers, followed by private, for-profit professionals. Health professionals in the nonprofit sector in Ethiopia are found to be less skilled, but more motivated. These are some of the exciting findings from the scarce existing evidence. Future work in this area will generate new insights on strengths and weaknesses of the different sectors and lay the ground for creatively combined or complementary approaches. Developing a finer typology to move beyond the simple categorization of rural–urban could also bring this work forward, as it will show more clearly where private, for-profit sector activity is viable.

Encourage and Support Self-Help among Rural Populations

Geographical imbalances in health care occur in all countries, regardless of whether they are low, middle, or high-income. This in itself suggests that they may be hard to solve. Moreover, among high-income countries, both those with more regulated labor markets (cf. Norway, France) as well as with weakly regulated labor markets (see section Encourage and Support Self-Help among Rural Populations) have imbalances, indicating that regulation of health worker labor markets might have limited impact. Policies have therefore typically focused on minimizing these imbalances, rather than eliminating them. The way forward seems to concentrate more on health outcomes, and make rural health care less dependent on the physical presence of health workers. The training of community health workers has been a tried model as a way to increase self-help by rural populations. Although there is no structured evaluation of these types of programs, existing overviews indicate that this is not the panacea it was once believed to be. Whereas the involvement of community health workers may help address needs for health care, including for infants and children, its scope remains limited. Past experience has also taught that there are many pitfalls for the implementation of such programs. A central concern is whether and how good quality of care can be guaranteed. Careful selection and training seem to be crucial. Another key for success seems to be whether the program is embedded not only in the community, but also in the health system. Initiatives that are implemented in parallel to the health system seem to be the least successful; whereas those integrated into the existing health system seem more effective. The Brazilian Family Health Program provides the largest and best known example of this approach. Although involving community health workers has potential, more structured evaluations are needed to increase the understanding of what works and why.

The renewed interest in community health work seems to lead to a new generation of community health work interventions. At least two potentially promising directions emerge. First, existing work suggests that community health care is more effective when built on existing institutions. An example is provided by a community-based approach grafted onto an existing network of women self-help groups making a substantial difference to maternal and infant health. Second, so far, little attention has been paid to how new technologies may further strengthen this approach. An impressive hands-on example is provided by the CARE foundation in India, where village health workers are equipped with a basic mini computer that can perform some basic tests, but also contains software with algorithms to support diagnosis and treatment. Moreover, the computer is connected via a mobile network to a doctor who can be consulted remotely by the village health worker; the doctor also monitors – and, if needed, intervenes – remotely. Although further work is needed to evaluate and explore these approaches, they open up promising avenues for future health care in rural areas.

How to Choose the Appropriate Approach?

None of the above approaches is a silver bullet, and in most cases the best way forward lies in a smart combination of approaches adapted to the local context and informed by past experience. There is currently limited understanding of the relative payoffs of these approaches to inform and guide tradeoffs between them. Budget constraints, often seen as a nuisance, may help focus minds and identify where returns are

highest. This question seems particularly relevant in light of the rapid urbanization in developing countries. How to balance the strive for equal access to quality health care with the concern of investing in geographical areas that may soon contain even less people?

Like in other areas of policy making, there is no 'one size fits all' approach. Increasing overall numbers of health workers may, for instance, sound attractive where there is a general shortage and existing capacity is fully utilized, but it remains unclear whether it will improve rural health care (see discussion in the section Implications of Imbalances for Health Outcomes). And even if it does, an equally hard question is whether this the most effective way to improve health outcomes. Would the same funding bring about more changes when used in another way? The key question remains where government expenditures – and aid – are best spent.

A useful illustration of how one can make *ex ante* tradeoffs is provided by a study focusing on Ethiopia in the early-2000s. It describes two potential ways to increase service delivery in rural areas: building more health clinics or improving and extending the quality of health care in existing facilities. Using a simple model and applying it to household data on health-care usage in Ethiopia, the study argues that additional expenditures to improve the quality of care will most likely be more cost effective than building more clinics. The conclusion sits well with earlier reported results, which show that patients bypass ill-performing facilities, and also provides deeper meaning to the results on the Ghana study mentioned earlier. The strength of this approach seems twofold. First, by providing a simple model, one can test *ex ante* what would be the most effective approach. Second, designing a simple model helps to generate well-defined hypotheses that can be tested empirically and can also help select the best empirical strategy to address identification challenges (e.g., randomized control trials (RCT)).

Summary and Conclusion

This article discusses the commonly observed discrepancies in the quantities of health workers and the qualities of health care between rural and urban areas in developing countries. The key question is how to close the gap in order to improve rural health outcomes. There is little doubt that human resources matter. Studies providing causal evidence are scarce, but they confirm the importance of human resources, which affect both the quality of care and several health outcomes. This has often been interpreted as evidence that important health gains can be made by increasing the quantity of professional health personnel in rural areas. However, the understanding of the optimal number of health workers remains limited. Although a focus on numbers and shortages may be warranted in some situations, it is by no means the silver bullet it is often claimed to be. One reason is that one also observes substantial underutilization of existing human resources, both in urban and rural areas. A small but increasing number of studies have shown a real a gap between the knowledge and the practice of health workers. Quality of care thus emerges as a real concern and deserves more attention. Future work will clarify whether quantity or quality is a more important binding constraint, and under what conditions.

One key observation shows the limitations of a single focus on increasing numbers of health workers: health professionals prefer to work in urban areas. Although studies indicate that it is possible to attract more health workers to rural areas, exploiting health worker heterogeneity in preferences, and making use of an appropriate mixture of supply, demand, and matching policies, the omnipresence of these imbalances in rich as well as poor countries suggests it is very unlikely that the gap between rural and urban areas can be closed. More creative approaches are needed. One way forward may lie in combining the different sectors – public, private, not-for-profit, and for-profit sectors, whose complentarity has been studied, but deserves more attention. Another way forward lies in the next generation of community health worker programs which are grafted on existing institutions, as well as applying new technologies. Undoubtedly, future work will pay more attention to comparing the cost effectiveness of different approaches. Here, RCT can help in building a better understanding, provided they are informed by theory and designed to reveal why some approaches work better than others.

See also: Equality of Opportunity in Health. Health Labor Markets in Developing Countries. Resource Allocation Funding Formulae, Efficiency of

Further Reading

Anand, S., Fan, V., Zhang, J., et al. (2008). China's human resources for health: Quantity, quality, and distribution. *Lancet* **372**, 1774–1781.

Banerjee, A., Deaton, A. and Duflo, E. (2004). Health care delivery in rural Rajasthan. *Economic and Political Weekly* **39**(9), 944–949.

Barber, S. L., Gertler, P. J. and Harimurti, P. (2007). The contribution of human resources to quality of care in Indonesia. *Health Affairs* **26**(2), w367–w379.

Basinga, P., Gertler, P. J., Binagwaho, A., et al. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: An impact evaluation. *Lancet* **377**(9775), 1421–1428.

Bjorkman, M. and Svensson, J. (2009). Power to the people: Evidence from a randomized experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics* **124**(2), 735–769.

Brock, J. M., Leonard, K., Masatu, M. C. and Serneels, P. (2012). Health worker performance. In Soucat, A., Scheffler, R. and Ghebreyesus, T. A. (eds.) *The labor market for health workers in Africa. A New Look at the Crisis*. Washington, DC: World Bank.

Collier, P., Dercon, S. and Mackinnon, J. (2002). Density versus quality in health care provision: Using household data to make budgetary choices in Ethiopia. *World Bank Economic Review* **16**(3), 425–448.

Das, J. and Gertler, P. J. (2007). Variations in practice quality in five low-income countries: A conceptual overview. *Health Affairs (Milwood)* **26**, w296–w309.

Das, J., Hammer, J. and Leonard, K. (2008). The quality of medical advice in low-income countries. *Journal of Economic Perspective* **22**(2), 93–114.

Dussault, G. and Franceschini, M. C. (2006). Not enough there, too many here: Understanding geographical imbalances in the distribution of the health workforce. *Human Resources of Health* **4**, 12, doi:10.1186/1478-4491-4-12.

Hanson, K. and Jack, W. (2010). Incentives could induce Ethiopian doctors and nurses to work in rural settings. *Health Affairs (Millwood)* **29**(8), 1452–1460.

Lavy, V. and Germain, J. (1995). Tradeoffs in cost, quality and accessibility in the utilization of health facilities: Insights from Ghana. In Shaw, R. P. and Ainsworth, M. (eds.) *Financing health services through user fees and insurance: Lessons from sub-Saharan Africa*. pp. 134–153. Washington, DC: The World Bank Publisher.

Lehmann, U., Dieleman, M. and Martineau, T. (2008). Staffing remote rural areas in middle- and low-income countries: A literature review of attraction and retention. *BMC Health Services Research* **2008**(8), 19, doi:10.1186/1472-6963-8-19.

Munga, M., Songstad, N. G., Blystad, A. and Mæstad, O. (2009). The decentralisation-centralisation dilemma: Recruitment and distribution of health workers in remote districts of Tanzania. *BMC International Health and Human Rights* **9**(9), doi:10.1186/1472-698X-9-9.

Serneels, P., Montalvo, J. G., Pettersson, G., et al. (2010). Who wants to work in a rural health post? The role of intrinsic motivation, rural background and faith based institutions in Rwanda and Ethiopia. *Bulletin of the World Health Organization* **88**, 342–349.

Serneels, P., Lindelow, M., Montalvo, J. G. and Barr, A. (2007). For public service or money: Understanding geographical imbalances in the health workforce. *Health Policy and Planning* **22**, 128–138.

Soucat, A., Scheffler, R. and Ghebreyesus T. A. (eds.) (2013) The labor market for health workers in Africa. A New Look at the Crisis. Washington, DC: World Bank.

Sousa, A., Tandon, A., Dal Poz, M. R., Prasad, A. and Evans, D. B. (2006). Measuring the efficiency of human resources for health for attaining health outcomes across subsnational unit in Brazil. *Back Ground Paper for World Health Report*. Geneva: WHO.

Speybroeck, N., Kinfu, Y., Dal Poz, M. R. and Evans, D. B. (2006). Reassessing the relationship between human resources for health, intervention coverage and health outcomes. *Evidence and Information for Policy*. Geneva: World Health Organization.

International E-Health and National Health Care Systems

M Martínez Álvarez, London School of Hygiene and Tropical Medicine, London, UK

Glossary

Bilateral A relationship, such as a trading relationship, involving two partners, such as countries.
e-health The application of information and communication technologies across a range of health care services.
General Agreement on Trade in Services (GATS) An outcome of the 1995 Uruguay Round Negotiations and the basis of the global multilateral sector trading system.
Multilateral A relationship, such as a trading relationship, involving many partners, such as countries, trading with many others.
Teleconsultation A medical consultation that takes place when the patient and doctor are not in the same physical location.
Telemedicine The use of information and communication technologies to deliver clinical health care services at a distance.
World Trade Organization The global institution that deals with the rules of trade between countries.

Introduction

With increasing globalization, countries have opened up their borders to trade in goods and services, often including health services. This has given rise to heated debates in the media and the academic and professional literature, with proponents arguing that it can improve efficiency and facilitate the sharing of ideas, although opponents argue that international trade in health services will result in increased privatization and hinder domestic decision making. In reality, lack of data makes it very difficult to ascertain the volume of trade in health services and the effect it is having on health systems.

There are different ways in which health services can be traded internationally, involving either patients or health professionals traveling to another country to obtain/provide health services, countries investing in other countries' health services, and through the remote provision of health services. This article is concerned with the latter form of trade, the remote cross-border provision of health services, also known as international e-health, and its impact on the national health system of the countries involved in it.

The article reviews both the positive and negative contributions that international trade in e-health services may offer to national health systems. In doing this, it will briefly comment on the different types of trade relationships the countries may engage in, and in turn, how this can affect the impact international e-health has on their health systems.

This article is structured as follows. First, it defines e-health and outlines examples of its different uses. This is followed by an account of how national health systems of countries engaging in international e-health (both as exporters and importers) are affected by it e-health, before outlining the different types of trade relationship e-health can be traded under. The article concludes with key messages.

What Is E-Health?

E-health can be defined as the application of information and communication technologies across the whole range of health care services. Given that the scope of this article is on international e-health, it will be defined as the use of information and communication technologies to deliver health services across an international border.

Although traditional communication technologies can be used to deliver health services remotely – for instance, by using the postal service to send samples to be analyzed in remote laboratories – the term e-health is concerned with the use of nontraditional information and communication technologies. As it will be seen, most e-health services take place through the use of the Internet.

Table 1 shows the different uses of e-health, which can be clinical and nonclinical. Nonclinical health services include medical transcription, where doctors record their notes and these are transcribed remotely, often overnight, and electronic patient records. However, the most potential use for this type of trade in health services lies within the provision of clinical services. This is known as telemedicine. Telemedicine can be divided into different subsets, depending on the type of care that is provided, as shown in Table 1.

An important use of cross-border telemedicine is the remote provision of diagnostic services. The most popular of these has so far been teleradiology, where images, such as X-rays, are transferred electronically to radiologists remotely for interpretation. Teleradiology is often done across different time zones, which allows for images to be processed overnight, a process known as 'nighthawking.' Similarly, telepathology involves sending images of processed samples (such as microscope images) for interpretation.

A final (and emerging) use of telemedicine is to provide consultations at a distance. This can be done when the experts are physically located far from the patients. This practice has given rise to specialties such as teledermatology, telepsychiatry, and teleophthalmology, and has the benefit of permitting access to expertise to patients who would not have otherwise been able to travel for it. The use of cross-border provision of surgery and emergency services has been considered as a potential area of growth in the global e-health market, but its use has not been explored as yet in any major initiatives.

Table 1 Types of e-health

Nonclinical	*Diagnostic telemedicine services*	*Teleconsultations*	*Potential telemedicine uses*
Medical transcription Patient records	Teleradiology Telepathology	Teleneurology Telepsychiatry Teledermatology Teleophthalmology Telesurgery	Emergency services

An example of the cross-border use of telesurgery is shown on the following YouTube excerpt: http://www.youtube.com/watchv=d7IojFFHtiA ('Telesurgery – "Lindbergh operation"' YouTube video, 3:40, posted by Justin Kochi, 23 June 2009).

To What Extent do Countries Engage in E-Health?

The size of the global e-health market is difficult to estimate, as there is currently no systematic collection of data on the amount of e-health trade that takes place or the revenues made from it. However, estimates in the literature indicate that it is happening on a large scale and generating significant revenues, with the global e-health market estimated to be worth between US$1 billion and US$1 trillion (Mutchnick *et al.*, 2005). This lack of reliable data poses problems for health planners, as they are not aware of how much e-health trade is taking place, and for policy makers, who then base their decisions on ideology rather than evidence.

The World Health Organization's Global Observatory for e-health conducted a global survey of e-health in 2009 to map out all e-health initiatives that are currently taking place across the globe. The results from this survey are summarized in Table 2; they include all e-health initiatives (national and international), so the true size of the international e-health market will be smaller.

Of these initiatives, teleradiology, some tertiary care, and telepathology are the areas that currently hold greatest promise for international e-health trade.

Although e-health is not bound by physical location, there are some factors that influence which countries trade with which, such as common language and data management protocols. This has resulted in a significant proportion of e-health trade taking place regionally. Examples of such regional trade initiatives are summarized in Boxes 1 and 2.

How can Countries Benefit from International E-Health?

When discussing cross-border e-health trade, countries can be divided into 'exporting' and 'importing,' depending on whether they provide or 'purchase' e-health services, respectively. Exporting countries tend to be low- and middle-income countries, which have invested in technology and can provide services for a fraction of the cost of their higher income counterparts. The top three exporters of e-health services are India, the Philippines, and Cuba. However, the importing countries tend to be high-income countries, whose health systems are facing budget restrictions and efficiency calls. The USA is the top importer of e-health services. Given that the impact e-health has on countries is dependent on whether they are importers or exporters, it will be discussed separately.

Table 2 Number of e-health initiatives reported by the global survey of e-health

Subset of telemedicine	*Number of initiatives*
Teleradiology[a]	61
Tertiary care[b]	25
Teleconsultation[c]	17
Telesurgery	15
Home care and patient monitoring	9
Telepathology[d]	7
Others	8
Total	142

[a]Ultrasonography, cardiology, scintillography, and mammography initiatives have been included as teleradiology.
[b]Diabetes, obstetrics and gynecology, oncology, pediatrics, urology, etc.
[c]Includes dentistry, ophthalmology, and otolaryngology.
[d]Includes biochemistry, cytology, hematology, hepatology, histopathology, immunology, and laboratory services.

Box 1 Case study 1: The Implementing Transnational Telemedicine Solutions project

The Implementing Transnational Telemedicine Solutions project is a European initiative started in September 2012. It aims to implement 10 demonstrator transnational telemedicine projects across Scotland, Norway, Finland, Sweden, Ireland, and Northern Ireland, including the use of video consultation, mobile self-management, and home-based health services. The key objectives of the project are to improve health service coverage for remote communities, thereby reducing hospital visits, enhancing the use of technology, and increasing and fostering transnational collaboration. The project is a pilot and will be evaluated, but it is hoped it will form a sustainable telemedicine network among northern European countries. More information on this project can be found on the following website: http://www.transnational-telemedicine.eu/

Importing countries

The most important benefit the importing countries stand to gain from outsourcing health care services to exporting countries is a financial one. This is because most exporting countries are low- and middle income, and can therefore provide health services remotely for a fraction of what they would cost in the importing country, mainly due to the fact that the health professionals' salaries can be up to 10 times lower. This is particularly relevant in the current financial situation, where many importing countries are facing budgetary restrictions and are looking to make their provision of health services more efficient.

Box 2 Case study 2: India

India is one of the world's key players in e-health. It has currently more than 400 e-health platforms, made up of both public and private actors. Although the main focus of its e-health industry is the domestic population, India is also viewed as an important provider of international e-health services. Some of the more headline-grabbing examples of India's international e-health concentrate on the provision of services to the US, often through 'nighthawking' (see http://articles.economictimes.indiatimes.com/2006-08-03/news/27444693_1_hospitals-tele-radiology-teleradiology-solutions as an example); however, India's key international e-health market is made up of its neighboring countries and Africa. As such, the Government of India, through its Ministry of External Affairs, has launched two initiatives: the SAARC Telemedicine Network and the Pan-African e-Network Project.

The SAARC Telemedicine Network is implemented by the Telecommunication Consultant India (Ltd.), and it links one or two hospitals in each of the countries forming the SAARC region with up to four superspecialty hospitals in India. The initiative involves the provision of teleeducation and teleconsultations.

The Pan-African e-Network Project's objective is also to provide education and telemedicine, through teleconsultations, from India's superspecialty hospitals (10 are involved in this initiative). The aim is to provide services to 53 African countries. At the time of the launch of the second phase in August 2010, 47 African countries had already joined the project. More information on this initiative can be found on the project's website: http://www.panafricanenetwork.com/

The second means by which the importing countries can benefit from outsourcing health care services internationally is by decreasing the waiting time. The health systems of many high-income countries suffer from long waiting lists, particularly for elective procedures. By outsourcing some of their health services, such as diagnostics, the importing countries can significantly reduce waiting lists. In addition, the fact that the importing and exporting countries are often situated on different time zones allows for services to be carried out overnight, greatly improving the efficiency of the health care system in the importing countries. Furthermore, due to the importance of early diagnosis in certain conditions such as cancer, patients can be diagnosed and started on treatment sooner, which will lead to improved prognosis and lower costs.

A further advantage of engaging in e-health trade facing the importing countries is the improvement in coverage of remote areas. Remote populations are very expensive to serve and often hard to access. Therefore, providing the services remotely would greatly reduce costs and improve the quality of health care coverage of remote populations.

Finally, outsourcing of routine diagnostic and curative services to the exporting countries can reduce the workload of health care professionals in the importing country and allow them to concentrate on the more complicated cases and therefore, improve specialization and skill set in the country.

Exporting countries

Exporting countries can also benefit greatly from engaging in international provision of e-health services. Similar to the importing countries, the key benefit is a financial one, as they can generate foreign income. As highlighted earlier, the e-health market is of substantial size; although no official figures are available, it is estimated that the telemedicine market holds a huge potential for the importing countries. For instance, in India, it is estimated to be worth €37.4 million, with projections to reach €374 million (Financial Express. Telemedicine: An answer to ailing India. 5 November 2007; http://www.financialexpress.com/news/telemedicine-an-answer-to-ailing-india/236263/0). It can therefore particularly benefit the exporting country's health system if it is invested back in it. This is of particular importance given that the exporting countries are typically low- and middle income and often have underfunded health systems.

Exporting countries can also benefit from providing e-health services by reversing their 'brain drain.' The brain drain is a phenomenon caused by health professionals migrating in the pursuit of higher salaries, improved quality of life and career prospects. It particularly affects the low- and middle-income countries, which suffer severe shortages in human resources for health. It is also some of these countries that have started exporting health services, such as e-health, and can therefore take advantage of the higher salaries and career opportunities the e-health posts offer to attract some of these workers back to the country and thereby increase their human resource base in the health sector.

To be able to provide e-health services to the importing countries, the exporting countries need to remain competitive and meet international standards. They therefore often make significant investments in technology and on improving the available skill set of their health workforce. This will also benefit the local population as the technology and health professionals available will unlikely devote all of their time to providing e-health services to other countries, and can then be used to provide domestic services, thereby providing the opportunity of using the international market to subsidize their domestic services. In fact, some of the key exporters of e-health services, such as India, have important domestic e-health services, with considerable potential for expansion.

What do Countries Risk by Engaging in E-Health?

The section Exporting countries has highlighted the great potential that both importing and exporting countries have for benefiting from international e-health trade. Next, the risks these countries face when entering this type of trade in health services are discussed.

Importing countries

The key risk the importing countries face when engaging in e-health trade is data security and privacy. Data sent over to the exporting countries are extremely sensitive in nature as they include health records, and there must therefore be absolute guarantee that confidentiality will be preserved. In fact this tends to be the main barrier to engaging in this type of trade, with countries only trading with those who have similar or trusted data management protocols.

In addition, the importing countries also face the risk that the quality of the services provided by the exporting countries would be lower than that they themselves can offer. This can be further compounded by language and cultural differences,

as well as the different training the health professionals receive in different countries, which hinder the ability of health professionals to communicate with each other and the patients and agree on a course of action. A related concern is liability: Who is responsible if something goes wrong? If countries engage in e-health trade, malpractice will eventually occur, and when it does, it is not clear whose responsibility it would be. There are concerns that the importing countries would face expensive lawsuits, which would offset any savings made from e-health trade.

Finally, the importing countries risk job losses if some health services are performed in other countries. Furthermore, whereas allowing health professionals in the importing countries to specialize and concentrate on complicated cases is clearly an advantage, if all the uncomplicated cases are dealt with abroad, this may hamper the ability of new health professionals to be trained as they will not be exposed to them.

Exporting countries

Exporting countries also face some risks when providing international e-health services. Given the revenues to be made by providing e-health services to other countries, there is a risk that resources will be diversified toward this, at the cost of health services the domestic population needs. This may worsen rather than improve the national health system. In addition, if e-health services are provided through the private sector (as is often the case), the revenues generated may not be invested back into the health system.

Another risk the exporting countries face is the creation of an internal brain drain. The higher salaries and career opportunities offered by international e-health may not just attract health workers who had migrated, but also health workers currently employed by the public health system. This may actually exacerbate rather than ameliorate shortages in health professionals and again, worsen the domestic health system.

Trade Agreements

It is important to note that the potential risks and benefits countries face when engaging in e-health international trade outlined in this article are influenced by the type of trade relationship they engage in. There are three types of trade relationships countries can engage in: multilateral, regional, and bilateral. This section briefly summarizes each type of trade relationship and highlights how they can each influence the extent to which national health systems are affected by international e-health.

Currently, most e-health takes place under a multilateral system, where many countries trade with each other. This takes place under the General Agreement on Trade in Services (GATS), under the auspices of the World Trade Organization. The GATS categorizes services into four modes, which can all be applied to health services. Mode one covers the cross-border provision of services, which in the case of health would be e-health. Mode two involves consumption of services abroad (in the case of health medical tourism). Modes three and four deal with foreign direct investment (for instance, in a hospital) and the movement of natural persons (health care professionals), respectively. Under this form of trade agreement, countries can freely trade with others. The benefits and concerns described above mainly apply to the current system of multilateral trade, where it is more difficult to implement safe guards on data safety and quality of care and countries may find it difficult to define litigation procedures.

Cross-border e-health trade can also take place regionally. In fact, this seems to be the case in many instances. Countries are more likely to import health services from countries that have similar language, culture, and training standards. This has led to the development of different regional e-health initiatives, such as the Implementing Transnational Telemedicine Solutions project, the South Asian Association for Regional Cooperation (SAARC) Telemedicine Network, and the Pan-African e-Network Project initiatives described in Case studies 1 and 2.

The final type of trade relationship countries may engage in when importing/exporting e-health services is a bilateral one. This would take place between two countries, an exporter and importer, where a contract would be drawn between the two outlining conditions under which trade will take place. The benefits outlined above would still apply to this type of relationship. However, there is potential to capitalize on them, for instance, by stating clearly in the contract what proportion of the revenues has to be invested back into the health care of the domestic population of the exporting country. Furthermore, some of the risks can be averted or reduced. For instance, the contract can state what data management protocols will be used, the minimum-required qualifications of the providers, and a program for the exchange or training of human resources to alleviate shortages in the exporting country. Despite these apparent benefits, bilateral relationships in e-health (and health services more generally) tend to be underresearched and underutilized.

Conclusion

This article has covered the definition and different uses of e-health before outlining how countries – and their health systems – stand to gain or risk losing from engaging in this type of trade. The article then briefly reviewed the different types of trade relationships and how these can affect the impact international e-health trade has on both the importing and exporting countries. It is important to emphasize the dearth of data on e-health trade (and trade in health services in general), which makes it difficult for the health planners to plan their services, and base their decisions on ideology rather than evidence. Notwithstanding this, countries considering whether to engage in international e-health should consider bilateral initiatives, as these offer the possibility of controlling some of the risks, while still reaping the benefits from this type of trade.

See also: International Movement of Capital in Health Services. International Trade in Health Services and Health Impacts. International Trade in Health Workers. Medical Tourism. Pharmaceuticals and National Health Systems

Reference

Mutchnick, I. S., Stern, D. T. and Moyer, A. (2005). Trading health services across borders: GATS, markets and caveats. *Health Affairs* **W5**, 42–51.

Further Reading

Blouin, C., Drager, N. and Smith, R. D. (2005). International trade in health services and the GATS: Current issues and debates. World bank, Washington, D.C.

Chanda, R. (2002). Trade in health services. *Bulletin of the World Health Organization* **80**(**2**), 158–163.

Gerber, T., Olazabal, V., Brown, K. and Pablos-Mendez, A. (2010). An agenda for action on global e-health. *Health Affairs* **29**(2), 233–236.

Khan, H. A., Qurashi, M. M. and Hayee, I. (2008). *Better healthcare through telehealth*. Commission on Science and Technology for Sustainable Development in the South. Islamabad: New United Printers.

Lougheed, T. (2004). Radiologists get that long distance feeling. *Canadian Medical Association Journal* **170**, 1523.

Mars, M. and Scott, R. E. (2010). Global e-health policy: A work in progress. *Health Affairs* **29**(2), 237–243.

Martínez Álvarez, M., Chanda, R. and Smith, R. D. (2011). How is telemedicine perceived? A qualitative study of perspectives from the UK and India. *Globalization and Health* **7**, 17–24.

McLean, T. R. (2006). The future of telemedicine & its Faustian reliance on regulatory trade barriers for protection. *Health Matrix* **16**(2), 443–509.

Scott, R. E. (2009). Global e-health policy: From concept to strategy. In Wootton, R., Patil, N. G., Scott, R. E. and Ho, K. (eds.) *Telehealth in the developing world*, pp. 55–67. Ottawa: International Development Research Centre (IDRC).

WHO (2010) Telemedicine: Opportunities and developments in member states: Report on the second global survey on eHealth 2009 (Global Observatory for eHealth Series, Volume 2). Available at http://www.who.int/goe/publications/goe_telemedicine_2010.pdf (accessed 10.04.13).

International Movement of Capital in Health Services

R Chanda and A Bhattacharjee, Indian Institute of Management Bangalore, Karnataka, India

Introduction

There has been considerable debate in recent years regarding globalization of health services and its implications for exporting and importing economies. This debate has been sparked by the growing scope for cross border delivery of health services due to advances in information and communication technology, growing mobility of healthcare providers and patients, and commercialization of health services through foreign direct investment (FDI) and entry of domestic private players. Today, trade in health services takes place through telemedicine, medical value travel, cross border flows of healthcare workers, international capital flows, and transnational corporations in the health sector. There are also emerging opportunities in information technology (IT)-enabled delivery of health-related services, such as medical coding, transcriptions, and back-office health support services.

In light of growing healthcare challenges confronting governments worldwide due to rising healthcare costs and strained public sector budgets, aging societies, and growing demand-supply gaps in healthcare, globalization of health services is a potentially important means of providing quality healthcare, of ensuring financial sustainability of health systems, and enabling equitable access. However, given the scant and often anecdotal nature of information on trade in health services and lack of primary evidence or case studies, it is difficult to understand the trends and characteristics in any detailed manner or to draw any concrete conclusions regarding the associated risks and challenges. Hence, the debate on the implications of globalization of health services remains polarized, mostly based on conjectures and opinions rather than factual and empirical analysis. One side stresses the potential to benefit from increased foreign exchange earnings with positive implications for the domestic health systems and another side voices concerns regarding the potential adverse effects on equity and access to healthcare.

The discussion in this article focuses on one form of globalization of health services, namely, international capital flows and foreign commercial presence in the provision of health services. It consolidates the scattered, secondary information that is available on such flows in order to outline the broad trends and characteristics of this mode of health services delivery and highlights the perceived, and where available, realized impact of such flows.

The Section on Overview of Trends provides an overview of trends in foreign financing of healthcare services, highlighting key source and recipient countries as well as major firms providing health services across borders through overseas commercial presence. It also discusses the nature of such capital flows. The Section on Policies Governing Foreign Investment in Health Services highlights the regulatory environment affecting foreign investment in health services for a sample of countries. It also compares national policies with multilateral commitments made by selected countries on foreign commercial presence (mode 3) in health services under the General Agreement on Trade in Services (GATS) (the other three modes of supply in health services, under the GATS are: Mode 1 (cross border supply)). The discussion highlights the regulatory and other concerns characterizing the liberalization of health services. The Section on the Impact of Foreign Investment in Health Services discusses the benefits and challenges associated with the movement of capital in health services, drawing on existing studies and discussions with healthcare providers. Given various classification issues and interdependencies between trade in health services and other related services such as insurance, education, and IT-enabled services, the discussion on impact primarily focuses on capital flows in health services establishments such as hospitals, clinics, and diagnostic facilities. The Section on Concluding Thoughts concludes with the key policy inferences.

Overview of Trends

Financing of health services can come from sources within a country such as taxes, insurance funds, and private investment, or from external sources in the form of portfolio and equity investments, commercial loans, FDI, official aid, and non-governmental financing. As per the GATS, cross border capital flows are also a form of services trade captured under mode 3, which refers to the establishment of any type of business or professional enterprise in the overseas market in order to supply a service (Table 1).

The following discussion provides an overview of recent trends in international capital flows in health services, drawing upon a variety of multilateral and company-level data sources. A few points are worth highlighting. The first relates to the severe data limitations that constrain any efforts to analyze international capital flows in health services. Mode 3 data are not readily available from the Balance of Payments statistics and comprehensive data on measures of health resource flows are lacking. The Foreign Affiliates Trade in Services statistics provide this information, but are available for only a small

Table 1 Various modes of foreign investment in health services

	Forms of foreign investments in health services
Mode 3 in Health Services	Joint Ventures Technology tie-ups Acquisition of facilities Health insurance services Management contracts and licenses Medical education/training centers/research facilities Foreign participation or ownership of hospitals, clinics, medical facilities

 doi:10.1016/B978-0-12-375678-7.00616-7

group of countries. Moreover, such data are not disaggregated by activities and segments within the health services sector and there are potential overlaps with health-related ancillary services and even products. Hence, a comprehensive understanding of the magnitude and breakdown of investment flows in health services is difficult. The second issue concerns the scope of the analysis. This article takes a broad definition of commercial presence and considers any form or level of commercial involvement through greenfield investments, mergers and acquisitions, offices or subsidiaries, or any form of juridical presence as constituting mode 3. The underlying assumption is that there are associated capital flows and authorizations from the host country. Hence, the scope of the analysis is not directly aligned with the GATS definition of mode 3, and terms such as foreign investment, mode 3, and international capital flows are used interchangeably.

The Broad Picture of Capital Flows in Health Services

Although it is difficult to build a comprehensive picture of foreign investment in health services, the available data indicate that health services play only a marginal role in international capital flows in services. Inward and outward FDI flows and stocks in health services accounted for a meager 0.17% and 0.02% in total services FDI for developed economies, and for only 0.06% of the total inward stock of services FDI in developing countries, in 2005. However, the significance of health services in total services FDI has grown over time. Between 1990 and 2005, inward and outward FDI stocks grew by 762% and 380%, respectively, in developed countries. Data sources for this section include the International Trade Centre (ITC) investment map, the United Nations Conference on Trade and Developments (UNCTAD's) FDI statistics and Trans-nationality Index online databases, and the Fortune Global 500 index. (UNCTAD World Investment Reports).

Foreign Affiliates Trade in Services Statistics, which cover a variety of indicators (exports, imports, sales, turnover, and employment) regarding the activities of foreign companies in overseas markets indicate that developed countries have been the leading sources and destinations for FDI in health services. In 2000, the US was the main source as well as destination market in terms of the number of transactions and was the main recipient in terms of the value of transactions. Countries with public sector dominated health systems such as the UK and those with commercially oriented health systems such as the US have featured among the leading exporters and importers of FDI in health services (Waeger, 2007, Table 4, p. 14).

Recent data from the ITC's investment map confirm that developed countries remain the leading sources for investment

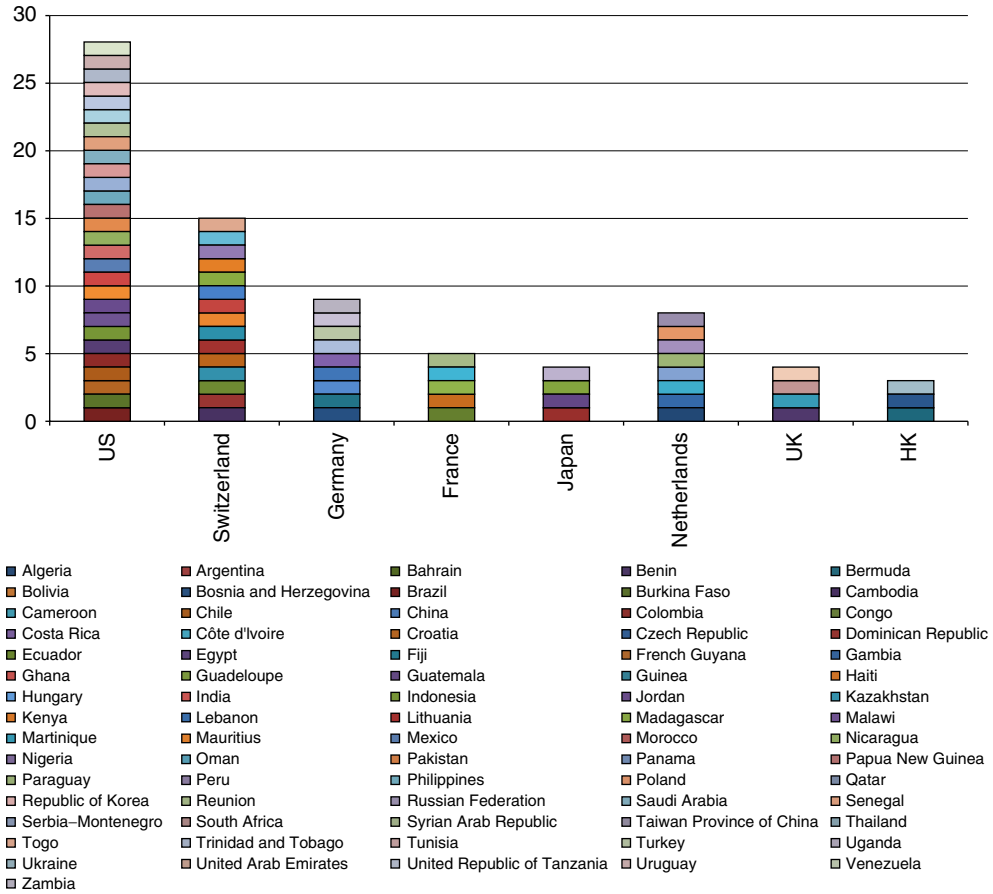


Figure 1 Top 8 leading home countries and their affiliates in host counties (health and social services). HK stands for Hong Kong (SAR China); UK for United Kingdom; US for United States. Available at: http://www.investmentmap.org/TimeSeries_Country_fdi.aspxprg=1 (accessed 24.05.11). Calculated from ITC investment maps.

in health and social activities, as measured by the number of overseas affiliates. The US is the leading investor, followed by Switzerland (Holden, 2002). However, the range of host countries for health services investment has grown considerably over the past decade. Of the 76 countries, which are host to health investments through affiliates, a large number are developing or least developed nations. Figure 1 shows the leading investor countries along with their corresponding developing country hosts for investment in health services.

ITC investment maps also provide information on FDI inflows, though this data is available for only a few countries over the 2000–10 period. The US is the leading recipient, with over US$300 million of FDI inflows (in 2010), in health and social services, significantly more than other countries. Figure 2 illustrates the FDI inflows in health services (including net sales of shares and loans to the parent company plus the parent firm's share of the affiliate's reinvested earnings plus total net intracompany loans – short- and long-term provided by the parent company) for some of the main recipient countries (excluding the US). The data show a sudden significant jump in inward FDI for some countries during this period, though it remains largely stagnant and low for many countries. It must be noted, however, that these FDI statistics pertain to health and social services. Hence, it is difficult to ascertain how much pertains to segments such as hospitals, diagnostics, and clinics directly related to healthcare provision and how much relates to health-related social services or sectors such as health insurance.

Figure 3 shows the number of foreign affiliates in the health and social services sector of the leading developing country hosts for health services FDI. The countrywise distribution indicates that the extent of foreign participation in the health services sector varies considerably across different developing countries, with Brazil, Reunion (Réunion is a French island in the Indian Ocean.), China, and Mexico hosting the largest number of such affiliates in 2009.

Transnational Activity in Health Services

Data on mergers and acquisitions in the health and social services sector show a similar upward trend in foreign commercial presence. The number of mergers and acquisitions in health and related social services reached US$14 billion in 2006 in terms of sales transactions, with an annual average value of M&A activity of US$3.9 billion during the 2004–06 period. The largest health services companies were based in and also operated in the developed countries such as the US, UK, and Canada (Cattaneo, 2009).

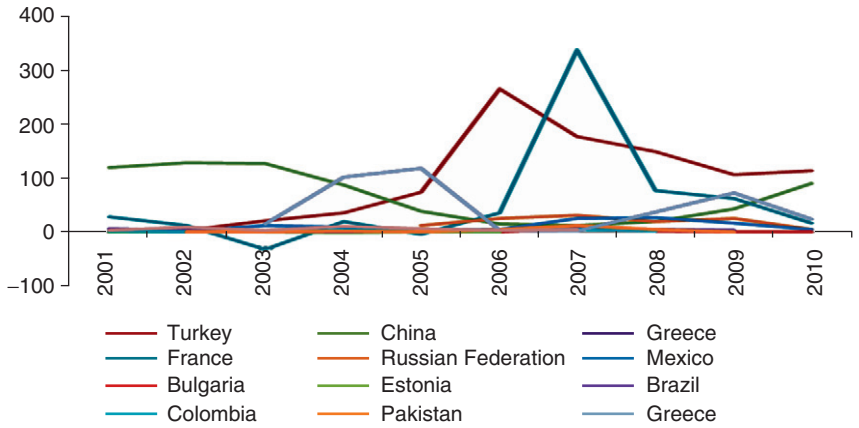


Figure 2 FDI inflows into health and social services of economies (in US$ million). Calculated from ITC investment maps for countries with data for at least 7 years. Available at: http://www.investmentmap.org/TimeSeries_Country_fdi.aspx?prg=1 (accessed 27.09.12).

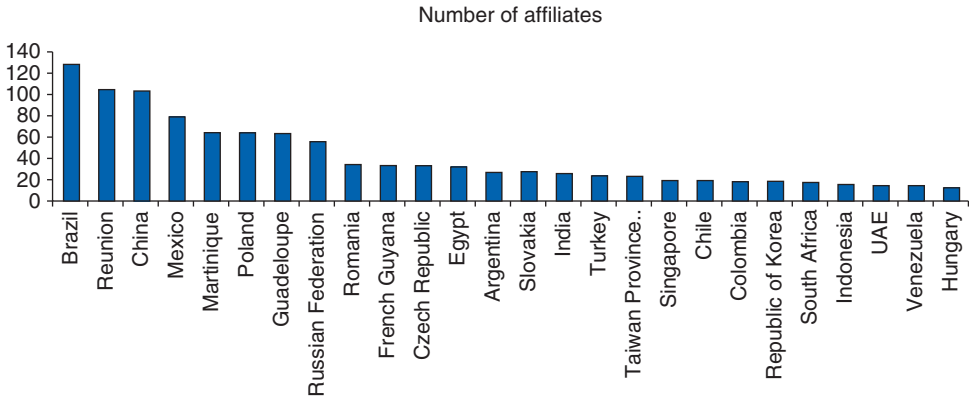


Figure 3 Number of foreign affiliates in host countries (health and social services). Calculated from ITC investment maps. Available at: http://www.investmentmap.org/TimeSeries_Country_fdi.aspx?prg=1 (accessed 24.05.11).

The growing internationalization of health services firms is also indicated by the Fortune Global 500 internationalization rankings of firms. Based on the Fortune Global 500 list for 2002, Holden (2005) had found that direct health services providers were the least internationalized whereas firms in areas like insurance and pharmaceuticals were the most prominent in internationalization rankings (Holden, 2005). Ten health service companies were listed on the Global 500 List in 205 and nine were ranked in 2006, 2007. The average ranking was 298, 262 and 245, respectively, for each of these years (http://money.cnn.com/magazines/fortune/global500/2010/index.html (accessed April 2011)).

Although the top ranked health services firms are mostly based in and also operate in developed countries (chiefly the US), M&A activity in the hospitals and clinical services segment reflects diversification of source and recipient markets. **Table 2** provides information on recent acquisitions of healthcare providers involving developing countries. It reflects the emergence of a small set of transnational hospitals and healthcare providers with both regional and global presence and the entry of several developing country health services firms based in Asia and Africa into foreign markets through M&As. It is also interesting to note the emergence of South–North and South–South flows of capital. The bilateral pattern of M&As indicates the significance of factors such as geographic and cultural proximity, regional markets, growth dynamics, and market size.

Box 1 highlights the growing regional and global presence of healthcare providers from selected developing and developed countries, and also outlines the formats and strategies adopted by these providers in overseas markets.

Entry into overseas markets is thus occurring through joint ventures, franchises, greenfield investments, acquisitions, tie-ups, contractual arrangements, and public–private partnerships. Linkages are also evident with other forms of health services trade.

Policies Governing Foreign Investment in Health Services

Increased transnational activity in direct health services reflects FDI liberalization in health services and the incentives given to private players in many developing countries. Since the 1990s, developing countries such as India, Indonesia, Thailand, Sri Lanka, Brazil, and South Africa have opened up their health service sectors to participation by foreign hospitals, diagnostic centers, and clinics. Privatization and deregulation of the healthcare sector in these countries have also contributed to the emergence of private healthcare providers who are globally or regionally competitive. Cambodia, for instance, permits cross border investment in hospital services and foreign ownership. It also permits management of private hospitals and clinics with the condition that at least one director is a national. Foreign firms are allowed to provide dental services through joint ventures with Cambodian legal entities. Similarly, Indonesia is open to foreign healthcare providers, allowing Singaporean, Australian, and Canadian firms to operate in its market. India has permitted automatic approval for 100% FDI in hospitals since 2000. Between 2000 and 2006, there were close to 100 approved FDI projects in hospitals and

Table 2 Recent acquisitions of healthcare providers involving non-OECD countries

Year	*Investor*	*Subsidiary*	*Exporting country*	*Importing country*	*Value of investment*	*Nature of investment*
2006	Netcare	General Healthcare	South Africa	The UK	GBP 2.2bn	52.6% stake
2007	Mediclinic	Emirate Healthcare Holdings	South Africa	United Arab Emirates	USD 46.4m	49% stake
2007	Mediclinic	Hirslanden	South Africa	Switzerland	USD 2.4bn	100% stake
2005	Bumrungrad International	Asian Hospitals	Thailand	Philippines		45.5% stake
2006	Bumrungrad International	Bumrungrad Hospital Dubai	Thailand	United Arab Emirates		49% stake (Joint venture with Istithmar)
2007	Bumrungrad International	Asia Renal Care	Thailand	Singapore (operates clinics in 6 Asian Countries)	USD 75m	100% stake
2005	Apollo Hospitals	Apollo Hospitals Dhaka	India	Bangladesh	USD 35m	100% stake
2005	Parkway Healthcare	Pantai Hospitals	Singapore	Malaysia	USD 139m	31% stake
2008	Siemens and Asklepios Kliniken	Sino-German Friendship Hospital	Germany	China	USD 145m	Public–private partnership with Tongji University, Shanghai

Source: Reproduced from Mortensen, J. (2008a). International Trade in Health Services – The trade and the trade-offs. *Working Paper 11*. Copenhagen: Danish Institute for International Studies, Table 8, p.26.

Box 1 International health service provider firms from developed and developing countries

Developing countries

Singapore: The Parkway Healthcare Group is the biggest investment group for healthcare in Singapore and one the largest healthcare organizations in Asia. It has created Gleneagles International as an international brand. The company has been interested in acquisition of hospitals in Singapore, building up a base, and entering countries like India, Indonesia, Malaysia, Sri Lanka, and the UK, mostly through joint ventures with local partners. It entered the Indian healthcare market in 2003 through a joint venture with the Apollo Group and built the Apollo Gleneagles Hospital, a multispecialty hospital at a cost of US$29 million (Chanda, 2007a). It has formed a joint venture with the Mumbai-based Asian Heart Institute and has established a research center to provide medical excellence. It is in the process of setting up a specialized heart hospital in London.(Source : http://portal.bsnl.in/bsnl/asp/content%20mgmt/html%20content/business/business56857.html)

The Singapore-based Raffles Medical Group is building strategic alliances through triangular business associations with healthcare organizations from developed countries and venturing into developing countries in partnership with host country investors.

Thailand: Bumrungrad Hospital in Thailand has entered into management contracts with hospitals in Bangladesh and Myanmar. It has formed a joint venture with a hospital in the Philippines. Bangkok Hospital has 12 branches in Southeast and South Asia, located mostly in tourist towns (Arunanondchai and Fink, 2007).

India:The Apollo Group of Hospitals has centers of excellence in several countries like Nepal, Sri Lanka, Ghana, and Bangladesh. It has also entered into contract-based management of hospitals or clinics in the United Arab Emirates, Oman, Kuwait, Mauritius, Malaysia, Sri Lanka, and Nigeria (Mortensen, 2008a, and http://www.thehindubusinessline.in/2005/12/03/stories/2005120303200200.htm). It has established a telemedicine center in Kazakhstan. Apollo Hospitals has entered into a joint venture with Amcare Labs, an affiliate of Johns Hopkins International, to set up a diagnostic laboratory in Hyderabad.

South Africa: South African health services firms are present in the UK, Switzerland and the United Arab Emirates, and are also the main source of regional FDI in Southern Africa (Mortensen, 2008a). Some major firms include Netcare, Mediclinic, Life Healthcare, and the Afrox Healthcare Group. Mediclinic owns private hospitals in Namibia; Life Healthcare operates private hospitals and clinics in Botswana; and the Afrox Healthcare Group has operations in Botswana, Namibia, Zambia, and Mozambique. Netcare has a public–private partnership with the Lesotho government to build a hospital and refurbish two feeder clinics and run clinical services for the government.

Developed countries

US corporations are major players in the hospital sector. Hospital corporations own the for-profit hospitals that they operate. In small specialized clinics like eye clinics, rehabilitation centeres, and outpatient clinics, US firms enter through joint ventures with local specialist doctors or surgeons.

Columbia Asia Group, a Seattle-based hospital services company, a worldwide developer and operator of community hospitals, has started its first American-style medical center in Bangalore.

The Fresenius Medical Care group (FMS) is headquartered in Germany and is one of the leading foreign healthcare providers in the US. FMS has operations Belgium, France, Italy, the Netherlands, Portugal, Switzerland, and the UK. It has affiliates in Australia, Singapore, Malaysia, Thailand, Korea,Taiwan, Philippines, Hong Kong, and Japan, and representative or branch offices in New Zealand, India, Indonesia, and China (Outreville, 2007).

Source: Based on company reports, country-specific studies, miscellaneous newspaper reports.

diagnostic centers for a total of US$53 million from both developed and developing country sources (Chanda, 2007a). Thailand's open FDI regime for hospital services has resulted in several part foreign-owned hospitals, mainly in the Bangkok area, with investments from Japan, Singapore, China, Europe, and the US.

Table 3 summarizes the FDI policies in the health services sector for a representative set of developing countries and their GATS commitments in mode 3. It highlights the extent of liberalization that has been undertaken autonomously in health services and the public policy considerations associated with opening up this sector.

Table 3 indicates that restrictions in the form of limits on foreign equity participation, type of foreign commercial presence, economic needs tests, authorization, certification, and licensing requirements, discriminatory taxes, technology collaboration, and transfer conditions apply in many countries. A comparison of the national policies with the GATS commitments reveals a general unwillingness to legally bind existing FDI regimes or to even undertake GATS commitments in health services.

Countries have also made commitments in health services under bilateral and regional agreements. Obligations of fair and equitable treatment, and pre- and postestablishment national treatment undertaken in Bilateral Investment Treaties (BITs), may also have a bearing on foreign investment in health services, to the extent that health services are covered under the BITs. Overall, however, countries tend to leave health services uncommitted and outside the purview of investment obligations under such agreements. Evidence also suggests that liberalization of FDI in health services has not necessarily translated into increased FDI inflows as structural and regulatory factors continue to constrain foreign providers. (High establishment costs, shortage of quality manpower, and low insurance penetration can constrain foreign investors.)

Impact of Foreign Investment in Health Services

Several studies have discussed the welfare implications of trade in health services, including foreign commercial presence in health services. The effects discussed relate to the resource allocation and accumulation effects of trade liberalization in health services and the likely equity-efficiency tradeoffs. With regard to foreign investment in health services, most authors conclude that the impact on national health systems is shaped by (1) the existing structure of the health system and the extent of commercialization and private sector participation rather than the extent to which the investment is foreign or domestic, and (2) the national regulatory environment.

Table 3 GATS commitments and unilateral FDI policies in health services for selected developing countries

	Hospital Services		*National FDI Policy*	*Other Human Health Services*	
	Mode 3 Commitment under GATS Limitations on Market Access	*Limitations on National Treatment*		*Mode 3 Limitations on MA*	*Limitations on NT*
Bangladesh	No commitment made in GATS	No commitment made in GATS	Commercial presence requires that Foreign Service providers incorporate or establish the business locally in accordance with the relevant provisions of Bangladesh laws, rules and regulations. There is no fixed ratio of equity between local and foreign investors. Foreign equity to the extent of 100% allowed	No commitment made under GATS	No commitment made under GATS
India	Only through incorporation with a foreign equity ceiling of 51%	None	Since 2000, 100% FDI under the automatic route permitted, no government approval required as long as the Indian company files with the regional office of the RBI within 30 days of receipt of inward remittances and files required documents within 30 days of issue of shares to nonresident investors. Foreign Investment Promotion Bureau approval currently only required for foreign investors with prior technical collaboration, but allowed up to 100%	No commitment	No commitment
Jordan	One of the owners must be a physician except in a public limited company. Commercial presence subject to 51% foreign equity limitation. Starting no later than 1 January 2004, 100% foreign equity will be permitted	None	Foreign investors are treated equally as Jordanians. The only difference is that non-Jordanians should deposit bank drafts of no less than 50 000 JD. Hospital sector-foreign investors can have 100% of the property	None, except lab director must be a Jordanian national. Foreign equity participation limits same as under the GATS, full liberalization effective January 1, 2004	None
Malaysia	Hospital Services Private hospital services: economic needs test; only through a locally incorporated joint-venture corporation with Malaysian individuals or Malaysian-controlled corporations or both and aggregate foreign shareholding in the joint	Establishment of feeder outpatient clinics is not permitted	Limited to 30% participation. But commitment under ASEAN Framework Agreement on Services (AFAS) permits higher equity stakes for foreign investors from the ASEAN member countries – 51% in 2008, 70% in 2010	No commitment	No commitment

(*Continued*)

Table 3 Continued

	Hospital Services		*National FDI Policy*	*Other Human Health Services*	
	Mode 3 Commitment under GATS Limitations on Market Access	*Limitations on National Treatment*		*Mode 3 Limitations on MA*	*Limitations on NT*
	venture corporation shall not exceed 30%; and the joint venture corporation shall operate a hospital with a minimum of 100 beds				
Nepal	Hospital services and direct ownership and management by contract of such facilities on a 'for fee' basis: none, except only through incorporation in Nepal and with maximum foreign equity capital of 51%	None	Needs to be registered and approved from Department of Investment, Department of Health and Company registrar's office. Foreign investors can own up to 100% equity in private health firms and are entitled to repatriate the investment and other earnings	No commitment	No commitment
Thailand	No commitment made under GATS	No commitment made under GATS	Must apply for and obtain a Foreign Business License before commencing operation. This category includes the business activity of leasing both fixed and nonfixed assets. Additionally, the activities in which representative offices and regional offices are allowed to engage in are all services that fall under this category	No commitment made under GATS	No commitment made under GATS
Vietnam	Foreign service suppliers are permitted to provide services through the establishment of 100% foreign-invested hospital, joint venture with Vietnamese partners or through business cooperation contract. The minimum investment capital for a commercial presence in hospital services must be at least US$20 million for a hospital, US$2 million for a policlinic unit, and US$200 000 for a specialty unit	None	Some foreign presence exists, though exact shares unavailable; largely complies with GATS commitments	No commitment	No commitment

Source: Based on GATS schedules of commitments in health services.

Hence, the consensus is that the costs and benefits may not be related to foreign investment per se but to the existing regulatory environment and the public–private mix characterizing the country's health system (Chanda, 2001; Smith, 2004; Janjararoen and Supakankunti, 2002).

Overall Cost-Benefit Dynamics

There are three dimensions along which the implications of foreign commercial presence in health services have been assessed. These relate to efficiency, equity, and quality.

Efficiency

Foreign commercial presence can help augment a country's health resources by bringing in additional financial resources, thereby enabling investment in capacity expansion and economies of scale, potentially alleviating the pressure on government budgets, and allowing public funds to be reallocated more efficiently. At the same time, foreign investment could create inefficiencies by encouraging overinvestment of resources in high-end and highly capital-intensive and specialized treatments and procedures with lower cost-effectiveness, while diverting funding from basic healthcare services. Inefficiencies may also arise, if domestic institutions compete by investing in such technologies and procedures at the expense of broader healthcare needs, and if the country's import burden increases. There could also be long-term outflows of payments to foreign investors. There may also be direct and indirect subsidization costs for incentives given to foreign investors. However, the efficiency gains or losses are likely to vary across different countries, depending on the regulatory environment governing such inflows and the infrastructural and human resource conditions, which would shape a country's ability to absorb foreign investment, and the extent to which the private healthcare segment is competitive and dynamic and in a position to derive benefits from the entry of foreign healthcare providers.

Quality

Foreign commercial presence in hospitals and health management may improve the quality of national health systems through the introduction of better management techniques and information systems, better technology, equipment, and infrastructure, and more opportunities for training and skill improvement of medical and management personnel. Foreign-owned or managed healthcare establishments are more likely to follow international standards and to get international certification. There could be positive spillover effects on domestic establishments, which may be incentivized to upgrade their standards, undertake technology investments, and get accredited. Investments in higher end technology and equipment could also provide greater exposure to healthcare professionals, thus helping improve their skills. Better quality small and midsize hospitals, diagnostic labs, and clinics are also likely to tie up with larger hospitals in terms of referral services, thus potentially improving outreach and quality of healthcare for all. Such improvements in the quality of the domestic healthcare system and presence of foreign healthcare providers of global standards could in turn benefit the country by reducing spending on expensive treatments overseas. Once again, these gains would be shaped by the regulatory environment for ensuring quality and standards, the ease with which technology and equipment can be accessed, and the dynamism of the domestic healthcare system.

Equity

Foreign commercial presence may have positive as well as negative implications for equity. Such establishments are more likely to cater to the urban and affluent segments of the population who can afford to pay, potentially aggravating existing inequities in access to healthcare between the rich and poor, between the urban and rural populations. Such establishments are more likely to focus on tertiary care, specialized treatments, and curative and intervention-oriented procedures rather than primary and preventive healthcare needs. There may be cost implications as foreign-owned and managed health providers are likely to be costlier given their higher capital intensity and focus on quality systems and processes and accreditation, which could adversely affect access by the poor out-of-pocket paying population. Foreign investment in health services, particularly in hospitals could also distort the healthcare market by encouraging brain drain of the most qualified and specialized health personnel toward such establishments and away from domestic establishments with offers of better pay and facilities. The latter could adversely affect the quality of medical manpower in competing institutions, particularly, public sector hospitals. Thus foreign commercial presence could accentuate the dualistic structure that often characterizes health systems. Such two-tiering could also weaken the constituency for improving public services.

But there are potential positive implications. The entry of foreign healthcare providers is likely to augment employment opportunities in the healthcare sector at all levels, with better remuneration, especially for specialized and senior medical professionals. Such establishments are also more likely to attract overseas medical professionals and returnees, who are internationally accredited, and could augment human resource capacity and quality in the host country. Some studies have highlighted that foreign healthcare providers may have greater scope to undertake cross-subsidization of poor patients, to do more outreach and extension services, and to establish themselves in second tier cities and towns, given their larger volumes and deeper pockets. Again, the equity implications, positive and negative, are likely to be contingent on factors such as the extent of health insurance penetration, how segmented is the prevailing healthcare system, whether there are regulatory requirements to cross-subsidize the poor and ensure access to the poor in foreign investor hospitals, and the overall quality and availability of human resources.

There are also externalities from foreign commercial presence in health services to other modes of health services trade. Foreign investment in health services can complement medical value travel, telemedicine, and movement of health personnel. Foreign commercial presence and setting up of internationally accredited and recognized hospitals could help attract foreign patients and augment medical value travel exports, reduce imports of health services through outflows of domestic patients, and encourage telemedicine exports.

Outward investment in health services through acquisitions, new ventures, and management and other tie-ups can also benefit exporting institutions through increased foreign exchange earnings, inflows of foreign patients, and greater exposure for their professionals.

Evidence from Selected Countries and Firms

The information in this section is based on company reports (Arunanondchai and Fink, 2007; Chanda, 2007a,b, 2010; Timmermans, 2002; Benavides, 2002; Janjararoen and Supakankunti, 2002; Wadiatmoko and Gani, 2002; Mortensen, 2008a,b), and miscellaneous newspaper articles.

Secondary evidence from a sample of transnational health services firms confirms the aforementioned implications of foreign investment in health services.

- South African hospital companies have succeeded in winning healthcare contracts abroad, including the UK's National Health Service. Netcare established its presence in the UK in 2001. It has helped in reducing wait lists in selected areas of the UK. In 2006, Netcare led a consortium that acquired General Healthcare Group owner of the largest independent hospital operator BMI Healthcare, making it one of the largest healthcare groups with 119 hospitals and almost 11 000 beds. Under this contract, Netcare sends teams of medical personnel from South Africa to its establishments in the UK for fixed periods, thereby enabling its employees to work 4–6 weeks at a time abroad, to get exposure to opportunities overseas, and to supplement their income with fixed term contracts abroad. Such ventures have also helped to reduce staff turnover and improve retention of skilled personnel in South Africa (see, http://www.netcareuk.com).
- Cuba has used joint ventures with Canadian, German, and Spanish companies to attract patients from these countries for specialized treatments. Such investments have helped it to become a hub for teleconsultation and telediagnostic services for the Central American and Caribbean market and have facilitated the establishment of specialized Cuban clinics in Central and Latin America where Cuban physicians and nurses are employed.
- In its bid to become the medical center of the Arab world, **Jordan** has provided incentives for national and foreign private investment in the health sector. This has led to the establishment of several private hospitals with foreign financing and tie-ups, benefiting the Jordanian health system through state-of-art technology, computerized links with prestigious health centers in Europe and North America, and medical value travel exports to the region.
- India's Apollo Group of Hospitals highlights the linkages between mode 3 and other forms of trade in health services. Apollo's mode 2 exports have been facilitated by its overseas marketing offices and management contracts with hospitals in the UAE, Saudi Arabia, Oman, Kuwait, Mauritius, Tanzania, UK, Sri Lanka, Bhutan, Nigeria, Pakistan, and Bangladesh. Apollo Gleneagles, which is a joint venture with the Singapore-based Parkway Group, exports health services to patients from neighboring countries like Bangladesh, Nepal, Bhutan, and Myanmar. It also provides telemedicine services such as medical consultation, diagnostic, telepathology, teleradiology, and scanning services. Apollo also provides contract research and medical education and training services through its overseas subsidiaries, using a combination of cross border supply (online training and research services) and temporary onsite deployment of professionals at its subsidiaries, thereby benefiting its own professionals and also host country professionals.
- In India, Hindustan Latex Ltd and Acumen Fund (USA) have created a joint venture to develop a small chain of high-quality and affordable (30–50% of regular price) maternity hospitals to serve the low-income population in underserved Indian regions. The aim is to make this a global model for increasing access to qualitative and affordable healthcare for the poor.
- The public and the private sectors in China have jointly developed a strategy to attract foreign health providers to set up commercial presence. Chinese institutions have entered into joint ventures with partners in the medical profession and with local authorities overseas. Traditional Chinese Medicine facilities have been established in more than twenty countries. Such joint ventures help spread Traditional Chinese Medicine overseas, enable the deployment of Chinese health workers and their exposure to other systems under contractual arrangements, and help attract patients to China.
- Evidence from some ASEAN countries shows that foreign investor hospitals can aggravate the existing inequities in the host country's healthcare system. Most of these hospitals have located in and around the main cities such as Bangkok and Jakarta and target the upper income segment. The Indonesian government has thus imposed fewer regulatory requirements on foreign investors in regions with weak public health infrastructure to attract foreign investors to islands other than Java and to the smaller cities. The Indonesian government has also imposed a requirement to accommodate at least 200 beds in foreign investment hospitals.

Primary Evidence on Impact: Case Study of India

A survey of 25 hospitals conducted in 2007 across several major cities in India examined the realized or perceived impact of foreign investment in Indian hospitals on quality, affordability, infrastructure, range of services, technology, accessibility, and prices. The survey findings largely corroborate earlier studies (Chanda, 2007a,b).

Services and infrastructure

Foreign investor hospitals were found to focus on more advanced and specialty services compared to domestic hospitals, indicating a greater emphasis on niche areas and on high revenue generating curative and surgical interventions as opposed to preventive care. The survey also revealed that foreign investor hospitals tend to invest more heavily in high-end technology and state-of-the-art equipment, which in turn leads to a difference in approach to medical care, with more intensive use of medical equipment in order to recover

investments. On average, foreign investor hospitals were also found to have more medical facilities, more equipment, and to be larger in terms of the number of beds, rooms, ambulances, and Intensive Care Unit infrastructure. Foreign funded institutions also reported greater availability of postoperative care facilities and critical care services. There was also greater availability of medical staff for critical care and specialized services as opposed to general care.

Human resources: Remuneration and quality issues

The survey findings showed that foreign investor hospitals pay higher salaries to their medical staff at all levels and particularly to senior specialists, suggesting the possibility of internal brain drain from domestic private as well as public sector hospitals to foreign investor hospitals. The findings on remuneration also suggest that there could be positive implications for employment and income opportunities for medical personnel. Hence, there is evidence on a likely two-tiering impact.

Costs of services

The data on costs of different procedures and treatments indicated that foreign investor hospitals tended to be more expensive than comparable domestic health providers. Indepth discussions with industry experts revealed that hospitals were on average 15–30% costlier than small and medium size healthcare providers.

Spillover effects

The study found a strong spillover effect on medical value travel. Increased foreign investor presence in hospitals was seen as facilitating medical value travel to India by enabling tie-ups with foreign health insurance providers and development of customized insurance products for elective surgeries by overseas patients in India with follow-ups abroad. Foreign investment in hospitals was seen as encouraging the entry of multinational insurance companies, which would be more comfortable in dealing with foreign funded corporate hospitals that were accredited and accountable. Respondents also noted that foreign investment in hospitals would spur expansion of activities in other areas such as medical transcriptions, back-office medical outsourcing, and telemedicine as well as promotion of opportunities in other areas such as clinical trials outsourcing, research and development, and medical training and education. Several respondents noted the likely boost to telemedicine from foreign commercial presence, given investments by foreign players in IT systems. Strong positive externalities were also perceived in the form of technology and knowledge transfer through tie-ups for research and development, technology sharing, professional exchange, and continuing medical education.

Concerns

The survey highlighted some areas of concern, along the lines suggested by other studies. Increased foreign investment in hospitals was seen as aggravating the internal brain drain of medical personnel from public to private healthcare establishments and making it more difficult for the public sector to retain doctors and teachers in affiliated medical colleges. It was noted that the increased focus on earning money would mean less focus on teaching and research, especially on issues relevant to local conditions. Several respondents highlighted the fact that small and medium size nursing homes would face greater competition from the large corporate hospitals, have difficulty retaining staff, and would become less attractive as they would not be able to provide many services under one roof. Hence, many would have to close down or would be acquired by the larger players. Similar concerns were expressed for independent pharmacies and diagnostics/labs. It was also felt that as foreign funded hospitals provide better remuneration, their expansion would put upward pressure on wages and salaries of medical personnel and thus increase competition for quality manpower. A third concern was related to costs, affordability, and relevance of healthcare following increased foreign investment in hospitals and possible adverse effects on the poor who might be squeezed out of the system. Several respondents expressed concern that foreign investment in hospitals and the focus on profits and returns for shareholders would lead to increased healthcare costs, increasing the existing income and geographic divide in healthcare delivery.

Concluding Thoughts

Foreign investment in health services has grown over the past decade, taking a variety of forms and involving a growing number of developed and developing countries. Although it is difficult to quantify the impact of foreign investment on national health systems, several general studies highlight the likely pros and cons of such investment. There is broad agreement on the various positive and negative implications. An important conclusion of these studies is that the impact on national health systems is a function of regulatory frameworks, the prevailing market structure, and the extent of commercialization. Although foreign investment may have adverse implications for equity, affordability, and on the public sector, the real underlying cause could be the prevailing distortions in the healthcare system and not foreign investment.

A key policy inference is that it is possible to shape the impact of foreign investment on national health systems and that possible negative fallouts should not lead to a restrictive approach to such investments. The negative effects can be mitigated and prevented. The positive effects can be facilitated through appropriate policies and regulations. For instance, public–private partnerships and facilitation of linkages between the public and private health services segments with regard to medical education, training, staff and information exchange, can be encouraged to reduce the scope for two-tiering. Initiatives to increase insurance penetration and conditions requiring foreign investors to provide medical outreach and extension services in less served areas, could mitigate the negative equity fallouts.

Clearly, more research is required across a mix of countries with different health systems and regulatory environments to draw more definitive, evidence-based conclusions. More dialog is also required between the commerce and health ministries and investment boards to enable an integrated social and economic perspective and to accordingly frame an appropriate mix of investment incentives and conditions to balance the tradeoffs.

See also: Health and Health Care, Macroeconomics of. International Trade in Health Services and Health Impacts. Medical Tourism

References

Arunanondchai, J. and Fink, C. (2007). Trade in health services in the ASEAN region. *World Bank Policy Research Working Paper No. 147*. Washington, DC: World Bank.

Benavides, D. (2002). Trade policies and export of health services: A development perspective. In Drager, N. and Vieira, C. (eds.) *Trade in health services: Global, regional and country perspectives*, pp. 53–69. Washington, DC: PAHO/WHO.

Cattaneo, O. (2009). Trade in health1 services: What's in it for developing countries. *World Bank Policy Research Working Paper No. 5115*. Washington, DC: World Bank.

Chanda, R. (2001). Trade in health services. *Paper No. WG4:5*. WHO, Geneva: Commission on Macroeconomics and Health.

Chanda, R. (2007a). Foreign investment in hospitals in India: Status and implications. New Delhi: WHO Country Office, India and the Ministry of Health and Family Welfare.

Chanda, R. (2007b). Impact of foreign investment in hospitals: Case study of India.*Harvard Health Policy Review* **8**(2), 121–140.

Chanda, R. (2010). Constraints to FDI in hospital services in India. *Journal of International Commerce, Economics and Policy* **1**(1), 121–143.

Holden, C. (2002). The Internationalization of long term care provision. *Global Social Policy* **2**(1), 47–67.

Holden, C. (2005). The internationalization of corporate healthcare: Extent and emerging trends. *Competition & Change* **9**(2), 185–203.

Janjararoen, W. and Supakankunti, S. (2002). International trade in health services in the millennium: The case of Thailand. In Drager, N. and Vieira, C. (eds.) *Trade in health services: Global, regional and country perspectives*, pp. 87–106. Washington, DC: PAHO/WHO.

Mortensen, J. (2008a). International Trade in Health Services – the trade and the trade-offs. *Working Paper 11*. Copenhagen: Danish Institute for International Studies.

Mortensen, J. (2008b). Emerging multinationals: The South African hospital industry overseas. *Working Paper 12*. Copenhagen: Danish Institute for International Studies, University of Copenhagen.

Outreville, J. (2007). Foreign direct investment in the health care sector and most-favoured locations in developing countries. *European Journal of Health Economics* **8**, 305–312.

Smith, R. (2004). Foreign direct investment and trade in health services: A review of the literature. *Social Science and Medicine* **59**, 2313–2323.

Timmermans, K. (2002). Overview of the South-East Asia region. In Drager, N. and Vieira, C. (eds.) *Trade in health services: Global, regional and country perspectives*, pp. 83–86. Washington, DC: PAHO/WHO.

Wadiatmoko, D. and Gani, A. (2002). International relations within Indonesia's hospital sector. In Drager, N. and Vieira, C. (eds.) *Trade in health services: Global, regional and country perspectives*, pp. 107–117. Washington, DC: Pan-American Health Organization/WHO.

Waeger, P. (2007). Trade in health services: An analytical framework. *Working Paper No. 441*. Kiel, Germany: Kiel Institute for world Economics, Advanced Studies Programme 2005/2006.

Further Reading

Blouin, C., Drager, N. and Smith, R. (eds.) (2006). *International trade in health services and the GATS, current issues and debates*. Washington, DC: World Bank.

Chanda, R. (2002). Trade in health services. *Bulletin of the World Health Organisation* **80**, 158–163.

Fortune Global 500 list. Available at: http://money.cnn.com/magazines/fortune/global500/2010/index.html (accessed on April 2011).

Gupta, I. and Goldar, B. (2001). *Commercial presence in the hospital sector under GATS: A case study of India*. New Delhi: WHO SEARO.

ITC. Investment map. Geneva: ITC. Available at: http://www.investmentmap.org/TimeSeries_Country_fdi.aspxprg=1%20 (accessed 24.05.11).

Mackintosh, M. (2003). Health care commercialisation and the embedding of inequality. *RUIG/UNRISD Health Project Synthesis Paper*. Geneva.

Maskay, N. M., R. K. Panta, and B. P. Sharma (2006). Foreign investment liberalization and incentives in selected Asia-Pacific developing countries: Implications for the health service sector in Nepal. *Working Paper Series No. 22*. Bangkok: Asia-Pacific Research and Training Network on Trade, ARTNeT.

Netcare (various) *Annual Report,* various years. Available at: http://www.netcareinvestor.co.za/rep_annual_reports.php

Smith, R., Chanda, R. and Tangcharoensathien, V. (2009). Trade in health-related services, Trade and Health Series. *The Lancet* 29–37.

UNCTAD (2004). The shift towards services. *World Investment Report*. Geneva: UNCTAD. Available at: http://unctadstat.unctad.org/ReportFolders/reportFolders.aspx

UNCTAD, Transnationality Index online database. Available at: http://unctad.org/en/Pages/DIAE/DIAE%20Publications%20%20Bibliographic%20Index/Transnational_Corporations_Journal.aspx

UNCTAD, FDI Statistics. online database. Available at: http://unctadstat.unctad.org/ReportFolders/reportFolders.aspx

Woodward, D. (2005). The GATS and trade in health services: Implications for health care in developing countries. *Review of International Poliltical Economy* **12**(3), 511–534.

WTO, GATS *Commitment schedules for selected countries*. Geneva. Available at: http://tsdb.wto.org/default.aspx

Relevant Websites

http://portal.bsnl.in/bsnl/asp/content%20mgmt/html%20content/business/business56857.html
Bharat Sanchar Nigam Limited.

http://www.thehindubusinessline.in/2005/12/03/stories/2005120303200200.htm
Business Line (Dec'05).

International Trade in Health Services and Health Impacts

C Blouin, Institut national de santé publique du Québec, Québec, Canada

Glossary

General Agreement on Tariffs and Trade (GATT) The predecessor of the World Trade Organization (to 1994).
General Agreement on Trade in Services (GATS) GATS were an outcome of the 1995 Uruguay Round Negotiations and basis of the global multilateral sector trading system.
Medical tourism A term describing the travel of individuals to countries solely for the purpose of receiving health care.
Multilateral A relationship, such as a trading relationship, involving many partners, such as countries, trading with many others.
Multiplier The fiscal multiplier measures the eventual change in national income that results from an initial change in a component of aggregate demand in the economy.
Telemedicine The use of information and communication technologies to deliver clinical healthcare services at a distance.
World Trade Organization The global institution that deals with the rules of trade between countries.

Introduction

The first section of this article reviews the risks associated with cross-border trade as well as legal consequences of trade treaties, focusing on World Trade Organization (WTO) agreements. It also discusses three features of the WTO agreements, which provide space for addressing the tensions between the economic objectives of trade policy and public health objectives. The second section of the article reviews how the WTO has been adjudicating disputes which have had health implications; indeed, the WTO dispute settlement mechanism is a venue whose explicit function is to weigh the objectives of facilitating international trade with other objectives included in these same treaties such as the protection and promotion of human health. The article concludes with some illustrations of ongoing exercises of global health diplomacy where tensions between trade and health policy objectives are being negotiated.

How Can Trade Affect Health?

When we examine the impact of trade on health, we are looking at two types of independent variables. First, it can refer to international trade rules as they are embodied in multilateral, regional, and bilateral trade and investment treaties negotiated by national governments. Second, it includes the impact of economic integration, i.e., increased cross-border flows of goods, services, and capital. Trade agreements can increase economic integration and the intensity of these cross-border

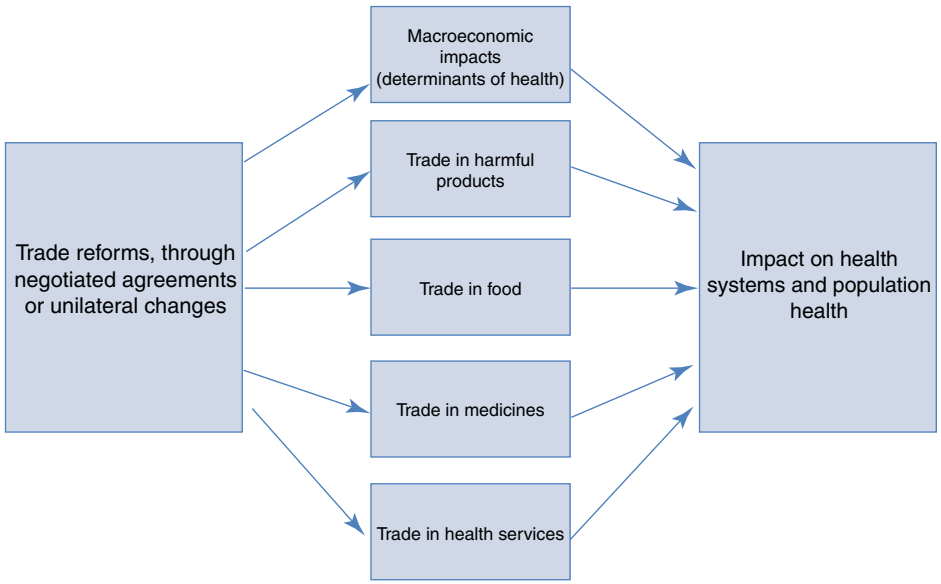


Figure 1 Trade and health key linkages.

flows; however, these may take place in the absence of treaties and should be considered as a separate analytical entity. Trade so defined can have impact on health systems and population health through a number of transmission channels. There are five main causal chains (see Figure 1). Trade reforms can have an impact on the macroeconomic conditions of a country to facilitate or hinder population health through changes in characteristics, such as poverty and inequality. Trade reforms can also ease or restrict access to harmful products, such as tobacco, weapons, or toxic waste. Third, trade policy in the agricultural sector can affect population health through its impact on food security, diet, and nutrition. Fourth, trade agreements also influence access to medicines by including patent protection such as we find in the WTO's Agreement on Trade-related Aspects of Intellectual Property Rights (TRIPS). This is the trade-health linkage that has received most attention from academics, policy makers, and civil society organizations in the last 15 years. The final section of the article focuses on a fifth channel, trade in health services.

National health systems can be transformed by the introduction of cross-border suppliers and investors. Trade in health services can take four different forms. First, the services can be provided electronically with both the patients and the providers remaining in their own jurisdictions; telemedicine across border is an example of such a trade. This is called Mode 1 of the General Agreement of Trade in Services (GATS) of the WTO.

Second, patients can travel abroad to receive care; health tourism or medical tourism has received a lot of attention in research and policy circles in recent years. There are good indications that there is a steady increase of health tourism, even though the actual scale is not well measured. Concerns have been raised regarding the quality of care and equity, in terms of the impact of such trade on the health system of the countries to which patients travel. Indeed, health tourism has been presented as an economic opportunity for many middle-income countries that are struggling with relatively weak health systems. The main concern is the risks of reallocation of resources away from local patients toward higher quality care supplied to affluent domestic and foreign patients.

However, Mode 2 trade can become an important source of foreign exchange earnings and add to the multiplier effects of tourism-related activities in the host economy. Promoting health tourism can also lead to efficiency gains for importing countries. According to one estimate, the health care system in the US would save $1.4 bn annually if only one in ten patients were to go abroad for a limited set of 15 highly tradable, low-risk treatments.

Another potential positive contribution is that some of the revenues from health tourism be harnessed to improve access to health care services for the local population. Typically, advocates of health tourism will recommend that governments in developing countries "put in place universal access policies that require private providers to contribute to a health care fund" (Mattoo and Rathintran, 2005). However, a review of the literature and the institutional frameworks related to health tourism failed to identify such a mechanism. Neither in the more established health tourism destinations like India, Jordan, Thailand, nor in countries that are more recently involved in this form of service exports (the Caribbean, Mexico, Costa Rica) has an explicit mechanism to allocate some of the additional income generated from health tourism been used to increase access to health care services for local patients. The only country found to have mentioned a specific tax on health tourism is New Zealand, where the government was considering in 2009 to apply on specific levy on private hospitals catering to foreign patients which would contribute to the Accident Compensation Corporation, a public agency which provides a comprehensive, no-fault injury insurance to all New Zealanders and visitors (reference http://www.imtj.com/news/EntryId82=166606).

The third mode of cross-border supply of health services according to the GATS relates to the movement of capital, such as foreign investors investing in the establishment or the management of a clinic or a hospital. The potential benefits of Mode 3 trade in health-related services are to generate additional investment in the health care sector, contribute to upgrading health care infrastructure, facilitate employment generation, and provide a broader array of specialized medical services than those available locally. However, the potential downside risks of Mode 3 trade once more include growing inequality in access and the emergence of a two-tiered health care system. This two-tiered system may result from an internal 'brain drain,' as foreign commercial ventures may encourage health professionals to migrate from the public to the private health care sector.

Trade in health services can also take place through the temporary movement of natural persons (so-called Mode 4 of the GATS); a nurse, physician, or other health professional practice abroad on a temporary basis. Mode 4 trade is still limited relative to its potential due to a number of regulatory barriers posed by recipient countries. These barriers include immigration rules, discriminatory treatment of foreign providers, and the nonrecognition of foreign qualifications. Virtually all countries impose restrictions on temporary migration and the quotas are usually substantially lower than the actual demand for entry. The cross-border movement of health care professionals may promote the exchange of clinical knowledge among professionals and therefore contribute to upgrading their skills and medical standards. The potential downside risks of Mode 4 trade arise from the danger that such mobility may be of a more permanent nature, such that health care professionals often trained at considerable home country expense are for ever lost, thus reducing the availability and quality of services on offer to home country consumers of health care services.

Trade rules can affect the cross-border supply of health services. Indeed, the main reason national governments agree to sign trade treaties is to increase access to foreign markets and facilitate international trade. Governments make commitments in trade agreements such as GATS, where they guarantee access for foreign investors interested in establishing a new clinic or health insurance company with a view to facilitate and increase cross-border flows of services and capital. Governments can unilaterally adopt reforms where they allow foreign services providers to compete in the domestic markets through one or all of the four modes of supply; however, including this reform into the binding commitments of a trade agreement decrease the likelihood that this policy will be reversed in the future.

Type restriction	
Market access	
# of beds	Austria, European Community, Latvia, Malaysia, Oman
Investment	Croatia, Vietnam
Needs test	Croatia, European Community, Lithuania, Malaysia, USA
Other regulation	Pakistan, St Vincent
Ownership	India, Jordan, Malaysia, Mexico, Nepal, Saudi Arabia, Chinese Taipei, USA, Vietnam
Prior authorization	Austria, European Community, Latvia, Lithuania, Slovenia, Turkey
Registration/licensing	Jamaica
Residence/nationality	Cambodia, Latvia, Poland, Chinese Taipei
National treatment	
Access to financial support	Lithuania, Poland, Slovenia
Legal entity	Malaysia
Other regulation	Oman
Registration/ licensing	Albania

Type of restriction	Defination
# of beds	Restriction on the number of beds in a health care facility.
Access to financial support	Restrictions on the access to financial support from public resources
Language	Requirement for the knowledge of a specific language.
Legal entity	Restrictions on the type of legal entity that can supply a service or benefit from a specific provision.
Investment	Requirements for the type and amount of foreign investment
Needs test	Requirements for local or economic needs tests.
Other regulation	Reference to a specific domestic legal act or regulation affecting market access or nationa l treatment.
Ownership	Requirements for the percentage or amount of foreign equity.
Prior authorization	Requirements for prior authorization for establishment or other activity from a ministry or another authoritative body.
Recognition of qualifications	Requirements regarding qualification and examination equivalents.
Registration/licensing	Registration or licensing requirements.
Residence/nationality	Requirements regarding the residency or nationality of service suppliers, the board of directors or other employees.
Operational experience	Requirement on the length of operational experience.

Figure 2 Summary of GATS Mode 3 commitments and restrictions in hospital services (updated in November 2009).

In the case of health-related services, WTO members have made relatively few and limited commitments in the GATS. One can argue that one option for policy makers to address the tensions between trade and health has been to prefer unilateral trade reforms rather than to include liberalization of health-related services into multilateral trade treaties. In that manner, they maintain a greater flexibility to experiment with cross-border health services provision and reverse reforms if they fail to deliver the desired outcomes. The detailed nature of these trade commitments in services provides a second avenue for government to address the tensions between their trade and public health objectives. Indeed, WTO member states can fine-tune their GATS commitments according to which of the four modes and specific health-related services they want to include in their list of commitments. They can also decide whether they want to commit to national treatment (no discrimination against foreign providers vs. domestic ones) or to market access (removing barriers to entry). They can also stipulate specific conditions for entry for foreign providers. For instance, the European Community and

Malaysia have stipulated in their commitments that entry of foreign investors in hospital services is subjected to an economic need test and to limits to the number of beds in the hospital (see Figure 2). These provisions can allow health authorities to channel foreign investments in hospitals in regions and of the size required as to their health care system planning.

Given the flexibility built into its design, it can be argued that the GATS provides the margin for maneuver for policy makers to harness the positive impacts of trade in health services, while mitigating the associated risks. However, we should note that other trade agreements do not have the same design and do not offer the same level of flexibility. For instance, governments have become parties to a vast network of investment agreements which aims to offer a predictable environment for foreign investors by protecting them against some kinds of state actions, such as discrimination and expropriation without compensation, and, as a result, to encourage foreign investment. These agreements, whether they are integrated in larger trade agreements or are stand-alone bilateral investment treaties, can exclude some sectors, but they tend to have a broader coverage with fewer exceptions and carve-outs, hence offering less space for addressing potential tensions between trade and health.

A third manner in which tensions between trade and health can be negotiated is at the national level with the adoption of domestic public policies which mitigate the negative impacts and harness the positive consequences. For instance, in Thailand, the increase in health tourism had a negative impact on the human resources for health in the country as nurses and physicians were attracted to work in the large urban hospitals catering to foreign patients, exacerbating the urban–rural gap in terms of access to health care services. To address this problem, the government significantly increased admissions in nursing and medical schools. These 'flanking' policies can take many forms, but they all require policy coherence, i.e., national authorities need to make their policy choices and the impacts of these choices explicit, realizing the potential divergence and trade-offs to be made between the realization of economic/trade policy objectives and public health goals, at least in the short term.

How has the WTO Managed the Tensions between Trade and Health?

When the agreements of the WTO came into force in 1995, they included a new dispute settlement mechanism which has become, since its creation, a key forum for managing the tensions between trade and health. Indeed, member states of the WTO have brought a number of disputes to the Panel and its Appellate body, which involved measures designed to protect human health, or claiming to do so.

How have these WTO panels and the Appellate body arbitrated disputes where the objectives of trade and the public measures to promote and protect public health clash? First, the WTO has defended the right of national governments to adopt public measures, even if they violate WTO rules, by claiming that these measures were necessary to protect human health under the exception found in General Agreement on Trade and Tarriffs (GATT) Article XX(b). Thus, when in 1998 Canada challenged France's ban on asbestos, the Panel estimated that the measure violated the national treatment principle in the GATT (Article III:4) which prevent parties from discriminating foreign products in favor of domestic products. In this case, the Panel judged that the French-made products containing polyvinyl acetate (PVA), cellulose, and glass fibers were similar to foreign products containing asbestos fibers (and therefore were like products as defined by Article III:4 of the GATT). Even though they deemed the measure discriminatory, the Panel agreed that the ban on asbestos was justified, given the health exception in Article XX(b) of the GATT. The Appellate body supported that view but went further and concluded that in determining whether products are similar, health impacts should be considered; hence, considering that products containing asbestos fibers should not be seen as similar to products containing PVA, cellulose, or glass fibers.

The WTO health exception specifies that the measures to protect human health should not be "applied in a manner which constitutes a means of arbitrary or unjustifiable discrimination or a disguised restriction on international trade." In 2007, the WTO concluded that Brazil was applying its ban on import of used tires in a discriminatory manner. The arbitrators did not challenge the right to adopt measures to protect public health, even though they were violating the national treatment principle. In this case, the ban on imports was adopted in order to limit the breeding grounds for diseases-transmitting mosquitoes created by stockpiling of discarded tires. The problem was that Brazil has allowed some imports from South American neighbors, whereas other countries such as the members of the European Union were under the complete import ban.

An earlier case involving an American regulation on gasoline had affirmed the capacity of WTO members to restrict trade to protect human health as long as trade-restricting health measures do not discriminate in violation of the national treatment principle (GATT Article III:4) by treating imported products less favorably than like domestic products. "Trade-restricting health measures that violate the national treatment principle may still be legitimate under GATT Article XX if such measures (1) fall within one of Article XX's enumerated exceptions, and (2) are applied in a manner that does not constitute unjustifiable discrimination or a disguised restriction on international trade." (Fidler, forthcoming).

Beyond the health exception, another principle which has been key in guiding WTO arbitrators when they have to manage the tensions between trade rules and public health objectives is the need for scientific risk assessment when adopting a sanitary and phytosanitary (SPS) measure. Indeed, the SPS Agreement (Article V) requires domestic regulation to be based on a risk assessment which takes into account available scientific evidence. The appropriate level of protection should be determined in consideration of economic factors such as the loss of production, the cost of control or eradication, the cost of alternative approaches, and with a view of minimizing negative trade effects.

The first WTO dispute involving the SPS agreement was initiated in 1996, by the US and Canada who complained that the prohibition enacted by the European Communities (EC) on

the importation and sale of meat treated with growth hormones in order to protect human health violated the SPS Agreement. The Panel and the Appellate body agreed with Canada and the US that the EC had violated Article 5.1 of the SPS Agreement and that the evidence presented by the EC's risk assessment did not support a total ban on meat with growth hormones. The other WTO dispute which related to the SPS agreement, the dispute around the EC *de facto* moratorium on biotech products such as genetically modified food, did not challenge the European assessment of the risks associated to the products. The European measures were violating procedural requirements of the SPS agreement.

Finally, the TRIPS agreement includes a clause (Article 30) which allows members to adopt policies that contravenes other provisions of the agreement, which has been used in a health-related dispute. This exception was tested by the disputes between Canada and the EU on patent protection for pharmaceuticals which began in 1999. With a view to reduce prices and improve access, generic pharmaceutical manufacturers in Canada were allowed to produce a drug under patent without the patent holder's permission in order to (1) obtain regulatory approval for the generic pharmaceutical product, and (2) produce a stockpile of generic drugs to sell when the patent expired. The government argued that even though these rules were violating some aspects of the TRIPS agreement, they fell within the exceptions provided by Article 30 of TRIPS. The WTO Panel partially agreed with Canada, ruling that allowing stockpiling before the expiration of the patent could not be justified under the public health exception.

Global Health Diplomacy in On-Going Trade Negotiations

Trade negotiations which can have an impact on health systems and population health through the five channels illustrated in Figure 1 are still on-going in a diversity of global and regional contexts. For instance, the European Union has been negotiating economic partnership agreements (EPA) with four regional groups in Africa since 2007. Because the EPAs touch on a wide range of trade-related issues, some have expressed concerns that they can potentially have a negative impact on health in sub-Saharan Africa. Four main areas of concern have been raised in this case: The impact of trade liberalization on public revenues and therefore the public expenditures for health; the risks of increasing patent protection in terms of access to pharmaceutical drugs; the opening of health services to foreign investment; and the impact of agricultural liberalization on food security and poverty.

What are the means of balancing these concerns against the potential economic benefits associated with trade liberalization? One means proposed is the use of health-impact assessments (HIAs) of proposed trade reforms. HIA are a set of procedures for assessing the potential impact of public policies on population health and the distribution of these effects on the population. It has been proposed they can be a useful tool as it can make the linkages between trade and health more visible to policy makers, it can improve the quality of evidence available to them, it can influence how the goals of trade policy are perceived by policy makers, and it can be used by various interest groups as an instrument for advocacy and mobilization. Equipped with the information from an HIA, trade negotiators are better positioned to decide to forgo certain trade commitments, to include some restrictions and limitations on their commitments or again, to adopt domestic policies which will ensure that trade policy does not impede the realization of national health objectives.

Except for the impact of patent protection on access to medicines, the linkages between trade and health are still an underexplored area of research and policy debates. Trade negotiations may be stalled at the WTO, but there are a number of trade and investment treaties being negotiated at the regional and bilateral levels which requires a more explicit approach to address the tensions between trade and health policy objectives.

See also: International E-Health and National Health Care Systems. International Movement of Capital in Health Services. International Trade in Health Workers. Medical Tourism

References

Fidler, D. Summary of key GATT and WTO cases with health policy implications. In Blouin, C., Richard S. and Drager N. (eds.) *Diagnostic tool on trade and health*. Geneva: WHO, forthcoming.

Mattoo, A. and Rathindran R. (2005) Does health insurance impede trade in health care services? *World Bank Policy Research Working Paper, No. 3667*, July.

Further Reading

Blouin, C., Drager, N. and Richard, S. (eds.) (2006). *International trade in health services and the GATS: Current issues and debates*. Washington, DC: The World Bank.

Fairman, D., Diane, C., McClintock, E. and Drager, N. (2012). *Negotiating public health in a globalized world: Global health diplomacy in action*. Dordrecht: Springer.

Hopkins, L. R. L., Vivien, R. and Corinne, P. (2010). Medical tourism today: What is the state of existing knowledge? *Journal of Public Health Policy* **31**, 185–198.

Lee, K., Ingram, A., Lock, K. and McInnes, C. (2007). Bridging health and foreign policy: The role of health impact assessment. *Bulletin of the WHO* **85**, 207–211.

Pachanee, C. and Wibulpolprasert, S. (2006). Incoherent policies on universal coverage of health insurance and promotion of international trade in health services in Thailand. *Health Policy and Planning* **21**, 310–318.

World Trade Organisation and World Health Organisation (2002). *WTO agreements and public health: A joint study by the WHO and WTO secretariats*, Geneva.

Relevant Websites

http://www.thelancet.com/series/trade-and-health
Lancet on Trade and Health.

http://www.ghd-net.org/
The Global Health Diplomacy Network (GHD-NET).

http://www.who.int/trade/trade_and_health/en/
The World Health Organisation.

International Trade in Health Workers

J Connell, University of Sydney, NSW, Australia

Introduction

The international migration of skilled health workers (SHWs) has grown rapidly since the 1970s, become more complex, more global, and of concern to countries that lose workers from fragile health systems. As health care has become more commercialized, so too has migration, as part of a wider globalization of health services. Few parts of the world, either as sources, destinations, or both, within a now global healthcare chain, are unaffected by the consequences. Most migration is to developed Organization for Economic Cooperation and Development (OECD) countries, in Europe, North America, and also the Gulf. Countries most affected by emigration are relatively poorly performing economies in sub-Saharan Africa, alongside some small island states in the Caribbean and Pacific, though absolute numbers are greatest from such Asian countries as India and the Philippines.

The international migration of SHWs parallels somewhat similar international migration of other professionals. The emergence of regional trading blocs and agreements, notably the European Union (EU), has expanded opportunities for international migration. International migration is linked to the General Agreement on Trade in Services, established in 1995, to liberalize international trade in services, including the movement of the so-called 'natural persons.' Many countries have eased their legislation on the entry of highly skilled workers, introduced points systems where skills facilitate entry, and actively recruited overseas. Such professional services as health care are part of the new internationalization of labor, and migration has largely been demand driven (or at least facilitated) by the growing global integration of healthcare markets. Forty years ago doctors – mostly men – were the main migrant group, but nurses – mostly women – have increasingly become dominant.

Demographic, economic, political, social, and, of course, health transformations have had significant impacts on international migration. Restructuring, often externally imposed, has affected health systems of developing countries, contributing to concerns over wages, working conditions, training, and other issues, all of which have stimulated migration. The health sector is different from other skilled sectors because most employment remains in the public sector. More dramatically, migration literally involves matters of life and death. Technology cannot easily replace workers, while the rise of human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) and non-communicable diseases and the aging of populations have placed new demands on health workforces. There is now a greater range of jobs for women, other than in a sector that is seen by some as dirty and dangerous (and unrewarding), sometimes difficult and demanding, and perhaps degrading. SHWs are in global demand.

Migrants move primarily for economic reasons, and increasingly choose health careers because they offer migration prospects. Migration has been at some economic cost, has depleted workforces, diminished the effectiveness of healthcare delivery, and reduced the morale of the remaining workforce. Countries have sought to implement national policies on wage rates, incentives, and working conditions, but these have usually been canceled out by global uneven development and national economic development problems. Recipient countries have been reluctant to establish effective ethical codes of recruitment practice, or other forms of compensation, or technology transfer, hence migration may increase further in future, despite the development of a Global Code.

Around the turn of the century, accelerated recruitment from developing countries, where populations are aging, expectations of health care are increasing, recruitment of health workers (especially nurses) is poor, and attrition considerable, contributed to a labor force crisis in source countries, raising complex ethical, financial, and health questions. The costs of training healthcare workers in developing countries are considerable, hence migration has been perceived as a subsidy from the poor to the rich. Migration issues are not only linked to financial issues, serious though these are, but are critical for the delivery of health care.

A Geography of Need

Human resources are central to healthcare systems, and have long been unevenly distributed. The need for health care is at least as uneven. Though definition and measurement of needs and shortages is complex, and the competence and effectiveness of workers hard to assess, demand for health care is greatest in the least-developed countries and regions, most of which are tropical, and, in a perfect example of the 'inverse health care law,' these needs are less well served than those in developed countries. The link between 'health workforce density' and health outcomes has been clearly demonstrated: Lack of health workers contributes to poor health status, and provision of such basic functions as adequate coverage of immunization or attendance at births. The disease burden is especially great in sub-Saharan Africa. World Health Organization (WHO) has shown that North and South America contain only 10% of the global burden of diseases, yet almost 37% of the world's health workers live in this region, whereas Africa has 24% of the global burden of diseases, but just 3% of the health workforce and less than 1% of global financial resources. At a national scale, the sub-Saharan countries of Uganda and Niger have 6 or 7 nurses for 100 000 people, whereas the US has 773, yet migratory flows – a perverse flow – are invariably from the former to the latter.

The WHO estimated that in 2005 some 57 countries had critical shortages of SHWs, equal to a global deficit of 2.4 million doctors, nurses, and midwives, let alone pharmacists, dentists, radiologists, and others. Some 36 of 47 sub-Saharan African countries fell short of the minimum. Moreover, most

 doi:10.1016/B978-0-12-375678-7.00614-3

SHWs are concentrated in urban areas and usually in the often primate city: A consequence of economies of scale, urban bias, and the social preferences of SHWs.

A Brief History of Migration

In the nineteenth century, migration of SHWs from more developed countries to their colonies was part of a colonial endeavor and missionary practice that remained in place until quite recently. In the 1940s and 1950s the direction reversed, from south to north, and the first flow of health workers began to migrate from developing countries, mainly to the UK and the USA, and mainly of doctors from larger countries such as India, Iran, and Pakistan. Nurses also migrated and were later recruited for emerging Gulf states. Britain, Australia, and Canada were experiencing both the immigration of doctors, mainly from the Philippines, India, Pakistan, Iran, and Colombia, and their emigration, usually to the USA. The Philippines had already contributed the largest number of overseas doctors in the USA, with training increasingly oriented to overseas needs. By then the ethnic distinctiveness of this skilled migration into Britain was evident, and a geographical pattern had emerged that scarcely changed substantially in later years. Over time what were then relatively simple migration flows, reflecting linguistic, colonial, and postcolonial ties, became steadily more complex. This new phase of migration was the start of what became widely recognized as a 'brain drain,' a term first applied by Oscar Gish at the end of the 1960s to the movement of doctors and scientists.

Such early flows were also characterized by active recruitment (notably of nurses from the Caribbean), and the employment of the migrants in the lower echelons of the health service. By the 1960s, the less-developed countries were experiencing the greatest costs from emigration, as SHWs left emphasizing the disparity in the number of medical workers per capita alongside the heavier burden of disease. In the 1970s, because of growing concerns over uneven flows and development, the WHO mounted a path-breaking study by Alfonso Mejia and others of migration from some 40 countries. Then, as now, the migration of SHWs was of greater concern than other skilled international migration flows, and the idea of a brain drain largely emerged from analysis of migratory health workers.

After a period of quiescence demand for SHWs in developed countries again increased in the 1990s, resulting from aging populations, growing demand and ability to pay, inadequate training programs, and high attrition rates (for reasons ranging from patient violence to discontent with working conditions, etc.), as jobs in the health sector were seen in many developed countries as too demanding, poorly paid, and lowly regarded (in line with reduced public sector funding, and disregard for the public sector). Reduced recruitment of health workers also followed declining birth rates in developed countries: There were fewer young people and more diverse employment opportunities for women, many with superior wages and working conditions, and greater prestige and respect. Significantly, these influences are similar to the reasons for attrition and migration in source countries.

Contemporary international recruitment of health workers is increasingly global. Where, a quarter of a century ago, it was mainly a movement from a few developing countries to a small number of developed countries, most countries are now involved. New movements of nurses occur between relatively developed countries, notably within the EU. Ireland, once an exporter of SHWs, has become an active recruiter. The new complexity of international migration is evident in Poland, as much a sending country as a recipient, where its source countries are eastern European countries (Ukraine, Belarus, Russia, and Lithuania) and the Middle East (Syria, Yemen, and Iraq), although Polish nurses migrate westwards. China has entered the market as a supplier of nurses, and its considerable interest in becoming more involved has the potential to profoundly influence the future system.

Over the past 30 years, the key receiving countries have remained remarkably similar, dominated by the UK and the USA. Whereas demand in the Gulf has stabilized, other European and global destinations (including Canada and Australasia) have grown in importance. Despite policies of localization, the Gulf states still employ 20 000 migrant doctors, and many more nurses, mostly from south Asia, but also from neighboring and poorer Middle Eastern states such as Egypt and Palestine.

In most developed countries, the proportion of foreign-trained medical workers in the health workforce has usually risen slowly: for example, in the USA and the UK, foreign doctors now represent approximately 27% and 33%, respectively, of their medical workforces; similar percentages occur in Australia and New Zealand, whereas comparable estimates are approximately 7% for Germany and France. Other OECD countries have become significant recipients. Hitherto Japan, virtually only one of the countries that have experienced substantial postwar economic growth and aging populations, has largely managed its health services without resorting to overseas workers, but has recently entered into agreements with the Philippines.

Throughout this time the Philippines has remained the main global source of SHWs for almost every part of the world, alongside India. Sub-Saharan Africa has emerged as a major supplier, and a major source of concern. Relatively recently other Asian states have become sources of SHWs, whereas much smaller Caribbean and Pacific states have become sources. Eastern Europe supplies Western Europe, whereas Latin America has tended to experience proportionately less emigration, though Latin America nurses have moved north to the USA and Europe, especially Spain.

Patterns of health worker migration from sources of supply such as sub-Saharan Africa have also changed. In the 1970s, SHWs were from a relatively small number of African countries (the larger states of South Africa, Nigeria, and Ghana) and predominantly went to a few developed countries outside Africa. Subsequently migration has become much more complex, involving almost all sub-Saharan countries, including intraregional and stepwise movement (e.g., from the Democratic Republic of Congo to Kenya, and from Kenya to South Africa, Namibia, and Botswana), because of targeted recruitment, by both agencies and governments, as much as individual volition. Globally, the 20 countries with the greatest emigration factors in the mid-2000s (the ratio of emigrant

to resident doctors) included 6 in Africa (Ghana, South Africa, Ethiopia, Uganda, Nigeria, and Sudan), 3 in the Caribbean (Jamaica, Haiti, and the Dominican Republic), the Philippines, India, and Pakistan, a cluster of countries perhaps best characterized by crisis (Sri Lanka, Myanmar, Lebanon, Iraq, and Syria), and also New Zealand, Ireland, Malta, and Canada. Migration is now shaped by both market forces and cultural ties, and deeply embedded in uneven global development.

The greater complexity of migration is evident in the interlocking chains of recruitment and supply, some of which were in place 30 years ago. Canada recruits from South Africa (which recruits from Cuba), as it supplies the USA. Kenyan nurses first went to southern African countries such as Botswana, Zimbabwe, and South Africa, and then moved on as 'step migration' to Britain. Something of a hierarchy of global migration – the global care chain – links the poorest sub-Saharan, Asian, and island microstates, to the developed world, culminating in the USA. New transport technology and reduced costs have produced variants of 'commuter migration' with SHWs taking on brief assignments elsewhere.

Migration is constantly in flux depending on labor markets, domestic pressures, evolving global legislation and codes of practice, and individual perceptions of amenable destinations. Migration links languages, training institutions, educational regimes, often in the context of other migration flows, sometimes characterized as chain migration in the context of a 'transnational corporation of kin.' Language proficiency is more crucial in the health sector than in any other arena of migration, skilled or unskilled. Although recruitment has crossed new borders, as trade barriers have disappeared and the Internet become accessible, potential migrants are also more likely to be informed about global job opportunities and be in some position to choose more widely than hitherto. Migration ranges from fixed-term contract migration (typified by that from the Philippines to the Middle East), usually negotiated between governments, and more personal, individual migration that may last a lifetime.

Rationales for Migration

Migration is primarily a response to global uneven development, usually explained in terms of such factors as low wages, few incentives, or poor social and working conditions. Poor promotion possibilities, inadequate management support, heavy workloads, and limited access to good technology including medicines have been widely recognized as 'push factors.' Such pressures are intensified in rural areas, where health workers feel they and their institutions are too often ignored, victims of institutionalized urban bias in development. Cultural factors have emphasized some migration flows. Tamil doctors have been more likely than majority Sinhalese to migrate from Sri Lanka for more than 30 years. Recruitment, by both agencies and governments, has played a critical facilitating role. However, all these various, specific factors are embedded in the broader context of social and economic life, family structures, and histories and broader cultural and political contexts.

Consequently, migration of SHWs occurs for many reasons, despite remarkable uniformity across quite different regions and contexts. Reasons include incomes, job satisfaction, and career opportunities, alongside social, political, and family reasons. The last of these factors, though often neglected, is particularly important since few migrants make decisions as individuals, but are linked to extended families and wider kinship groups. The migration of SHWs is rarely unique but exists within the context of wider migration flows. This is evidently so in India, the Philippines, and most small island states, like those of the Caribbean and Pacific, where there have been steady and diverse migration streams for several decades. In such circumstances, there is effectively a 'culture of migration' where most individuals at least contemplate migration at some time in their lives.

Yet migration is usually constrained in certain ways. Even for those with skills it is rarely easy to cross political boundaries. Where political circumstances have changed, as in the expansion of the EU, migration from poorer eastern states to those in western Europe quickly became substantial. Violence, coups, crime, warfare, and persistent social unrest have predictably hastened migration from countries such as Zimbabwe, Fiji, and Lebanon.

Intention to migrate may occur even before entry into the health system. In the Philippines, at least some people sought to become nurses, partly and sometimes primarily, because that provided an obvious means of international migration. By the end of the 1980s, a medical degree at the Fiji School of Medicine was widely seen as a 'passport to prosperity' and in Kerala (India) a nursing diploma is considered an 'actual passport for emigration' thereby raising the status of nursing. Specific careers may be chosen that optimize migration opportunities; in the Philippines and Pakistan, male doctors have retrained as nurses, and fewer people choose a medical career, as nurses have superior migration opportunities. The initial overseas destination may not be the intended final destination, especially for health workers in the Gulf, who seek to move on to the USA. Migration is not solely of SHWs; for some SHWs a career in health is seen as a way to move the whole family. This step migration points to the challenges in source countries of trying to develop an effective national workforce, when substantial proportions of those being trained may migrate.

Health workers have not usually entered the profession solely for income benefits, but also out of some desire to serve and be of value in the community. However, such feelings do not sustain a career, as workers become frustrated by low pay and poor (or biased) promotion prospects, especially in remote areas. As, increasingly, people do join the health sector for economic reasons, migration becomes even more likely. Income differentials are therefore invariably key factors in migration, as they are in decisions to join or later leave the health profession. Many decisions are simply rationalized in this way, since income differences between countries are often increasingly evident. Income differences are often such that even significant wage increases have had little effect on reducing the extent of migration. Econometric studies, at least for the Pacific island states, have shown that migration demonstrates considerable sensitivity to income differences, but complicated by the structure of household incomes.

In countries where there have not been specific surveys of migration, anecdotal evidence and, in some cases, the rationale for strikes by health workers, emphasize the significance of wage and salary issues. Similarly, the general movement of doctors, dentists, and others from the public to the private sector marks the quest for better incomes and conditions.

Income is firmly linked to the structure of careers and promotion, which many health workers see as being more about 'who you know than what you know' – nepotism and favoritism – and longevity in the system, rather than ability. SHWs have been critical of the lack of a transparent career structure, preferring to move to a meritocracy where skills and accomplishments will be rewarded. Where health workers are stationed outside the main national urban center, the perception that they are being ignored for promotion is even stronger as many consider themselves to be 'out of sight and out of mind.' Inadequate opportunities for promotion constitute not only an incentive to migration, but a constraint to productivity and innovation in the health system.

After income, the actual conditions of employment are influential for migration. Migrants, and potential migrants, frequently complain about the work environment in terms of insufficient support, through inadequate management (lack of team work, poor leadership and motivation, limited autonomy and support, and little recognition and access to promotion and training opportunities) or through the outcome of poor 'housekeeping' (limited access to functioning equipment and supplies). A desire to acquire further training and gain extra experience is a key factor influencing migration. Long hours of overtime, double shifts, working on the early morning 'graveyard' shift or on weekends, especially when these do not receive proper income supplementation, further influence migration. Shift work is a universal source of complaint, and particularly so in more remote places, where fewer staff are available and pressures on those remaining are greater. Inadequate working conditions may also entail the risk of contracting disease. The rise of HIV/AIDS made the nursing profession especially much less attractive than hitherto and, notably in Africa, created a more difficult working climate as the workload increased.

In several developing countries economic restructuring, sometimes externally imposed by international agencies, has led to reductions in the size of the public sector workforce and restrictions on the hiring of new workers. Changes in the health sector take place in a wider context where negative balances of payments and high levels of debt servicing place huge resource constraints on many developing countries. This has sometimes meant the deterioration of working conditions rather than the greater efficiency it was intended to encourage. Ironically, in the mid-2000s, in Kenya, for example, though half of all nursing positions were unfilled, a third of all Kenyan nurses are unemployed, as International Monetary Fund pressure encouraged national wage restraint. In several countries lack of resources, or alternative priorities, has resulted in low wages and poor conditions, with simultaneous vacancies, unemployment, and migration.

Many migrants have left rural areas to take advantage of superior urban and international educational, social, and employment opportunities. These factors reinforce each other, especially in the health sector. The widespread education bias enables young and skilled migrants, with fewer local ties, to migrate more easily. Most nurses, and many other SHWs, are women and may face particular constraints related to partners' careers and family obligations, which may make remote postings and overseas migration difficult. Consequently, the most likely migrants are young single workers followed by married workers without children. In contrast, Indian nurses from Kerala have migrated because their ability to earn and retain significant incomes gave them high status and the consequent ability to find high-status partners in the 'matrimonial market.' In many contexts, gender relations have been restructured following migration. Social ties may result in pressure to migrate, to support the extended family, but may sometimes make migration more difficult to achieve.

Recruitment

Developed destination countries offer real alternatives to political and economic insecurity in many source countries. A high standard of living with higher wages, better career prospects, good education, and a future for children are offered in recruitment campaigns, and often verified by those migrants established overseas. The structure of migration has become increasingly privatized through the expansion of recruitment agencies, and their regular use by recipient countries and by particular hospitals. Recruitment has existed since the 1940s but grew rapidly around the turn of the century. Irrespective of any existing intent to migrate, active recruitment has put growing pressure on, and impressive opportunities in front of, potential migrants. Recruitment agencies smooth the way in attending to bureaucratic issues, satisfying concerns over distant and different countries and cultures, and sometimes providing their own induction training in destinations.

Little information exists on the operations of recruitment agencies, and therefore there is no evidence on whether they exaggerate the potential of overseas employment, although they increase its probability. Recruitment has been particularly significant in sub-Saharan Africa, though there, as elsewhere, it would not have been successful unless other reasons for migration existed. In the early 2000s, half of all overseas nurses in Britain were there because they had been recruited. Recruitment has significantly extended migration beyond its postcolonial routes, for example, taking Chinese nurses to the Gulf and Fijian nurses to the Bahamas and the United Arab Emirates.

Recruitment is competitive, resulting in 'selective depletion' of the more qualified workers from several countries. In recruiting health workers for the UK many agencies engaged in some forms of exploitation. Both in source and recipient countries agencies operate beyond the extent of effective regulation. Such issues resulted in regional attempts to construct and use codes of practice for ethical recruitment, spearheaded by the Commonwealth Secretariat for former British colonies, thus covering significant parts of the Caribbean, Pacific, and sub-Saharan Africa.

The finalization by WHO of a Global Code in 2010 emphasized continued migration concerns and universal agreement to mitigate its harmful effects, notably that migration

did not disrupt health services in source countries. However, migration is a human right and occurs in contexts that do not necessarily involve health issues; there are no incentives for recipient countries and agencies to be involved in ethical international recruitment and all codes are voluntary which limits their impact. Recruitment and migration are both likely to continue.

Consequences of Migration

The trade in, and migration of, SHWs has diverse impacts, from more obvious effects on the delivery of health services and the economic consequences of the loss of locally trained skilled workers, to more subtle social, political, and cultural impacts. Migrants tend to be relatively young and recently trained, compared with those who stay. Many leave after relatively short periods of work, but long enough to gain important practical experience. They often include the best and the brightest. Because migrants move to improve their own and their families' livelihoods, they are usually the key beneficiaries of migration. Recipient countries benefit from having workers who fill shortages in the healthcare system. Conversely, sending countries and their populations, especially in remote areas, lose valuable skills unless those skills are an 'overflow' or are otherwise compensated for.

Healthcare Provision

Migration affects the provision of health care both in quality and quantity. Links exist between migration and the reduced performance of healthcare systems, though actual correlations between emigration and malfunctioning healthcare systems are difficult to make, because it is impossible to quantify what is not there. However, India and the Philippines, both long-term providers of migratory health workers, in circumstances initially described as an overflow, now appear to have become negatively affected, whereas sub-Saharan Africa and many small states experience critical problems, but not simply or even primarily because of migration.

In some circumstances, the quantitative outcome of migration is obvious. In Malawi, the loss of many nurses to the UK in early 2000s brought the near collapse of maternity services even in Malawi's central hospitals, with 65% of nursing positions being vacant. Maternal health care has been similarly affected in Gambia and Malawi with increased workloads, waiting and consultation times, and poorer infection control. In Jamaica, wards have been closed, male and female units have been merged raising cultural issues, and immunization coverage and *in situ* training have both been declined. Although such data are fragmentary, and often depict worst-case scenarios reported in the media, and are not solely the outcome of international migration, they point to difficult circumstances.

Reduced staff numbers mean that workloads of those remaining become higher, and less likely to be accomplished successfully. Many anecdotal reports emphasize longer waiting times with the implication that this raises opportunity costs of medical care, and may also result in medical attention coming too late. In Zimbabwe, in the 2000s, over a quarter of health workers believed that longer waiting times, and shorter opening times, had resulted in unnecessary deaths that prompt attention could have prevented. Foreign aid programs expanded in sub-Saharan Africa in the mid-2000s, to provide drugs to millions affected by tuberculosis and AIDS, yet were hard to implement because too few nurses existed to administer them effectively.

A further consequence of health worker migration is that of some patients traveling overseas for health care, as part of the growing phenomenon of medical tourism. Where such referrals are paid by the state, the cost is considerable. Even where they are not, as is usually the case, resources are nevertheless transferred overseas. In several African countries, referrals have increased at the same time as health worker migration, resulting in an unprecedented increase in the expense of care to fewer people and in the use of foreign currency, which could have been used for other development programs or for the motivation and retention of the country's health workers. The lack of health personnel may not always be the primary motivation for traveling overseas for treatment, but it nonetheless represents a substantial loss of scarce resources, especially because some of the source countries of medical tourist are impoverished nations such as Yemen. Even in countries that are relatively well supplied with health personnel, the cost of referrals is considerable, making the task of financing local health systems and organizing more labor-intensive preventive health care more difficult.

Rural and Regional Issues

The impact of emigration is usually most evident in remote regions, where losses tend to be greater (and where resources were initially least adequate), and has therefore fallen particularly on the rural poor (and sometimes therefore on cultural minorities) who are most dependent on public health systems, and where health needs are often greater, further emphasizing urban bias and the 'inverse care law.' The impact of emigration is complicated and compounded by ubiquitous internal migration, and a parallel movement from the public to the private sector. The movement of SHWs to the private sector has disadvantaged the poor, most of whom cannot afford higher private sector costs, alongside growing evidence of less adequate public sector services. This is poorly documented and it is primarily the evidence of inadequate stocks of health workers in the regions, and very different staff: patient ratios, which suggests the extent of adequate provision and migration (and attrition) in remote areas. The WHO has developed distinct strategies for developing and stabilizing regional workforces. Internal migration exhibits a similar rationale to international migration, but poses distinct problems where the internal migration is of those with particular skills, such as radiologists or pharmacists, and where few are required; hence the loss of even a small number may be crucial.

The Economics of Migration

Training SHWs is costly because of the long duration and high costs and is a burden on relatively poor states, whether directly

or through overseas scholarship provision. When trained workers migrate and the process is repeated, costs mount further. However, there have been few estimates of the costs of the ensuing brain drain, or the possible gain in skills through return migration, and a variety of methodologies and conclusions. The impact on healthcare provision of the emigration of doctors may be remarkably slight, compared with that of nurses, who provide the bulk of health care in many places, and especially in regional areas, where needs are considerable, but not necessarily complex.

A series of estimates of training costs suggest that low-income African countries subsidize high-income countries by as much as $500 million a year through the migration of SHWs, whereas equally fragmentary data from developed countries indicate considerable cost savings involved in hiring overseas-trained SHWs rather than training locally. This has been described as a perverse and unjust subsidy from relatively poor countries to relatively rich ones. These estimates are based solely on the costs of training rather than additional costs based on foregone health care, lost productivity, the under use of medical facilities, etc. However, they usually ignore possible remittances and their consequences. Where the remittances of health workers have been calculated, as in the Pacific island states, they are substantially above training costs, though they flow into the private sector rather than the public sector where most training takes place, and make no contribution to equitable human development.

Where return migration of SHWs occurs, the relationship between income losses, return, and the acquisition of human capital becomes more complex. Return migration of SHWs is relatively limited in many countries; however, if migrants return from overseas, with enhanced skills, knowledge, experience, and enthusiasm (and perhaps also some capital), there can be major gains from migration, including a positive transfer of technical knowledge. However, significant return migration fails to occur for the same reason that migration occurs: Migrants are less likely to be tempted back by a system they left because of its perceived failings. The overall number of return migrant nurses and doctors is modest, and many return because of perceived benefits, such as business opportunities, outside the healthcare system.

A further outcome of migration can be a skill loss when migrants with specific skills do not use them, which may result from failure to recognize qualifications, discrimination, or a preference for jobs with better wages and conditions. The most significant skill loss comes where nurses are employed as caregivers in nursing homes rather than working in hospitals. Expensive training is largely wasted and neither health systems, the migrants, nor their kin at home, who wait for remittances, make real gains.

Social Costs

The social costs attached to the migration of SHWs are complex but often considerable, especially where women move as individuals, leaving families at home. Many migrant workers, especially women within and outside the health sector, experience deprivation and discrimination. Recruitment agencies may impose unforeseen costs, and SHWs experience difficult circumstances, especially where cultures differ from those at home. Numerous examples exist of their experiencing racism in developed countries, and being ignored or experiencing reprisals when complaining of such problems, alongside being denied parity with local workers, promotion, or wage gains.

Health workers are often recruited for, and directed into, positions and locations that are unattractive to local health workers, and peripheral geographical placement is common. Consequently, new migrants are unlikely to be involved in specialist activities despite previous experience, and are most likely, at least initially, to be in the least attractive fields of health care and in outlying parts of the country, and with limited autonomy and authority. Stresses may occur for the families of migrants. Children may have to make complex adjustments to parental absences, and experience what has been called a 'care deficit.' However, migrants and their families usually gain in status through the material benefits of migration.

Migration of SHWs has made it necessary for less or non-qualified people, such as nurses' aides, to perform tasks that are normally beyond their training. This poses risks of incorrect diagnoses and inappropriate treatment. Patients have also reverted to the informal sector with sometimes costly, uncertain, and ineffective outcomes. In many countries, migrant nationals have been replaced by other international migrants, as part of the cascading global care chain, though the direct economic costs may be considerable (in both recruitment and salaries) and they may be less effective because of language and cultural differences, which restrict their ability to provide health services, contribute to training, and enable sustainability.

The Future Global Healthcare Chain

Shortages of SHWs exist in most countries in the world, and have been remedied mainly by migration from poorer countries rather than by strategies for improved retention and recruitment, hence the development of a Global Code of recruitment by WHO to encourage a more regulated migration, bilateral reciprocity, and greater international cooperation. Countries such as India and the Philippines, that previously exported an 'overspill,' have experienced some adverse effects from their 'export policies.' Migration has been problematic for relatively poor countries as the costs of mobility are unevenly shared, and the care chain becomes more global and hierarchical. Greater complexity increases the challenge of achieving more equitable outcomes.

An open international market is said to offer efficiency and economic gains. However, gains in economic efficiency tend to be localized in receiving countries and, as the evidence of costs to national health, economic, and social systems has mounted, there has been somewhat greater interest in developing policies to diminish and mitigate the impacts of migration. Nonetheless, international migration is not the main cause of healthcare shortages in developing countries, nor would a significant reduction in emigration remove human resource problems.

The onus for a more equitable global distribution of SHWs has gradually shifted toward recipient countries, where demand occurs. Few recipient countries have taken effective measures to increase recruitment and reduce attrition of SHWs, at a time of greater demand, either by increasing the number of training places or improving wages and working conditions. Continued migration has thus led to be renewed calls for ethical recruitment guidelines, adequate codes of practice binding countries, and/or compensation for countries experiencing losses; yet compensation is inherently implausible and impractical, although ethical arguments confront political realities. Better regulation, and more ethical recruitment, alongside bilateral relationships suggest some partial solutions, in terms of more effective managed migration.

The principal occupational flows of SHWs are primarily of nurses, where the evidence of losses in developing countries is substantial; however, there are more poorly documented flows of all cadres of health workers, such as radiologists and pharmacists. Failures of governance, broadly the inadequate delivery of services, whether health or education, and weak or nonexistent political will, constrain the development and retention of national workforces. Various possibilities exist for more effective production and retention of SHWs, ranging from diverse financial incentives (inside and outside the health system), strengthening work autonomy, and improving the status of health workers, increasing recruitment capacity, introducing intermediate categories of workers, such as nurse practitioners, and ensuring an effective 'fiscal space' for health services, but only rarely have these been effectively implemented in a concerted manner.

The international migration of SHWs has increased because perceptions of inadequate local conditions have grown, diaspora 'host' populations are generally increasing in destination states, demand has increased and recruitment intensified, and because health skills are valuable commodities in international migration. Yet paradoxically almost everywhere fewer people are being attracted to health careers. Wages and conditions are increasingly seen as deterrents to entry as other sectors become more attractive. Potential employees witness the frustrations of health workers and there is a wider range of job options. In both developed and developing countries, careers in health are now less attractive, other than as a means to migration.

Sending countries have not always been able to discourage migration, which is widely perceived as a human right. Indeed, several remittance-dependent countries, such as Cape Verde, the Philippines, and Kiribati, have not challenged migration but nurtured it because of its economic role. Unions have supported the rights of members to better their circumstances by migration, while also pressing governments to act locally to improve working conditions. Migration is increasingly embedded in national and international political economies. It is more resilient to cyclical downturns than other sectors. Few recipient countries have taken realistic and effective steps to increase national market supply, and any solution requires multilateral consensus rather than a national or bilateral approach. Migration of SHWs, and its complex consequences, will probably continue.

See also: International E-Health and National Health Care Systems. International Movement of Capital in Health Services. Medical Tourism

Further Reading

Bach, S. (2008). International mobility of health professionals: Brain drain or brain exchange? In Solimano, A. (ed.) *The international mobility of talent*, pp 202–235. Oxford: Oxford University Press.

Brown, R. and Connell, J. (2004). The migration of doctors and nurses from South Pacific island nations. *Social Science and Medicine* **58**(11), 2193–2210.

Clark, P., Stewart, J. and Clark, D. (2006). The globalization of the labour market for health-care professionals. *International Labour Review* **145**, 37–64.

Connell, J. (2008). *The global health care chain: From the Pacific to the World*. New York: Routledge.

Connell, J. (2010). *Migration and the Globalisation of Health Care*. Cheltenham: Edward Elgar.

Connell, J. and Buchan, J. (2011). The impossible dream? Codes of practice and the international migration of skilled health workers. *World Medical and Health Policy* **3**(3), 1–17.

Connell, J., Zurn, P., Stilwell, B., Awases, M. and Braichet, J-M. (2007). Sub-Saharan Africa: Beyond the health worker migration crisis? *Social Science and Medicine* **64**, 1876–1891.

Gish, O. (1971). *Doctor Migration and World Health*. London: Bell.

Ho, C. (2008). Chinese nurses in Australia: Migration, work and identity. In Connell, J. (ed.) *The International Migration of Health Workers*, pp 147–162. London: Routledge.

Kingma, M. (2006). *Nurses on the move. Migration and the global health care economy*. Ithaca: Cornell University Press.

Mackintosh, M., Mensah, K., Henry, L. and Rowson, M. (2006). Aid, restitution and international fiscal redistribution in health care: Implications of health professionals' migration. *Journal of International Development* **18**, 757–770.

Mejia, A., Pizurski, H. and Royston, E. (1979). *Physician and Nurse Migration: Analysis and Policy Implications*. Geneva: WHO.

Percot, M. and Rajan, S. (2007). Female emigration from India. Case study of nurses. *Economic and Political Weekly* **42**, 318–325.

Vujicic, M. and Zurn, P. (2006). The dynamics of the health labour market. *International Journal of Health Planning and Management* **21**, 101–115.

World Health Organization (2006). *Working together for health*. Geneva: WHO.

Latent Factor and Latent Class Models to Accommodate Heterogeneity, Using Structural Equation

AJ O'Malley, Harvard Medical School, Boston, MA, USA
BH Neelon, Duke University, Durham, NC, USA

Heterogeneity

In statistics and econometrics, heterogeneity typically refers to a random variable, parameter, or distribution that varies across a population of interest. It can involve the mean, variance, or other features of a distribution and may arise from observed and unobserved causes.

Observed heterogeneity is variability in an outcome (or dependent variable) attributable to observed predictors (Skrondal and Rabe-Hesketh, 2004). In the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \boldsymbol{\varepsilon}_i \quad [1]$$

heterogeneity in the expected value of the outcome y_i, denoted $E[y_i|x_i]$, is accounted for by the predictor x_i across subjects $i=1,\ldots,n$. The parameter β_1 quantifies the magnitude of heterogeneity. Random variability in y_i that cannot be explained by x_i is denoted by $\boldsymbol{\varepsilon}_i$, the error term, which is assumed to have mean zero and a constant (or homogeneous) variance $\text{var}(y_i|x_i)=\sigma^2$. In parametric modeling, the most common distribution assumed for $\boldsymbol{\varepsilon}_i$ is a normal or Gaussian distribution, which has many appealing features including characterizing the ordinary least squares (OLS) estimation method.

Now consider a model with two predictors and an interaction effect,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \boldsymbol{\varepsilon}_i \quad [2]$$

In eqn [2], the effect of a one-unit change in x_{1i} on $E[y_i|x_i]$ is $\beta_1 + \beta_3 x_{2i}$, illustrating that the effect of x_{1i} depends on x_{2i}. A consequence of effect heterogeneity is that any statement of the effect of x_{1i} must be accompanied by the value(s) of x_{2i} at which it is computed and vice-versa. If x_{2i} is not observed and the model in eqn [1] is estimated, the OLS estimate of β_1 is a weighted average of the true heterogeneous effects of x_{1i} with respect to the likelihood of each value of x_{2i} (Angrist, 1998). When x_{1i} and x_{2i} are uncorrelated by design (e.g., x_{1i} is assigned at random in a randomized trial), the OLS estimate under eqn [1] corresponds to the average effect of x_{1i} over the individuals in the sample, otherwise being more difficult to interpret.

Other forms of heterogeneity are accommodated by relaxing the assumptions of the linear regression model, yielding a wider array of models and possibly requiring specialized estimation methods. For example, if $\text{var}(y_i|x_i)$ depends on x_i (directly or via $E[y_i|x_i]$), the assumption of equal variance at all values of the predictors required by OLS is violated. This phenomenon, referred to as heteroscedasticity, may be accommodated in the context of OLS by dividing y_i and x_i by the standard deviation of the residuals, $\text{var}(y_i|x_i)^{0.5}$, and then applying OLS (weighted least squares). If $\text{var}(y_i|x_i)$ is known, the process is straightforward, otherwise $\text{var}(y_i|x_i)$ must be estimated. Point estimates of β that do not account for heteroscedasticity are estimated imprecisely whereas confidence intervals (frequentist inference) and credible intervals (Bayesian inference) are likely to be incorrectly calibrated. When the objective is to estimate a tail probability or quantile (e.g., in immunoassays seeking to determine whether the concentration of a substance in blood serum exceeds a critical threshold), estimation of the variance function is key. Substantial progress on variance function estimation methods has been made in the context of analyzing assays (Davidian *et al.*, 1988; O'Malley *et al.*, 2008).

Unobserved Heterogeneity and Measurement Error

Unobserved heterogeneity, the variability in y_i arising from unobserved sources, cannot be accommodated without much difficulty as direct adjustment for the cause of the heterogeneity is not possible. To illustrate the difficulties that may arise from unobserved heterogeneity, suppose that to relate an individual's health, y_i, to his/her intelligence quotient (IQ), u_i, but in lieu of u_i, the educational attainment, x_i, is observed. Because u_i is unobserved directly but essential to the model, it is referred to as a latent variable. The situation is represented by the following equations:

$$\begin{aligned} &\text{health:} \quad y_i = \beta_0 + \beta_1 u_i + \boldsymbol{\varepsilon}_i \\ &\text{education:} \quad x_i = u_i + \delta_i \end{aligned} \quad [3]$$

where, by assumption u_i is unrelated to $(\boldsymbol{\varepsilon}_i, \delta_i)^T$, and δ_i is unrelated to $\boldsymbol{\varepsilon}_i$. Equation [3] is a classical measurement error model (Carroll *et al.*, 1995). The observed data regression,

$$y_i = \beta_0 + \beta_1 x_i + \tilde{\boldsymbol{\varepsilon}}_i \quad [4]$$

is problematic because x_i is correlated with the error, $\tilde{\boldsymbol{\varepsilon}}_i = \boldsymbol{\varepsilon}_i - \beta_1 \delta_i$, in violation of the OLS assumption that the predictors are unrelated to the errors. Here x_i is said to be endogeneous. It can be shown that the quantity being estimated by applying OLS to eqn [4] is $\rho\beta_1$, where $\rho = \text{var}(u_i)/(\text{var}(u_i) + \text{var}(\delta_i)) < 1$ is the attenuation factor (Bedeian *et al.*, 1997). Thus, if the heterogeneity in x_i, arising from δ_i is ignored, the estimated coefficient of x_i will be an inconsistent estimator of β_1.

An alternative model arises when u_i varies according to x_i; i.e., $u_i = x_i + \delta_i$, where x_i is independent of δ_i. For example, x_i is the setting of a machine, the control variable, and u_i is the actual level at which the machine operates. This situation, known as Berkson measurement error (Berkson, 1950) is less problematic, at least in linear models, because the OLS estimate of β_1 under eqn [4] is unbiased, and so the only consequence of this form of measurement error is that $\text{var}(\tilde{\boldsymbol{\varepsilon}}_i) \geq \text{var}(\boldsymbol{\varepsilon}_i)$, which leads to a reduction in statistical power.

Returning to the classical measurement error model, the availability of replicate observations on x_i allows u_i and hence var(δ_i) to be identified, thus enabling an estimate of $\rho\beta_1$ to be decomposed into estimates of ρ and β_1. If replicate observations are not feasible or available, an instrumental variable (IV) – a variable z_i that is related to u_i conditional on u_i being unrelated to γ_i – facilitates estimation. In the case of linear relationships, the first condition for z_i to be an IV implies $u_i = \theta_0 + \theta_1 z_i + \gamma_i$ with $\theta_1 \neq 0$ and u_i being uncorrelated with the random error γ_i. Substituting for u_i in eqn [3] yields:

$$\gamma_i = \tilde{\beta}_0 + \tilde{\beta}_1 z_i + \tilde{\varepsilon}_i \quad [5]$$

$$x_i = \theta_0 + \theta_1 z_i + \tilde{\delta}_i \quad [6]$$

where $\tilde{\beta}_0 = \beta_0 + \beta_1\theta_0$, $\tilde{\beta}_1 = \beta_1\theta_1$, $\tilde{\varepsilon}_i = \beta_1\gamma_i + \varepsilon_i$ and $\tilde{\delta}_i = \theta_1\gamma_i + \delta_i$. Under the second IV condition, z_i is uncorrelated with the errors $(\tilde{\varepsilon}_i, \tilde{\delta}_i)^T$, ensuring that OLS yields unbiased estimates of the parameters in eqns [5] and [6]. Hence, consistent estimates of β_0 and β_1 can be deduced from the relations $\beta_0 = \tilde{\beta}_0 - (\tilde{\beta}_1/\theta_1)\theta_0$ and $\beta_1 = \tilde{\beta}_1/\theta_1$. Alternatively, one may use two-stage least squares (2SLS): apply OLS to eqn [6] and compute predicted values of x_i, denoted $\hat{x}_i$, then apply OLS to eqn [4] but with x_i replaced by $\hat{x}_i$. The impact of measurement error is a decreasing function of the fraction of variation in x_i is explained by z_i. The readers are referred to the article on instrumental variables and to the econometric text by Wooldridge (2002) and that by Angrist and Pischke (2009) for further discussion of IVs, 2SLS, and related methods.

Classic Structural Equation Models

Broadly speaking, a structural equation model (SEM) is a model involving relationships between latent variables. Latent variables generally represent true values of a variable and so relationships between them are often considered to be truisms or causal (Lee, 2007). The use of SEMs to estimate causal relationships has a long history (Pearl, 2000). Latent variables must have associated observed (or manifest) variables in order to identify the model. Traditionally, an SEM is characterized by continuous-valued observed (or manifest) variables, continuous-valued latent (or unobserved) variables, and linear relationships among the latent variables. The linear SEM has the form

$$\mathbf{y}_i = \mathbf{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad [7]$$

$$\mathbf{x}_i = \mathbf{\Lambda}_x \boldsymbol{\mu}_i + \boldsymbol{\delta}_i \quad [8]$$

$$\boldsymbol{\eta}_i = \mathbf{A}\boldsymbol{\eta}_i + \mathbf{B}\boldsymbol{\mu}_i + \boldsymbol{\upsilon}_i \quad [9]$$

where $\boldsymbol{\varepsilon}_i$, $\boldsymbol{\delta}_i$, and $\boldsymbol{\upsilon}_i$ are mutually independent error terms with zero means and constant covariance matrices (Jöreskog, 1973). Equations [7] and [8] are measurement models relating the observed variables, $\mathbf{y}_i$ and $\mathbf{x}_i$, to their latent counterparts, $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$, whereas eqn [9] contains the structural model relating the latent construct $\boldsymbol{\mu}_i$ to the latent construct $\boldsymbol{\eta}_i$. Here $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$ and $\mathbf{B}$ are matrices of regression coefficients whereas $\mathbf{A}$ is a matrix of parameters that affects both the mean and covariance of $\boldsymbol{\eta}_i$. The involvement of $\boldsymbol{\eta}_i$ on both sides of eqn [9] allows for direct relationships between its elements, inducing correlations between them and imposing correlation structure on $\boldsymbol{\eta}_i$ and thus $\mathbf{y}_i$.

The measurement error model in eqn [3] is a special case of an SEM in which the effect variable is $\mathbf{y}_i$ (as observed). Therefore, with replicated measurements, the classic measurement error model in eqn [3] corresponds to $\gamma_i = \eta_i$, $\mathbf{A} = 0$, $\mathbf{B} = \beta_1$, and $\boldsymbol{\mu}_i = \mathbf{1}u_i$, i.e., eqns [8] and [9] reduce to $\mathbf{x}_i = \mathbf{1}u_i + \boldsymbol{\delta}_i$ and $\mathbf{y}_i = \beta_1\mathbf{1}u_i + \boldsymbol{\varepsilon}_I$, respectively, where $\mathbf{1}$ denotes a vector of 1's. For model identifiability, the dimensions of $\mathbf{y}_i$ and $\mathbf{x}_i$ must exceed those of $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$ respectively; the larger the differences, the better. The regression models in eqns [1] and [2] are simple cases of SEMs.

SEMs have been used extensively in the social (e.g., economics, sociology) and behavioral (e.g., psychiatry, psychology) fields. For example, in an analysis of the relationship between job satisfaction and organization commitment to job turnover, Williams and Hazer (1986) use a SEM having the exact forms of eqns [7]–[9]. The measurement models relate observed values of the final outcome (job turnover), intermediate outcomes (intention-to-quit, job satisfaction, organizational commitment), and four exogeneous measures of work environment to their true values. The structural model relates the true values of the outcomes (the endogeneous variables) both to outcomes themselves and to the true values of the work environment variables in order to test hypothesized causal models as depicted by a flow diagram. For a thorough description of the traditional SEM, readers are referred to the classic text by Bollen (1989), the manual of the Linear Structural Relationships (LISREL) software package (Jöreskog and Sörbom, 1996), and the recent text by Lee (2007). Modern SEMs extend well beyond linear models, including a wide-range of generalizations of SEMs to outcomes that are not normal (e.g., binary, ordinal, categorical outcomes) (Rabe-Hesketh *et al.*, 2004). Next, models with continuous-latent variables in linear and nonlinear contexts are discussed (see Section 'Latent Factor Models'), following the same trend for discrete-latent variables (see Section Latent Class and Finite Mixture Models').

Latent Factor Models

Exploratory factor analysis (EFA) decomposes the covariance or correlation matrix of the centered values (residuals if the model includes covariates) of a sample of multivariate observations by relating these values to a smaller number of latent variables ('factors') that are interpreted on the basis of their relationships ('loadings') with the observed variables. Among various applications, EFA is used to generate hypotheses with regard to the dimensions underlying the data, to construct summary scales for reporting information, and to eliminate redundant items from questionnaires or survey instruments. The EFA model has the form

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad [10]$$

where $\boldsymbol{\eta}_i \sim N(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{\Psi})$, $\mathbf{0}$ is a vector of m zeros, $\mathbf{I}$ is an $m \times m$ identity matrix, $\mathbf{\Psi}$ is a diagonal matrix, and $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$

are independent vectors of $m<r$ and r random variables, respectively (Johnson and Wichern, 1998). Therefore, $\text{cov}(\mathbf{y}_i,\boldsymbol{\eta}_i)=\boldsymbol{\Lambda}$ and $\text{var}(\mathbf{y}_i)=\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi}$.

Ambiguity arises with factor decompositions when $m>1$ as the model is not identified by the data. To illustrate, let $\mathbf{T}$ be an $m\times m$ orthonormal matrix; i.e., $\mathbf{T}\mathbf{T}^T=\mathbf{I}$. Then $\boldsymbol{\Lambda}\boldsymbol{\eta}_i=\boldsymbol{\Lambda}\mathbf{T}\mathbf{T}^T\boldsymbol{\eta}_i=\boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^*$, where $\boldsymbol{\Lambda}^*=\boldsymbol{\Lambda}\mathbf{T}$ and $\boldsymbol{\eta}^*=\mathbf{T}^T\boldsymbol{\eta}_i$, illustrating that the factor loadings can be 'rotated' using an orthonormal basis without changing the fitted values of $\mathbf{y}_i$ in eqn [10]. In practice, factor rotation is useful as it provides a means to obtain more interpretable factor loadings. For example, the commonly used factor rotation procedure Varimax seeks to split the factor loadings into two groups, the elements of the one tending toward zero, and the elements of the other toward unity, thereby making it easier to align variables with factors. In an analysis of Joint Committee data on the Accreditation of Health Care Organizations in the US, a hospital-level EFA with factor rotation was integral to developing two optimal scales (treatment and diagnosis, counseling and prevention) for the quality of hospitals' treatment and care delivered to patients with acute myocardial infarction, congestive heart failure, and pneumonia (Landon *et al.*, 2006).

Latent factor models generalize eqn [10] by allowing the expected value of $\mathbf{y}_i$ to depend on a matrix of covariates $\mathbf{X}_i$ for subject i; i.e.,

$$\mathbf{y}_i=\boldsymbol{\Lambda}\boldsymbol{\eta}_i+\mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i \qquad [11]$$

The latent variables $\boldsymbol{\eta}_i$ are known as latent factors due to the joint dependence of the multiple elements of $\mathbf{y}_i$ on the elements of $\boldsymbol{\eta}_i$. To identify the model, the variances of the latent variables may be set to 1 (typical in EFA), or an element of each row of $\boldsymbol{\Lambda}$ may be set to 1 (typical in latent factor models). The latter anchors the model and makes the variance parameters representative of the strength of the correlation between the outcomes (Skrondal and Rabe-Hesketh, 2004). One of the appealing features of eqn [11] is that model estimation is simplified because the independent assumptions on $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$ imply that the elements of $\mathbf{y}_i$ are conditionally independent given $\boldsymbol{\eta}_i$. Therefore, the distribution of $\mathbf{y}_i$ conditional on $\boldsymbol{\eta}_i$ has the convenient form

$$f(\mathbf{y}_i|\boldsymbol{\eta}_i,\mathbf{X}_i)=\prod_j f(y_{ij}|\boldsymbol{\eta}_i,\mathbf{x}_{ij}) \qquad [12]$$

where $f(y_{ij}|\boldsymbol{\eta}_i,\mathbf{x}_{ij})$ is the probability distribution of $\mathbf{y}_i$ given $\boldsymbol{\eta}_i$. In SEM terminology, latent factor models are measurement models in which the outcomes are directly affected by covariates and jointly dependent on shared latent traits. Because $\boldsymbol{\eta}_i$ is in the model for outcome j ($j=1,\ldots,r$), models that factorize like eqn [12] are referred to as shared-parameter models (Vonesh *et al.*, 2006; Reich and Bandyopadhyay, 2010). In practice, there may be little interest in the factor structure, in which case if $m>1$, the nonuniqueness of the fitted model is a nuisance. A simple uniqueness condition such as $\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}=\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix, may be imposed to identify the model.

Latent factor models are being used increasingly in applications involving complex data and study designs and, therefore, apply to a broader array of settings than EFA. For example, Hogan and Tchernis (2004) used a latent factor to obtain a model-based index of material deprivation at the census tract level in Rhode Island. They supposed that for each area on a map, four manifest variables (standardized to z-scores) are conditionally independent given a one-dimensional latent factor with spatial correlation incorporated through the latent factor. The model was fit using Bayesian methods and the model-based material deprivation index was defined as the posterior expectation of the latent factor given the observed data. A model-based index confers several advantages over ad hoc methods of combining indices into a single score, including optimally weighting the constituent indices and the computation of their inferences.

Hierarchical Models

Latent factor models accommodate clustered data and longitudinal data. To illustrate, the authors have presented the latent factor model in terms of individual observations

$$y_{ij}=\lambda_j^T\boldsymbol{\eta}_i+\mathbf{x}_{ij}^T\boldsymbol{\beta}+\varepsilon_{ij}$$

where j denotes measurement type (ordered the same across subjects), and $\boldsymbol{\lambda}_j^T$ and $\mathbf{x}_{ij}^T$ are the jth rows of $\boldsymbol{\Lambda}$ and $\mathbf{X}_i$ respectively. The random intercept model

$$y_{ij}=\eta_i+\mathbf{x}_{ij}^T\boldsymbol{\beta}+\varepsilon_{ij}$$

where $\eta_i\sim N(0,\tau^2)$ is then seen to be the special case of the latent factor model in which $\lambda_j=1$, a scalar, $j=1,\ldots,m$. The importance of the latent factor is quantified by τ^2; larger variances are indicative of individuals that differ extensively in unmeasured ways (widely varying η_i). The random intercept–random slope model

$$y_{ij}=\mathbf{z}_{ij}^T\boldsymbol{\eta}_i+\mathbf{x}_{ij}^T\boldsymbol{\beta}+\varepsilon_{ij}$$

is a latent factor model with known factor loadings if $\mathbf{z}_{ij}=\mathbf{z}_j$; i.e., the covariates with random coefficients are balanced in that they do not vary across subjects. Balanced covariates for the random effects may arise when each subject is evaluated by the same set of raters (e.g., radiologists evaluating images) or, in a longitudinal setting, when observations are made at regular intervals. Regularly spaced longitudinal data is common. For example, participants in a study are examined weekly; the Federal Reserve sets interest rates quarterly; quarterly financials are released by companies.

In hierarchical models, latent variables can be incorporated at multiple levels to account for correlation within clusters (Raudenbush and Sampson, 1999). For example, in a study of academic achievement, students may be nested within classes and classes are nested within schools. A hierarchical latent factor model contains latent factors at one or more levels of the hierarchical structure.

Multivariate Mixed Outcome Models

A natural extension of the latent factor model is ascribed to situations where the outcome variables have mixed types (Dunson, 2000). Under the decomposition in eqn [12], a

generalized linear model is used to model y_{ij} with the link function

$$h_j(E[y_{ij}|\boldsymbol{\eta}_i, \mathbf{x}_{ij}]) = \boldsymbol{\lambda}_j^T \boldsymbol{\eta}_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} \quad [13]$$

where $h_j(\cdot)$ is a monotone function that maps its argument to an unrestricted random variable. For example, in medical device trials of coronary-artery stents, the key outcomes often include a clinical (binary) and an angiographic (continuous) outcome, leading to one binary and one linear equation (O'Malley *et al.*, 2003). The logit link function was used for the binary component (the probit link would have been an alternative) whereas the linear link (or identity function) was used for the continuous outcome. Other common link functions (associated data types) include the logarithmic link (count data, nonnegative data such as costs or expenditures, time-to-event or survival data) and the log–log link (extreme-event or maximal-outcome data such as in flood prediction).

For discrete-valued (e.g., binary, count) outcomes, additional identifiability constraints are required as the mean and variance are no longer free parameters. However, it is important not to inadvertently restrict the parameter space of the model when imposing identifiability conditions. As a general rule, λ_j should only be constrained if the variance of outcome j is determined by its mean (e.g., as for a binary random variable), in which case setting $\lambda_j = 1$ is appropriate. But when applied, such an identifying constraint can lead to insolvable identifying conditions (e.g., equations that can only be solved by allowing a variance to be negative).

Joint modeling of multiple outcomes may yield more precise results than separate analyses as information on one outcome can be brought to bear on the analysis of another outcome. However, if the sets of covariates depend on the information by which two outcomes are found to be identical, then point estimates are minimally affected, and unaffected for all outcomes in the case of linear models (Teixeira-Pinto and Normand, 2009).

In a related family of models, the covariates may be modeled through the latent variable (Sammel *et al.*, 1997). Therefore, the coefficients of the covariates are proportional across outcomes and thus represent overall associations with the underlying construct generating the data (O'Malley *et al.*, 2003). Although such proportionality makes it easier to summarize the impact of a covariate, it might be too restrictive in many applications. Furthermore, the regression coefficients affect the marginal variance of the outcomes, because of which estimates are more sensitive to the correlation structure than under eqn [13].

Joint Models Involving Censored or Missing Data

In longitudinal analyses where outcomes may be censored due to death, the censoring mechanism is nonignorable (i.e., informative) if unobserved factors are correlated with those outcomes that are correlated with survival. One approach for overcoming this problem is to jointly model the outcome and survival time, conditioning on a latent factor to account for unmeasured common causes (Vonesh *et al.*, 2006).

The above approach may be adapted to account for missing values of the outcome (or potentially of covariates) in other contexts. In general, if the number of distinct missing data patterns across the sample is small (e.g., if the outcome is the only variable ever missing, there are only two missingness patterns), missing data can be modeled using a categorical random variable that is the subject of one (set of) equation(s) whereas the outcome is the subject of another equation, both equations depending on observed predictors and a latent factor (Tsonaka *et al.*, 2009). Shared-parameter models provide one of the few methods applicable when the missing data mechanism is not-missing-at-random (Little and Rubin, 2002). For example, in the case when the outcome is the only variable with missing values, a binary regression equation relates $d_i = I(y_i = \text{missing})$ to the observed covariates and a latent variable whereas a second model relates the nonmissing values of y_i to d_i, the observed covariates and the same latent variable.

In observational studies, latent factor models may be used to account for unmeasured confounders affecting the selection of treatment and the outcome. In place of a censoring or missing data indicator, a problematic (i.e., endogeneous) predictor is modeled in conjunction with the outcome. Therefore, models to account for nonignorable treatment selection emulate SEMs by modeling the relationships between latent variables and, like the measurement error model in the Section 'Unobserved Heterogeneity and Measurement Error', these models involve simultaneous equations. This scenario is expanded in the Section 'Bivariate Probit Type Models'.

Categorical Outcome Variables

When the outcomes under the model in eqn [13] have the same form but are binary (or ordinal) as opposed to continuous, the model reduces to an item response theory (IRT) model. The most common IRT model assumes a single latent factor with r categories, $\mathbf{x}_{ij} = (I(j=1), \ldots, I(j=r))^T$, and a logit link:

$$\text{logit}\{\Pr(y_{ij} = 1|\eta_i, \mathbf{x}_{ij})\} = \gamma_j(\eta_i - \tilde{\beta}_j) \quad [14]$$

where $\tilde{\beta}_j = \beta_j/\gamma_j$. Thus, the model includes an intercept and slope parameter for every measurement type (e.g., a type of test) along with a single underlying latent variable for each subject (e.g., true level of ability). The specification of the model in eqn [14] is completed by assuming η_i is normally distributed. The same form of model is commonly used to model ordinal responses (see article on models for ordered data).

Although eqn [14] generalizes to allow an m-dimensional latent factor, it is rare to have more than two dimensions. If η_i is treated as a fixed effect parameter, then eqn [14] is the well-known Rasch model (Rasch, 1960), often used in education or other situations where multiple informants provide ratings of an individual (Horton *et al.*, 2008).

Bivariate Probit Type Models

An alternative to the method in Section 'Categorical Outcome Variables' for modeling when outcomes are not normal is to define latent continuous variables y_{ij}^* that underlie a discrete-valued y_{ij}. The multivariate normal latent factor model is assumed for y_{ij}^*. The bivariate probit is an example of this type of

model. Let y_{i1} and y_{i2} denote binary realizations of underlying normally distributed random variables y_{i1}^* and y_{i2}^* respectively. The bivariate probit model can then be defined as

$$y_{ij}^* = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + \rho_j \eta_i + \boldsymbol{\varepsilon}_{ij} \quad [15]$$

where $y_{ij} = I(y_{ij}^* > 0)$ for $j = 1,2$ and $\rho_1 = 1$ for identifiability. The latent factor η_i denotes an unmeasured confounding variable and $\rho^* = \rho_2/(1-\rho_2^2)^{0.5}$ is a measure of the extent of confounding (the selection effect) standardized to [–1,1]. In the absence of covariates, ρ^* is the correlation between two continuous random variables that are estimated from observing two binary realizations, commonly referred to as a tetrachoric correlation (Bonett and Price, 2005). A $\rho^* > 0$ indicates that unobserved factors are such that larger values of y_{i2}^* are associated with larger values of y_{i1}^*. Bivariate probit models are often used when observations of a binary outcome are available only for a subset of individuals in a study. For example, in a study of the impact of financial incentives on quality of care delivery by physicians, a quality indicator may be available only for individuals with certain health experiences. Whenever a quality measure for an individual depends on unmeasured factors possibly relating to quality of care, then a bivariate model can be fit to account for nonrandom selection into the sample. The outcome equation is augmented with a latent variable that being also a predictor in an equation, describes the likelihood that individual with certain characteristics is sampled. If the regression equations and the probability distribution of the observations are correct, unbiased estimates of the effects of interest (in this case, physician financial incentives) are obtained.

A generalization of the bivariate probit in eqn [15] yields the family of models developed in Heckman (1978), in which one or both of y_{i1}^* and y_{i2}^* may be observed, y_{i2}^* may be a predictor of y_{i1}^* and vice-versa, and y_{i2} may be a predictor of y_{i1}^* and y_{i2}^*:

$$y_{i1}^* = \alpha_1 y_{i2}^* + \theta_1 y_{i2} + \mathbf{x}_{i1}^T \boldsymbol{\beta}_1 + \eta_i + \boldsymbol{\varepsilon}_{i1} \quad [16]$$

$$y_{i2}^* = \alpha_2 y_{i1}^* + \theta_2 y_{i2} + \mathbf{x}_{i2}^T \boldsymbol{\beta}_2 + \rho_2 \eta_i + \boldsymbol{\varepsilon}_{i2}$$

The model in eqn [16] accommodates both continuously valued and discrete-valued endogeneous variables, the latter being referred to as a structural-shift. In general, for the model to be identifiable, restrictions on the parameters are needed. The special case of $\alpha_1 = \alpha_2 = \theta_2 = 0$ (all predictors observed) is a parametric alternative to nonparametric IV methods when the endogeneous predictor is binary. In addition, if y_{i1}^* is observed but y_{i2}^* is not observed, the Heckit model (Arendt and Holm, 2006) arises. If $\theta_1 = \theta_2 = 0$ and y_{i1}^* and y_{i2}^* are observed, a linear simultaneous equations model is obtained. The article on discrete outcomes includes a detailed review of discrete outcome models with endogeneous predictors.

Latent Class and Finite Mixture Models

In many applications, the study population can be decomposed into a finite number of distinct groups with respect to a variable, y_i. If the variability in the data arises primarily from differences between groups rather than those within groups, the marginal distribution of y_i can be represented by a mixture of distributions of the same parametric form but with unique parameters, $\{\theta_k\}_{1:K}$:

$$p(y_i|\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k(\mathbf{x}_i) p(y_i|\theta_k, \mathbf{x}_i) = \sum_{k=1}^{K} Pr(C_i = k|\mathbf{x}_i) p(y_i|\theta_k, \mathbf{x}_i), \quad i = 1, \ldots, n \quad [17]$$

where $\pi_k(\mathbf{x}_i) = Pr(C_i = k|\mathbf{x}_i)$ is a latent class probability or mixing weight associated with latent class k, and C_i indicates, conditional on subject i having covariates $\mathbf{x}_i$, the subpopulation k ($k = 1, \ldots, K$) to which subject i belongs.

The model in eqn [17] is referred to as a finite mixture model, a latent class model (because it partitions subjects into one of K latent classes), or a discrete-latent variable model (because C_i denotes a latent variable with a finite number of values). When x_{1i} is observed and x_{2i} is a discrete-valued unobserved covariate, the interaction effect model in eqn [2] is a latent class model (i.e., the coefficient of x_{1i} takes on different values across the latent classes that are defined by the unobserved x_{2i}).

Latent class models are useful when the research goal is to cluster patients into distinct subpopulations, or if one believes that the data-generating process can be modeled by first assuming that subjects fall into one of K latent classes; then, conditional on class-k membership, the outcome y_i is drawn from $p(y|\theta_k)$ for subject i. When $\mathbf{x}_i$ only consists of the scalar 1, eqn [17] assumes that the class-membership probabilities are identical for all n subjects – that is, $Pr(C_i = k|\mathbf{x}_i) = \pi_k$ for all i – and the model reduces to the model-based clustering approach of Fraley and Raftery (2002). In general, however, these probabilities vary as a function of subject-level predictors, $\mathbf{x}_i$. In this case, the class-membership probabilities are typically assumed to follow a multinomial logit or multinomial probit model.

As an illustration, consider a hypothetical study in which y_i denotes the annual medical expenditures for the ith patient. Suppose, further, that the investigators propose to model the data using the following two-component mixture of normal distributions:

$$(1-\pi) N\left(y_i; \mu_1, \sigma_1^2\right) + \pi N\left(y_i; \mu_2, \sigma_2^2\right), \quad i = 1, \ldots, n$$

where π denotes the probability of class-2 membership and $N\left(\mu_k, \sigma_k^2\right)$ denotes a normal distribution with mean μ_k and variance $\sigma_k^2 (k = 1,2)$. Note that $\mu_2 > \mu_1$ implies that the mean expenditures for class 2 are higher than those for class 1. Further, $\sigma_2^2 > \sigma_1^2$ implies that class 2 is more dispersed than class 1. Then subjects for whom $\pi > 0.50$ are more likely to be in a class characterized by high average spending and increased variability relative to class 1. A comprehensive review of latent class models is given in the texts by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). In the remaining part of this section, four types of latent class models have been considered.

Latent Growth Models

Latent class models can be applied to longitudinal and clustered data. In the longitudinal setting, the classes are

characterized by average trajectories (or 'growth curves') over time. Consequently, these models are often referred to as latent growth models (LGMs). For example, a basic linear LGM for a normal outcome variable would take the form:

$$f\left(y_{ij}|\eta_{ijk},\sigma_k^2\right)=\sum_{k=1}^{K}\pi_k N\left(y_{ij}|\eta_{ijk},\sigma_k^2\right) \quad [18]$$

where $\eta_{ijk}=\beta_{0k}+\beta_{1k}t_{ij}$ and y_{ij} denotes the response at observation j for individual i; t_{ij} denotes the time (e.g., from baseline) of the ijth observation; β_{0k} and β_{1k} denote the intercept term and the trajectory slope for class k respectively; and σ_k^2 is the class-k variance of y_{ij}. Such models presume the existence of an unobserved discrete-valued variable that has both a main effect and an interaction effect with t_{ij} on the outcome. If the discrete-valued variable is observed, then eqn [18] reduces to a linear regression model with both main effects and time-interaction effects as yielded by the levels of that variable. Therefore, the defining feature of the latent growth model in eqn [18] is that the class to which an individual belongs is a discrete-valued latent variable that is unknown.

Extensions to nonlinear trajectories are straightforward. LGMs are especially popular in developmental psychology, where they have been used to model the progression of physical violence (Nagin and Tremblay, 1999) and criminal behavior (Roeder *et al.*, 1999). They have also been applied to joint longitudinal outcome–survival models (Lin *et al.*, 2002), where latent factor models are an alternative to the shared-parameter model approach to censored or missing data.

Growth Mixture Models

LGMs can be broadened to include subject-specific random effects. Such models are called growth mixture models (Muthén *et al.*, 2002) or heterogeneity models (Verbeke and Lesaffre, 1996). Growth mixture models assume that individuals are first placed into one of K latent classes that are defined by a mean trajectory curve (as in LGMs); then, around these mean trajectories, individuals are given their own subject-specific trajectories that are defined by a set of random effects with class-specific variance parameters. As such, growth mixtures can be viewed as finite mixtures of random effect models.

To continue with our previous example, the authors can extend the model in eqn [18] to include subject-specific intercept and slopes:

$$\begin{aligned} f\left(y_{ij}|\eta_{ijk},\sigma_k^2\right)&=\sum_{k=1}^{K}\pi_k N\left(y_{ij}|\eta_{ijk},\sigma_k^2\right),\\ \eta_{ijk}&=(\beta_{0k}+b_{0i})+(\beta_{1k}+b_{1i})t_{ij},\\ \begin{pmatrix} b_{0i}\\ b_{1i}\end{pmatrix}|C_i=k&\sim N_2(\mathbf{0},\boldsymbol{\Sigma}_k) \end{aligned} \quad [19]$$

where b_{0i} and b_{1i} denote, respectively, a random intercept and random slope for subject i. Conditional on subject i belonging to class k (i.e., $C_i=k$), the vector $(b_{0i}, b_{1i})'$ is assumed to follow a bivariate normal distribution with mean $\mathbf{0}=(0,0)'$ and class-specific covariance matrix $\boldsymbol{\Sigma}_k$. Growth mixture models have been applied in various contexts, including studies of class-specific prostate specific antigen trajectories (Lin *et al.*, 2000), daily affect (i.e., emotional expression) scores (Elliott *et al.*, 2005), and mental health expenditures (Neelon *et al.*, 2011).

Causal Inference via Latent Class Models

Latent class models have similarities with the principal stratification approach to causal inference (Frangakis and Rubin, 2002). The connection is illustrated in the context of a randomized controlled trial compromised by noncompliance. A subject's compliance status is formulated as a categorical variable with four levels on the basis of compliance behavior of the subject under both potential treatment assignments (Frangakis and Rubin, 1999). Under both assignments, compliers take the assigned treatment, always-takers take the experimental treatment, never-takers take the control, and defiers take the opposite treatment to that assigned. For example, in a randomized trial comparing the efficacy of two antipsychotics for refractory schizophrenia, compliers might be defined as individuals who would take the assigned treatment for the entire follow-up period whereas the other three groups characterize those individuals who would switch treatment under one (always- and never-takers) or both (defiers) treatment assignments (O'Malley and Normand, 2005). Because compliance status does not depend on the outcome, the same definitions would apply to a health economic outcome such as aggregate mental health cost of treatment. In general, compliance status can be considered as a latent class because it is unobserved (compliance behavior is observed only under the assigned treatment) and the expected outcome from treatment may vary with compliance status (compliers form one principal stratum, never-takers form another stratum, etc.). Thus, a model such as that in eqn [17] could be used. However, in causal inference, the more common approach is to identify the model by imposing structural assumptions as opposed to parametric assumptions, which cannot be completely tested from the data. It is typically assumed that defiers do not exist, an assumption referred to as monotonicity, and that treatment assignment only affects outcomes through the treatment received, the exclusion restriction. Unbiased nonparametric moment-based estimators are then available for the effect of the treatment received on the outcome. In this sense, treatment assignment is an instrumental variable and, under the additional assumption that one individual's treatment does not affect another's outcome (the stable unit treatment value assumption), the estimand is a local-average treatment effect (Angrist *et al.*, 1996).

Model Fitting

Two techniques that are especially well-suited to estimation of models with latent variables are the expectation-maximization (EM) algorithm for frequentist inference and Markov-chain Monte Carlo (MCMC) for Bayesian inference. Their suitability arises from the fact that the values of latent variables (continuous-latent factors or categorical latent classes) can be considered as the missing data. Estimation for latent factor and latent class models proceeds by treating the latent

variables as missing data and applying either the EM algorithm (Dempster *et al.*, 1977) or, in the Bayesian context, MCMC (Gelfand and Smith, 1990).

EM and MCMC computations can be conceptualized as applying a regular regression (linear or otherwise) on the basis of imputed values of the latent variables (the complete data analysis) and subsequently using all of the information in the data as well as the fitted model to impute the latent variables. In the second step, the EM algorithm yields 'best' values of the latent variables by maximizing the complete data likelihood function whereas the MCMC algorithm yields random realizations of the latent variables from the corresponding joint posterior distribution (also referred to as data augmentation) (van Dyk and Meng, 2001).

Prior Distributions for Bayesian Modeling

To fit a Bayesian model, prior distributions need to be specified for the model parameters that have not already been assigned probability distributions (essentially all parameters other than the latent variables). As for the regression coefficients of the observed predictors, the coefficients of the latent factors (the factor loadings) are often assigned normal distributions. When the outcomes y_i or their unobserved continuous counterparts (e.g., in the bivariate probit model) follow a normal distribution, a normal prior yields a normal posterior distribution, which simplifies the MCMC procedure by allowing posterior samples to be drawn directly. However, unless the parameters are restricted to ensure that the model is identifiable, the computational issues discussed in the Section 'Computational Challenges' can hinder convergence of model fitting algorithms. An advantage of Bayesian modeling is that prior distributions can often easily accommodate constraints for making the model identifiable by data. For example, in an analysis of monthly international exchange rates, Lopes and West (2004) specify independent normal priors for the factor loadings with two restrictions: the loadings above the diagonal are 0 (thus the factor loading matrix is block lower triangular); and the diagonal elements are nonnegative.

In the latent class model, the class prevalence or mixing probabilities, the π_ks, are typically assigned a Dirichlet prior, which leads to a convenient closed-form conditional posterior distribution. Covariance matrices are often assigned inverse-Wishart prior distributions or variants thereof, although alternative specifications are becoming more common (O'Malley and Zaslavsky, 2008).

Computational Challenges

Several numerical challenges that arise in fitting SEMs, inclusive of latent factor and latent class models, are due to the fact that the parameterization of latent factor and latent class terms being permuted without affecting the fitted model. This problem is partly resolved in the latent factor case through the use of a uniqueness condition (see Section 'Latent Factor Models'). However, under MCMC estimation, a related problem called label-switching (Stephens, 2000) arises when the order of the factors varies across the draws from the joint posterior distribution, in which case, posterior summaries resulting from naïve Monte Carlo averages will be nonsensical. To enable inference concerning latent factors, postprocessing is necessary in order to obtain a consistent order of the factors across the posterior draws before computing posterior summaries. Postprocessing may also be applied to the draws of the latent class parameters to account for the possibility that 'Class 1' in one draw is 'Class 2' in another, corresponding to equivalence classes at which the likelihood function has equal maximal values.

Software for implementing latent factor and latent class models includes Mplus, Latent Gold®, WinBUGS, and R packages for specific families of models. The SAS procedure Proc Traj fits LGMs for panel data, including whenever observation times are unevenly spaced (Jones and Nagin, 2007).

Model Comparison and Checking

A general way of comparing single-level models (models that do not include random effects or latent variables) is the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), also known as the Schwarz Criterion. The AIC and BIC balance the level of fit (quantified in terms of the log-likelihood) with model complexity (a penalty for using the sample data to estimate the model parameters). A challenge in applying these methods to SEMs lies in the estimation of latent variables and their effects wherein amounts of information (i.e., degrees-of-freedom) being used are different from those utilized during the estimation of observed predictors and their effects. Therefore, assessing model complexity on the basis of the number of estimated parameters is not appropriate.

In Bayesian analysis, model comparison on the basis of Bayes factors (Kass and Raftery, 1995) is the most principled approach though computational problems may be encountered. Because Bayes factors rely on the marginal likelihood of the data under a presumed model, they only exist if the prior on the model parameters is proper. To allow the use of improper priors, an alternative to Bayes factors, such as the intrinsic Bayes factor (Berger and Pericchi, 1996) has been proposed. The pseudo Bayes factor (Gelfand and Dey, 1994) offers a computationally convenient numerical approximation but has been criticized due to its dependence on the harmonic mean (Neal, 2008). An alternative to Bayes factors is the Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002), which can be regarded as a Bayesian counterpart to the AIC. In response to ambiguity over the appropriate way of accounting for latent variables in finite mixture models, Celeux *et al.* (2006) have proposed several alternative DIC measures with improved inferential properties. However, discerning and implementing the appropriate measure of DIC is not straightforward in many situations.

Once a model is selected, Bayesian posterior predictive checks can be used to compare the observed data to the one replicated from the posterior predictive distribution under the model (Gelman *et al.*, 1996). If the model fits well, the replicated data would resemble the observed data. To quantify the degree of similarity, the percentile of the predictive distribution corresponding to the observed value of a discrepancy measure

that reflects an aspect of model-fit important to the area of study is evaluated. If the percentile, a Bayesian predictive *p*-value, is near 0 or 1, the model exhibits lack-of-fit. For more details of Bayesian model specification, model fitting, and model checking, refer to the article on Bayesian analysis.

Limitations of SEMs

One of the most common criticisms of models involving latent variables is that model identifiability stems from the distribution specified for unobserved variables. Because such assumptions cannot be completely tested by the data, there is a concern that models with latent variables are unscientific. This is of particular concern in models involving structural assumptions such as instrumental variable assumptions. Such concerns have motivated research on nonparametric and semiparametric methods (Lee, 1995), including alternatives to parametric hierarchical (or mixed effect) models (Heagerty and Zeger, 2000).

In studies involving structural assumptions, deciding between nonparametric and parametric SEMs entails a trade-off between assumptions. For example, in IV analyses, the trade-off is between identifiability of model parameters via the exclusion restriction (typically supported by a theoretical model and, in the case of multiple IVs, partially tested empirically using a test of over-identifying conditions) or via the joint probability distribution of the outcome variable and the endogeneous predictor (O'Malley *et al.*, 2011). In practice, one approach may be a sensitivity analysis for the other.

Summary

The intersection of heterogeneity and SEM encompasses a diverse range of models. Before concluding, several models that are equally important though more loosely connected to the central theme of this article are mentioned. These include two-part models and spatial models. Two-part models account for outcome distributions having multiple parts, distributional heterogeneity. For example, when analyzing medical costs, it is often the case that the outcome distribution is part discrete (zero costs arising when no service is performed) and part continuous (a broad range of nonzero costs associating with different services). Such 'semi-continuous' data may be modeled using two-part models with one component of the model being dedicated to the likelihood that the outcome (e.g., cost) is 0 while the other one to the expected outcome of being nonzero (Neelon *et al.*, 2011; Olsen and Schafer, 2001). For more details, refer to the article on modeling expenditure and utilization data.

Analogous models exist for zero-inflated count data. One such model is the Poisson hurdle model, which is a two-component mixture consisting of a point mass at zero, followed by a truncated Poisson for nonzero observations (Mullahy, 1986). Other count distributions, such as the negative binomial, can alternatively be used. A related model is the zero-inflated Poisson (ZIP) model, which consists of a degenerate distribution at zero and is mixed with an untruncated Poisson distribution (Lambert, 1992). The ZIP partitions the zeroes into two types: 'structural' zeroes (e.g., those that occur because patients are ineligible for health services) and 'chance' zeroes (e.g., those that occur by chance among eligible patients) (Neelon *et al.*, 2010). For more details on modeling count data, refer to the article on modeling ordinal outcomes.

In spatial analysis, heterogeneity refers to how a variable y_i varies across a region of space. Two common types of spatial data are point-referenced data and areal data. For point-referenced data, y_i is measured at a set of geo-referenced locations, $\mathbf{s}$, and the covariance of y_i at locations $\mathbf{s}_1$ and $\mathbf{s}_2$ is assumed to be a function of the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$. For areal data, the spatial unit is an aggregated region of space, such as a Census block or a county, and y_i is typically a count or average response among individuals residing in that region. Popular models for analyzing areal data include the simultaneously autoregressive (SAR) and conditionally autoregressive (CAR) models. Foundational work in the field of spatial modeling has been conducted by Whittle (1954) and Besag (1974). The field of spatial econometrics includes a literature on network autocorrelation as well as other models for the sake of estimating peer effects (Anselin, 1988). For a comprehensive discussion of spatial models, see the text by Banerjee *et al.* (2004) and the Encyclopedia entry on spatial analysis.

In line with the general emphasis in the statistics and econometric literature, our focus has been on models for the mean (or transformations thereof). One of the few exceptions are mixed effect location-scale models, where the variance as well as the mean of the outcome depends on latent variables (Hedeker *et al.*, 2009). Such models allow shrinkage to an overall variance in addition to shrinkage to an overall mean.

Although it is always possible to specify SEMs by writing out a series of equations or a path diagram, the recent explosion in computing power and development of computer software programs to harness this power have made it possible to fit a wide range of models. This has enabled many extensions to traditional models including accounting for missing data, clustered or hierarchical data, and other heterogeneous features of models to be accommodated. SEMs can yield powerful improvements over traditional approaches to regression, covariance decomposition (or factor) analysis, grouping (or clustering) subjects, and separating cause from association. In the future, the authors predict that the uptake of SEMs will continue to expand into new areas of application.

Acknowledgment

The authors thank Alan Zaslavsky and Jaeun Choi for comments on an earlier draft of the article. A. James O'Malley's effort was in part supported by NIH Grant IRC4MH092717-01.

See also: Analysing Heterogeneity to Support Decision Making. Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap. Instrumental Variables: Informing Policy. Instrumental Variables: Methods. Missing Data: Weighting and Imputation. Modeling Cost and Expenditure for Healthcare. Models for Count Data. Models for Discrete/Ordered Outcomes and Choice Models. Models for Durations: A Guide to

Empirical Applications in Health Economics. Observational Studies in Economic Evaluation. Panel Data and Difference-in-Differences Estimation. Primer on the Use of Bayesian Methods in Health Economics. Risk Selection and Risk Adjustment. Sample Selection Bias in Health Econometric Models. Spatial Econometrics: Theory and Applications in Health Economics

References

Angrist, J. D. (1998). Estimating the labor market impact on voluntary military service using social security date on military applicants. *Econometrica* **66**, 249–288.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Ch. 4. Princeton, NJ: Princeton University Press.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht. The Netherlands: Kluwer Academic Publishers.

Arendt, J. N. and Holm, A. (2006) Probit models with dummy endogenous variables. CAM Working Papers. Available at: http://EconPapers.repec.org/RePEc:kud:kuieca:2006_06 (accessed 17.04.13).

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall: Boca Raton, FL.

Bedeian, A. G., Day, D. V. and Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: Some important reminders. *Educational and Psychological Measurement* **57**, 785–799.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association* **45**, 164–180.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological* **36**, 192–236.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bonett, D. G. and Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics* **30**, 213–225.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement error in nonlinear models*. New York: Chapman and Hall.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–674.

Davidian, M., Carroll, R. J. and Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika* **75**, 549–556.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B* **62**, 355–366.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1–50.

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R. and Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6**, 119–143.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–379.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–760 (Discussion: pp. 760–807).

Heagerty, P. and Zeger, S. (2000). Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science* **15**, 1–26.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46**, 931–960.

Hedeker, D., Demirtas, H. and Mermelstein, R. J. (2009). A mixed ordinal location scale model for analysis of ecological momentary assessment (EMA) data. *Statistics and Its Interface* **2**, 391–401.

Hogan, J. W. and Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association* **99**, 314–324.

Horton, N. J., Roberts, K., Ryan, L., Suglia, S. F. and Wright, R. J. (2008). A maximum likelihood latent variable regression model for multiple informants. *Statistics in Medicine* **27**, 4992–5004.

Johnson, R. A. and Wichern, D. W. (1998). *Applied multivariate analysis*. Ch. 9. Upper Saddle River, NJ: Prentice-Hall.

Jones, B. L. and Nagin, D. S. (2007). Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods and Research* **35**, 542–571.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In Goldberger, A. S. and Duncan, O. D. (eds.) *Structural equation models in the social sciences*, pp. 85–112. New York: Seminar Press.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Landon, B. E., Normand, S.-L. T., Lessler, A., et al. (2006). Quality of care for the treatment of acute medical conditions in United States hospitals. *Archives of Internal Medicine* **166**, 2511–2517.

Lee, M.-J. (1995). Semi-parametric estimation of simultaneous equations with limited dependent variables: A case study of female labour supply. *Journal of Applied Econometrics* **10**, 187–200.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Ch. 2. New York: Wiley.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H. and Clark, L. C. (2000). A latent class mixed model for analyzing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine* **19**, 1303–1318.

Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97**, 53–65.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, Chichester: John Wiley & Sons.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.

Muthén, B., Brown, C. H., Masyn, K., et al. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3**, 459–475.

Nagin, D. and Tremblay, R. E. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development* **70**, 1181–1196.

Neal, R. (2008). The harmonic mean of the likelihood: Worst Monte Carlo method ever. citeulike: 5738012.

Neelon, B. H., O'Malley, A. J. and Normand, S.-L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to psychiatric outpatient service use. *Statistical Modelling* **10**, 421–439.

Neelon, B. H., O'Malley, A. J. and Normand, S.-L. T. (2011). A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. *Biometrics* **67**, 280–289.

O'Malley, A. J., Frank, R. G., Normand, S.-L. T. (2011). Estimating cost-offsets of new medications: Use of new antipsychotics and mental health costs for schizophrenia. *Statistics in Medicine* **30**(16), 1971–1988.

O'Malley, A. J. and Normand, S.-L. T. (2005). Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics* **61**, 325–334.

O'Malley, A. J., Normand, T. and Kuntz, R. E. (2003). Application of models for multivariate mixed outcomes to medical device trials: Coronary artery stenting. *Statistics in Medicine* **22**, 313–336.

O'Malley, A. J., Smith, M. H. and Sadler, W. A. (2008). A restricted maximum likelihood procedure for estimating the variance function of an immunoassay. *Australian and New Zealand Journal of Statistics* **50**, 161–177.

O'Malley, A. J. and Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association* **103**, 1405–1418.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika* **69**, 167–190.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.

Raudenbush, S. W. and Sampson, R. (1999). Assessing direct and indirect associations in multilevel designs with latent variables. *Sociological Methods and Research* **28**, 123–153.

Reich, B. J. and Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *The Annals of Applied Statistics* **4**, 439–459.

Roeder, K., Lynch, K. G. and Nagin, D. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association* **94**, 766–776.

Sammel, M. D., Ryan, L. M. and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B* **59**, 667–678.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*, pp. 9, 66. Boca Raton, FL: Chapman and Hall/CRC.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64**, 583–616.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **62**, 795–809.

Teixeira-Pinto, A. and Normand, S.-L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in Medicine* **28**, 1753–1773.

Tsonaka, R., Verbeke, G. and Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* **65**, 81–87.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Vonesh, E. F., Greene, T. and Schluchter, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine* **25**, 143–163.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–449.

Williams, L. J. and Hazer, J. T. (1986). Antecedents and consequences of satisfaction and commitment in turnover models: A reanalysis using latent variable structural equation methods. *Journal of Applied Psychology* **71**, 219–231.

Wooldridge, J. L. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Learning by Doing

V Ho, Rice University, Houston, TX, USA

Glossary

Endogeneity A situation where an explanatory variable in a regression equation is correlated with the error term, because of an omitted variable, measurement error, or simultaneity.

Fixed effects In a dataset with multiple observations of the same health care facility over time, unobserved factors specific to that facility that do not change over time can be modeled in a regression equation by specifying a dummy variable (fixed effect) for each facility.

Instrumental variables (IV) In a regression equation with an endogenous explanatory variable, an IV is a variable that is correlated with the endogenous variable, but is uncorrelated with the error term in the regression and is excluded from the regression.

Introduction

Learning by doing is viewed as an important determinant of success for many professions requiring high skill. Over the years, researchers have come to realize that teams and firms can also exhibit learning by doing. Even in cases where annual output does not increase over time, a firm can experience reductions in unit costs or improvements in quality that cannot be attributable to economies of scale, but cumulative experience instead. The presence of learning can have important implications for overall growth in a nascent industry. Differential learning across workers and firms can also have important implications for competition in the market. Health economists have been particularly interested in learning, because current and emerging medical technologies are complex, requiring both individual and team-based skills which are likely to benefit from experience.

Social scientists have been examining the impact of learning by doing on production technology for several decades. The concept of a learning curve was first described in 1936, when a study determined that as the quantity of manufactured units doubled, the number of direct labor hours required to produce an individual unit decreased at an uniform rate (Wright, 1936). Another early study concluded that the aircraft industry's rate of learning, or reduced labor requirement, was 80% between doubled quantities of airframes (Alchian, 1963).

The standard equation that is used in the literature to characterize a learning curve takes the form:

$$y_i = \alpha x_i^{-\beta} \tag{1}$$

where y represents the resources (hours or costs) required to produce the ith unit, α is the amount of resources required to produce the first unit, x is the cumulative number of units produced through the current time period i, and β is the learning rate (Argote, 1999). Taking natural logs yields a regression that can be readily estimated:

$$\ln y_i = \alpha - \beta \ln x_i \tag{2}$$

Most economic studies of learning by doing have employed this general framework by estimating the effectiveness of cumulative output production in reducing average costs. For example, it has been determined that on average each doubling of plant scale was accompanied by a 11% reduction in unit costs in the chemical industry (Lieberman, 1984). In the health economics literature, learning by doing has been tested for insurance plans, hospitals, and doctors. One study determined that clinical costs decline 10–15% with each doubling of experience for insurers administering managed behavioral health plans (where experience is measured as the cumulative number of managed care claims processed in a state by a particular health plan) (Sturm, 1999). However, most health economic studies have attempted to measure the effects of cumulative experience on patient outcomes (primarily mortality) rather than unit costs.

Hospital-Level Studies

Hospital-level studies of learning by doing have examined specific complex operations or procedures performed on patients. These studies focus on specific procedures in order to control for heterogeneity across treatments provided in hospitals. Patient outcomes (mostly patient mortality) are the dependent variable of interest, and regressions include both cumulative output and annual output as explanatory variables. Cumulative output is hypothesized to represent learning by doing, whereas annual output is hypothesized to reflect economies of scale. Because many of these studies do not have access to patient data dating back to when a particular operation was initially introduced for medical care, these studies often proxy for cumulative output using lagged values (e.g., number of operations performed 1, 2, or 3 years ago at a particular hospital) as a proxy for cumulative output.

Most published studies of learning by doing at the hospital level are based on two procedures for heart disease: Coronary artery bypass graft (CABG) surgery, and percutaneous transluminal coronary angioplasty (PTCA) (Gaynor *et al.*, 2005; Ho, 2002; Pisano *et al.*, 2001; Sfekas, 2009). CABG is a form of open heart surgery in which the rib cage is opened and a section of a blood vessel is grafted from the aorta to the coronary artery. PTCA is a procedure performed to improve blood supply to the heart. A balloon-tipped catheter is inserted into

an artery in the groin or shoulder and threaded to the blocked artery. The balloon is then inflated to flatten atherosclerotic plaque against the artery wall, reopening the artery. Health economists have focused on these procedures, because heart disease is the leading cause of death in the United States. Therefore, these procedures are performed frequently by many hospitals, so that data are readily available. The data are usually derived from one or multiple US states, which allow researchers access to detailed data from hospital discharge abstracts for all admissions for several years. These data specifications are required, so that researchers can accurately count the cumulative and annual number of procedures performed for each hospital. CABG and PTCA have also been the focus of interest for learning by doing studies, because there is a large body of medical literature that specifies the information that is necessary to control for patient characteristics that influence patient mortality and other outcomes for these two procedures. The required variables, which include multiple demographic and clinical characteristics, are also available in hospital discharge datasets.

Multiple studies find no support for learning by doing at the hospital level for either CABG or PTCA. Lagged volume or cumulative volume tends not to be statistically significant in explaining mortality for patients who undergo these procedures. This conclusion has been found for studies analyzing data from Arizona, California, and Maryland for various sample periods spanning the years 1983 through 2001 (Gaynor *et al.*, 2005; Ho, 2002; Sfekas, 2009). All of these studies include hospital specific fixed effects (dummy variables for each hospital in the sample) in the regression specifications in order to control for unobserved heterogeneity across hospitals which are constant over time. For example, some hospitals may benefit from exceptional and long-tenured nursing staff, or highly talented administrative staff. These factors can influence patient outcomes, but they are not observable in hospital discharge abstracts. The inclusion of hospital fixed effects means that the regressions estimate the effects of increases or decreases within the hospital in procedure volume over time, rather than the effects of differences in cumulative volume across hospitals on patient mortality. The fixed effect specification will more accurately capture the learning by doing effect which is hypothesized in the underlying economic model. However, precise estimation requires a sample of hospitals with data from a sufficient number of time periods in order to observe significant variation in procedure volume across time. Thus, these prior studies may have failed to precisely estimate a learning by doing effect. The samples in these studies contained data from only one or two states. Samples of this size may not have enough hospitals that experienced noticeable changes in procedure volume across years.

One study of 16 institutions that began performing a new procedure for minimally invasive heart surgery found that the amount of time required to perform the operation declined as the cumulative number of procedures increased (Huckman and Pisano, 2006). This result is tangible, even with the inclusion of hospital fixed effects in the regression models. The patient-level data were collected during the first 2 years after which the procedure was first approved by the Food and Drug Administration. Thus, this study may have been able to detect an institution-specific learning by doing effect, because the analysis was performed just after the technology was introduced; when the greatest amount of learning most likely occurs.

Although health economists have had little success identifying a tangible effect of cumulative volume on patient outcomes, a large literature in both medical and health services research journals finds a significant association between procedure volume and patient mortality. In addition to CABG and PTCA, this 'volume-outcome' relationship has been documented for a wide range of procedures and treatments, including carotid endarterectomy, hip replacement, lung cancer resection, liver transplantation, and neonatal intensive care (Halm *et al.*, 2003; Luft *et al.*, 1979; Birkmeyer *et al.*, 2002). When this relationship was first identified in the medical literature, learning by doing was mentioned as a likely explanation for this finding. Researchers suggested that if experience was the underlying source of the volume-outcome effect, then complex operations should be 'regionalized,' so that patients would benefit from improved outcomes at a select number of facilities that would be able to gain greater experience. The absence of a significant effect of cumulative volume on patient mortality for CABG and PTCA casts doubt on the learning by doing hypothesis, particularly for common cardiac procedures.

One other challenge faced by health economists trying to identify a learning by doing effect is that cumulative volume and annual volume are highly correlated. Hospitals that perform a large number of procedures in 1 year tend to do so in subsequent years. In at least one instance, an analysis of hospital data for PTCA could not explicitly test for the effect of learning by doing on average costs per patient, because inclusion of both cumulative and annual volume as explanatory variables led to multicollinearity (Ho, 2002). This issue might be resolved if researchers were able to analyze data from multiple states simultaneously, with data stretching over many years. Gathering a much larger sample would increase the likelihood that one could find hospitals that experienced sufficient variation in volume (e.g., due to entry or exit of competitors), which would weaken the collinearity between cumulative and annual volume.

It is also interesting to note that learning by doing studies in the economics literature have tended to focus on patient outcomes as the dependent variable of interest rather than costs. This focus contrasts with the general industrial organization literature, where there is much less research on the relationship between learning by doing and product quality. Some research has analyzed data on nuclear power plants (Lester and McCabe, 1993). Both reactor-specific learning and spillovers across reactors have been found to be important determinants of nuclear reactor performance. Learning by doing as measured by cumulative output has also been associated with fewer complaints in the aircraft production industry (Argote, 1993).

There may be fewer studies of the effect of learning by doing on costs in the health economics literature, because it is difficult to obtain datasets that provide both detailed information on patient outcomes and the costs of care. Hospital discharge abstracts often contain information on the total charges for a patient admission. These data can be linked with

hospital cost reports that contain the cost-to-charge ratio for each hospital, so that an estimate of costs per patient admission can be calculated. However, the saliency of patient mortality as a dependent variable of interest may have led to the greater focus of learning by doing studies on health outcomes for patients.

Physician-Level Studies

Many fewer published studies have attempted to estimate the volume-outcome effect at the surgeon level. The lack of studies stems in part from the fact that it is difficult to identify hospital datasets that provide consistent identifiers of physicians across patients and time. Only one published study included cumulative surgeon volume as an explanatory variable to explain patient mortality for CABG, and it finds no evidence of learning by doing (Huesch, 2009). In fact, this study also finds no association between annual surgeon volume and patient outcomes, although a small number of studies in the medical literature find that surgeons who perform more complex operations achieve lower mortality rates. Another study of approximately 4000 patients who received LASIK surgery in the early 2000s in the country of Colombia also found no effect of cumulative surgeon procedure volume on patient outcomes (Contreras *et al.*, 2011). The presence of learning by doing effects at the hospital versus the individual doctor level is likely to vary by medical intervention. For some operations, the surgeon's technical skill and discretion over specific intraoperative processes are likely important determinants of patient outcome. In other operations, hospital-based services (intensive care, pain management, respiratory care, and nursing care) are more likely to determine inpatient mortality.

Endogeneity

One may be concerned that the absence of a learning by doing effect may reflect endogeneity in the volume-outcome effect. There may be factors that are unobservable to the researcher, which influence both procedure volume and patient outcomes, leading to an observed association between these two variables in a regression model. For example, some facilities may be more quick to invest in newer surgical devices, which allow them to treat more patients and achieve better outcomes simultaneously. Endogeneity may also result from selective referral. The reputation of higher quality hospitals or surgeons may become well known in the community, attracting more patients seeking care.

Some learning by doing studies have accounted for potential endogeneity using instrumental variables techniques (Gaynor *et al.*, 2005). The variables that are hypothesized to influence procedure volume but are otherwise uncorrelated with patient outcomes include: The number of patients residing within a fixed geographical radius of a hospital, the number of other hospitals offering the same procedure within a fixed geographical radius of a hospital, and the predicted number of patients to choose a hospital for treatment, based on distance from the patients' residences to each particular hospital. These instruments are significant predictors of patient volume, but specification tests cannot reject the null hypothesis that procedure volume is exogenous in explaining patient outcomes. Therefore, concerns regarding the potential endogeneity of procedure volume are not supported by current empirical analyses.

Forgetting

The general industrial organization literature has also tested for the presence of forgetting in firm production (Benkard, 2000; Thompson, 2007). This literature considers the possibility that productivity gains from learning can depreciate over time. More flexible regression specifications capture the fact that cost per unit of output can rise during significant production troughs that may occur in the life cycle of a product. Only one published paper has attempted to test for forgetting in the health economics literature, and it found almost complete forgetting from prior experience among recently trained surgeons performing CABG (Huesch, 2009). More studies need to be performed to validate this finding. The industrial organization literature identified forgetting in the context of airplane manufacturing, where there can be noticeable declines in production in the life cycle of a particular model of airplane. In contrast, most hospitals are not likely to experience noticeable troughs in the performance of a procedure. It would be useful to identify a large sample of hospitals that had experienced the entrance of a nearby competitor for the same procedure to precisely estimate a forgetting effect. Determining the extent to which forgetting exists in the performance of complex medical treatments has important implications for patient care. If there is little depreciation in learning, then one can be more certain that hospitals or surgeons who are currently high quality will remain so in the future. If forgetting does exist, further studies would be needed to determine why learning depreciates. Quality could depreciate over time, because the skill set of surgeons could depreciate with lack of use, or because multidisciplinary teams of caregivers become less coordinated if they treat fewer patients.

Other Forms of Learning

Past industrial organization studies have also identified forms of learning other than learning by doing. For instance, in a study of the semiconductor industry, firms learn three times more from an additional unit of their own cumulative production than from an additional unit of another firm's cumulative production (Irwin and Klenow, 1994). The reductions in unit costs associated with increases in other firms' cumulative production or industry cumulative output are referred to as spillover effects. In this context, a firm's own learning by doing is referred to as proprietary learning. It is plausible that spillover learning could occur in the context of complex medical procedures. For example, a hospital performing a small volume of procedures in a city with several large facilities nearby may have better outcomes than a

comparatively small facility in a rural area. The small urban hospital may be able to benefit from nearby expertise.

Cost reductions associated with calendar time rather than production quantity have been referred to as 'learning by watching.' Hospitals may be able to improve outcomes by learning from the experience of other facilities. For example, a hospital which began performing 50 PTCAs per year in 1996 is likely to have better outcomes than a comparable hospital in 1986, because the former facility could benefit from the knowledge and experience gained over the previous decade. One study that found little evidence of learning by doing based on the cumulative number of PTCAs performed by hospitals over time found substantial evidence of learning by watching for this procedure (Ho, 2002). Outcomes improved year by year for all hospitals, regardless of the cumulative number of angioplasty procedures they performed. Learning by watching has also been identified for the performance of LASIK (Contreras *et al.*, 2011). Significant improvements in outcomes were observed at two points in the sample period analysis when all physicians in a practice performing this procedure met to update surgical plans based on patient characteristics.

Determining the relative magnitude of learning by doing, spillover learning, and learning by watching is important for assessing the relative success of small versus large firms. If most learning is nonproprietary and few economies of scale exist, then small firms can more easily compete with large firms. Health economics lacks a comprehensive set of studies that test for learning by doing for a range of procedures and for hospitals or physicians in multiple states. The studies so far find little evidence for learning by doing, whereas there is more convincing evidence for learning by watching. These findings suggest that there is little support for 'regionalizing' complex surgical procedures at a select number of high volume hospitals that would benefit from greater experience.

Conclusion

Researchers have identified learning by doing that reduced unit costs in industries ranging from chemical processing to semiconductors. And there are hundreds of papers in the medical literature finding an association between higher hospital or surgeon procedure volume and lower mortality rates. However, most rigorous econometric analyses of health care data have been unable to formally identify learning by doing. Perhaps health economists lack sufficient data to distinguish between annual and cumulative output measures when testing for learning by doing in mortality and/or costs. Analysis of a wider range of newly emerging medical treatments, as well as more detailed data on costs would help to explain the role of learning in influencing the costs and quality of medical care.

In the meantime, policy makers should be cautious of recommendations to centralize complex surgical procedures based on existing volume-outcome studies. Although larger providers tend to yield better patient outcomes, making them even larger will not likely lower hospital mortality rates further. More research is required to determine the underlying reasons for the volume-outcome relationship. One should also keep in mind that learning by watching effects appear to be significant in health care. All providers tend to improve over time, regardless of volume. Given the potential beneficial effects of competition in maintaining quality and lower costs, patients may in fact be better off without centralization of complex treatment.

See also: Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Heterogeneity of Hospitals. Instrumental Variables: Informing Policy. Panel Data and Difference-in-Differences Estimation. Production Functions for Medical Services

References

Alchian, A. (1963). Reliability of progress curves in airframe production. *Econometrica* **31**(4), 679–693.

Argote, L. (1993). Group and organizational learning curves: Individual, system and environmental components. *British Journal of Social Psychology* **32**, 31–51.

Argote, L. (1999). *1st Organizational learning: Creating, retaining and transferring knowledge*. Norwell, MA: Kluwer Academic Publishers.

Benkard, C. L. (2000). Learning and forgetting: The dynamics of aircraft production. *American Economic Review* **90**(4), 1034–1054.

Birkmeyer, J. D., Siewers, A. E., Finlayson, S. R., et al. (2002). Hospital volume and surgical mortality in the United States. *New England Journal of Medicine* **346**(15), 1128–1137.

Contreras, J. M., Kim, B. and Tristao, I. M. (2011). Does doctors' experience matter in lasik surgeries? *Health Economics* **20**(6), 699–722.

Gaynor, M., Seider, H. and Vogt, W. B. (2005). The volume-outcome effect, scale economies, and learning-by-doing. *American Economic Review* **95**(2), 243–247.

Halm, E. A., Chassin, M. R., Tuhrim, S., et al. (2003). Revisiting the appropriateness of carotid endarterectomy. *Stroke*. **34**(6), 1464–1471.

Ho, V. (2002). Learning and the evolution of medical technologies: The diffusion of coronary angioplasty. *Journal of Health Economics* **21**(5), 873–885.

Huckman, R. S. and Pisano, G. P. (2006). The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science* **52**(4), 473–488.

Huesch, M. D. (2009). Learning by doing, scale effects, or neither? Cardiac surgeons after residency. *Health Services Research* **44**(6), 1960–1982.

Irwin, D. A. and Klenow, P. J. (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political Economy* **102**(6), 1200–1227.

Lester, R. K. and McCabe, M. J. (1993). The effect of industrial structure on learning by doing in nuclear power plant operation. *RAND Journal of Economics* **24**(3), 418–438.

Lieberman, M. B. (1984). The learning curve and pricing in the chemical processing industries. *RAND Journal of Economics* **15**(2), 213–228.

Luft, H. S., Bunker, J. and Enthoven, A. (1979). Should operations be regionalized? An empirical study of the relation between surgical volume and mortality. *The New England Journal of Medicine* **301**(25), 1364–1369.

Pisano, G. P., Bohmer, R. M. J. and Edmondson, A. C. (2001). Organizational differences in rates of learning?: Evidence from the adoption of minimally invasive cardiac surgery. *Management Science* **47**(6), 752–768.

Sfekas, A. (2009). Learning, forgetting, and hospital quality: An empirical analysis of cardiac procedures in Maryland and Arizona. *Health Economics* **18**(6), 697–711.

Sturm, R. (1999). Cost and quality trends under managed care: Is there a learning curve in behavioral health carve-out plans? *Journal of Health Economics* **18**(5), 593–604.

Thompson, P. (2007). How much did the liberty shipbuilders forget? *Management Science* **53**(6), 908–918. Available at: http://mansci.journal.informs.org/cgi/doi/10.1287/mnsc.1060.0678 (accessed on 19 February 2013).

Wright, T. P. (1936). Factors affecting the cost of airplanes. *Journal of Aeronautical Sciences* **3**(4), 122–128.

Further Reading

Halm, E. A., Lee, C. and Chassin, M. R. (2002). Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Annals of Internal Medicine* **137**(6), 511–520.

Huesch, M. D. and Sakakibara, M. (2009). Forgetting the learning curve for a moment: How much performance is unrelated to own experience? *Health Economics* **18**(7), 855–862.

Long-Term Care

DC Grabowski, Harvard Medical School, Boston, MA, USA

Introduction

Long-term care is a set of services delivered over a sustained period of time to people who lack some degree of functional capacity. Put alternatively, long-term care is the help needed to cope, and sometimes to survive, when physical and cognitive disabilities impair the ability to perform activities of daily living (ADL), such as eating, bathing, dressing, using toilet, and walking. Unlike the provision of general health services, which are often targeted toward acute medical problems, long-term care must be continually provided and is, thus, continually expensive.

Long-term care services are needed by a diverse set of individuals who receive care from an equally wide array of providers. As the result of declining functioning, older individuals – especially the very old – are the primary recipients of long-term care services, but in some instances, younger individuals with physical or cognitive limitations also require services. The primary providers of long-term care services in most countries are 'informal' providers such as family members and friends. Formal providers include nursing homes, board and care homes, home health care agencies, assisted living facilities, adult foster and day care homes, home- and community-based providers, and continuing care retirement communities (CCRCs). Across these different formal providers, a number of different payer types exist including out-of-pocket, public and private insurance. Because consumers are often thought to lack information regarding the quality of services provided, an immense amount of government regulation exists within institutional long-term care settings in countries such as the US.

Long-term care has been an active and distinct subfield of health economics for some time. To paraphrase an old line 'long-term care economics is like health economics, only more so,' several of the key features that make the economics of health different from the economics of other goods and services are even more pronounced in the study of long-term care. That is, the assumption of the well-informed, rational consumer is more dubious; the role of government as a payer and regulator is more prominent; the response to financial incentives such as insurance is exacerbated for certain services; and the external costs of illness are often more formidable.

This article provides a broad discussion of the basics of long-term care: who needs it; who provides it; who pays for it; and some background on government regulation of these services. Next, this article provides an overview of some of the central issues in the economics of long-term care: The nonpurchase of private long-term care insurance; long-term care quality; pay-for-performance in long-term care; cost-effectiveness of home- and community-based services (HCBS); effects on informal caregiving; and the integration of long-term care with other health care services.

Who Needs Long-Term Care?

The key to long-term care is functioning. Unlike acute health care where a number of highly technical medical services are typically provided to patients, long-term care is assistance with daily tasks of living. Long-term care personnel have divided these tasks into ADLs, such as eating, using toilet, dressing, bathing, and locomotion and instrumental ADLs (IADLs), such as cooking, cleaning, doing laundry, handling household maintenance, transporting themselves, reading, writing, managing money, using equipment such as the telephone, and comprehending and following instructions. Clearly, the need for assistance with multiple ADLs might necessitate more intensive long-term care such as a nursing home, whereas the need for assistance with one or two IADLs may potentially be provided in the home or community. However, more than health dictates the need for more intensive long-term care services, an individual's wealth, and presence of family caregivers will also influence the site of care. For example, disabled individuals who are married and have children have been found to have a lower risk of nursing home entry.

Most elderly persons are physically active, able to care for themselves, and do not need long-term care. However, the prevalence of disability rises steeply with age. For example, in the US, only approximately 1 in 10 individuals aged 65–74 years is disabled, but roughly 7 in 10 individuals aged 85 years and older are disabled. Additionally, not all disabled persons are old. For example, individuals under the age 65 years with spinal cord injuries, advanced multiple sclerosis, traumatic brain injuries, developmental disabilities, and mental illnesses may all require some form of long-term care.

Who Provides Long-Term Care?

Although many people associate long-term care with nursing homes, the predominant provider of long-term care is the family. The predominant providers of care within a family have historically been spouses and adult children of elderly individuals and parents of younger individuals in need of services. Although several recent societal trends have worked against informal provision of services (e.g., greater female labor force participation and geographic dispersion of families), this is still the dominant type of long-term care. Among the community-dwelling US elderly with long-term care needs, 95% receive some informal care and two-thirds rely solely on informal care.

Elderly individuals almost universally prefer receiving care in their homes from family members. However, health, familial, and financial issues often precipitate the need for care from a formal provider. A broad continuum of services constitutes the formal long-term care marketplace. Although

 doi:10.1016/B978-0-12-375678-7.01002-6

nursing homes serve less than a quarter of the disabled elderly in the US, they are certainly the most expensive long-term care option and thoroughly studied.

In the US, roughly 1.6 million residents live in nearly 17 000 nursing homes. About two-third of all nursing homes are investor owned, about a quarter are nonprofit, and the remaining are government-owned facilities. Roughly half of all nursing homes are members of a chain and approximately 6% are hospital-based facilities. The average-sized facility has approximately 100 residents and the overall occupancy rate is approximately 88%. Historically, occupancy rates have been much higher within this industry because of the presence of supply constraints such as certificate-of-need (CON) and construction moratorium laws that attempt to limit the growth in beds in an effort to hold down the Medicaid expenditures. However, the recent growth in alternatives to nursing home care has likely competed away some of the 'healthier' nursing home residents to other care settings and lowered nursing home occupancy rates.

For individuals who can still live on their own, home care can range from periodic help with shopping and cleaning to full-time nursing help. Social support services such as meals on wheels, adult foster care, and adult day care, may enable individuals in need of long-term care to remain in the community. Assisted living facilities are residential settings that provide more supportive services than boarding houses but less medical care than a nursing home. Assisted living may provide lodging; meals; protective oversight; activities; and some assistance with medications, personal care, and ADL.

From an economic perspective, one intriguing development within the US long-term care market is the blurring of the roles of provider and insurer in CCRCs. Under this model, residents pay a large initial fee on entry and rent an apartment for an additional monthly fee in a community setting designed specifically for elderly individuals. As health declines, the individual may move on from the independent living section of the CCRC to onsite-assisted living and onsite nursing home care for additional charges. Given that this model is typically geared toward wealthier individuals, CCRCs make up a relatively small part of the US market.

Who Pays for Long-Term Care?

Similar to acute health care services, long-term care is paid by a number of sources. What is most striking about this sector in the US, relative to the acute health sector, is the lack of private insurance coverage. Less than 5% of all long-term care expenditures are paid by private insurance. Individuals who use long-term care typically pay it out of their own (or their family's) income and assets, or they must qualify for public coverage. Thus, long-term care represents the largest source of catastrophic costs for the elderly in the US. Although Medicare does cover some rehabilitative (or short-stay) nursing home care, the primary payers of long-term care services are state Medicaid programs. For example, Medicaid accounts for about half of all expenditures on long-stay nursing home services, which amounted to approximately $45 billion nationwide in 2009. Individuals must qualify for Medicaid by meeting income and asset criteria at the time of nursing home entry or by 'spending down' during their stay.

Although HCBS have been found to be associated with lower long-term care expenditures for individuals with certain care needs, most state Medicaid programs are more generous in covering nursing home services because of a perceived moral hazard problem. Individuals generally do not want to enter a nursing home, with research suggesting that these services are relatively price inelastic with respect to Medicaid eligibility policy. However, HCBS attracts some individuals who otherwise would have received care from family members and friends in the community. Thus, in recognition of this potential moral hazard problem, states are more likely to cover nursing home services relative to HCBS. Nevertheless, spending for Medicaid HCBS has grown substantially, increasing from $4 billion in 1992 to $22 billion in 2007. Nursing home expenditures have also increased over this period, HCBS grew from 14.5% to 31.6% of Medicaid long-term care spending between 1992 and 2007.

The emotional, physical, and financial burden on informal caregivers can be quite high. Historically, US long-term care policy has not financially reimbursed informal care provision by family members and friends. Although such services are not reflected in the national health accounts, never trigger a payment from an insurer; do not inflate the federal deficit, and are rarely included in any calculation of the overall cost of long term care; they nonetheless represent a genuine opportunity of cost burden. For example, if an adult child is taking care of an elderly parent, this individual is forgoing other work and leisure opportunities. Policymakers in the US have experimented with several measures to support informal care by family and friends with the idea that these savings might offset higher cost institutional services. For example, the 'cash and counseling' program, currently active in 15 states, provides Medicaid beneficiaries with a budget to hire their own personal care aides. Recent economic research in the US and elsewhere has begun to calculate the direct (e.g., opportunity cost) and indirect (e.g., health implications) costs of informal caregiving.

Government Regulation of Long-Term Care

Reflective of government spending over the past several decades, regulation in the US long-term care sector has largely been defined by the regulation of nursing homes where government continues to play a vital role in protecting a potentially vulnerable resident population. The reason for the high degree of government intervention and oversight is often thought to relate to the inability of many nursing home consumers to monitor quality effectively. Dating back over three decades, a number of reports and studies have documented low-quality care within this industry. In response to this issue, the US government has placed a number of restrictions on the industry. For example, the Nursing Home Reform Act was passed in 1987 mandating that nursing facility care should be more consistent with expert recommendations for assuring quality care. These recommendations included reduction in the use of physical restraints, prevention of

pressure ulcers, reduction of psychoactive medications, and some minimal staffing standards including the stipulation that a registered nurse must be on duty 24 h a day and all nurses' aides must be certified.

The US government is also an overseer of care via the survey and certification process. To accept the Medicaid and Medicare recipients, a nursing home must be annually certified via Centers for Medicare & Medicaid Services survey. Several alternative remedies may be imposed on facilities that receive a high number of deficiencies. These punishments include civil money penalties of up to $10 000 a day, denial of payment for new admissions, state monitoring, temporary management, and immediate termination. In addition to this survey process, certified nursing homes must fill out Minimum Data Set (MDS) assessments for every resident on a quarterly basis. Thus, the government generates an immense amount of quality information at a substantial cost. One estimate suggested that the survey and certification process costs the government nearly $400 million annually, which equates to approximately $22 000 per nursing home or $208 per nursing home bed. This figure does not include the indirect costs to the facility of the certification process, such as interacting with the regulatory agency, preparing for and hosting survey visits, gathering and providing data, and responding to complaint investigations. Experience from other sectors of the economy suggests that the indirect costs of the certification process to the nursing home are likely greater than the direct costs to the government.

Beyond setting and enforcing quality standards, examples also exist of market entry and price regulations in the US long-term care sector. Regulated barriers to entry are present in many long-term care markets via state CON laws and construction moratoria. Most of these state laws focus on nursing home beds, although states are increasingly grappling with whether and how to expand these policies to other long-term care settings, such as assisted living facilities. A CON law constrains market growth by employing a need-based evaluation of all applications for new construction. A construction moratorium is even more stringent in that it effectively prevents any market expansion. The stated rationale for these regulations is that lower capacity ultimately results in lower public expenditures, although research suggests that the repeal of these policies does not lead to increased state Medicaid long-term care expenditures. As a related barrier to entry, some states exercise greater scrutiny over the ownership status of nursing homes (e.g., New York State does not allow out-of-state for-profit chains to operate facilities in the state). Although infrequently used, an example of price regulation in the US long-term care sector is nursing home rate equalization laws. Both North Dakota and Minnesota prohibit nursing homes from charging a private-pay price above the state Medicaid rate.

Key Economic Questions in Long-Term Care

Nonpurchase of Private Long-Term Care Insurance

As noted above, relatively few individuals in the US purchase private long-term care insurance. Researchers have explored a number of potential supply- and demand-side explanations for this nonpurchase. On the supply side, research has observed that long-term care insurance premium pricing has relatively high loads compared to other types of insurance – that is, a lower portion of the premium dollar translates into benefits. These high loads are consistent with several supply-side market failures including transaction costs, imperfect competition, asymmetric information, and a range of dynamic contracting problems. Empirical support exists for the asymmetric information and dynamic contracting explanations. However, these supply-side factors cannot entirely explain the limited size of the market. Research suggests that even if actuarially fair policies (i.e., policies with zero load) were made available, the majority of elderly individuals would still not purchase these policies. Thus, research suggests that most of the nonpurchase relates to demand-side factors.

On the demand side, one explanation for the nonpurchase of long-term care insurance is incomplete information on the part of consumers. Many studies have found that individuals underestimate their need for long-term care or mistakenly assume it is covered by Medicare. Another possible demand-side explanation is that the form of the utility function may not be constant in the context of chronic health conditions. That is, individuals may place a lower value on consumption while in a nursing home than when healthy at home, which would serve as a disincentive to purchase long-term care insurance. Demand for long-term care insurance may also be limited by the availability of imperfect but less costly substitutes such as unpaid care provided by family members. Another potential explanation is that illiquid housing wealth can be used to insure long-term care. An individual may prefer to use their housing wealth in the event of a health shock rather than pay long-term care insurance premiums out of liquid wealth.

One prominent demand-side theory is that the Medicaid program 'crowds out' the purchase of long-term care insurance. Using simulation models, one study found that the implicit tax imposed by Medicaid (i.e., the part of the premium going to benefits Medicaid would have otherwise provided) explains why more than 60% of the wealth distribution does not purchase a policy. Importantly, the same researchers note that reducing the implicit tax of Medicaid on long-term care insurance would likely be an insufficient mechanism to expand the market, in part, because of the consumer misperceptions and supply-side failures described above.

Long-Term Care Quality

A number of studies have suggested poor quality of care in long-term care markets, especially the US nursing home market. A large health economics literature has focused on the economic explanations for low-nursing home quality. Economists have generally focused on four explanations for variation in the quality of nursing home care: public payment generosity; supply constraints; asymmetric information between nursing homes and patients; and macroeconomic factors.

The health economics literature on nursing home quality of care in the 1980s and 1990s was largely based on Scanlon's model in which nursing homes face two markets. One market

is for private-pay residents with downward sloping demand, and the other is for Medicaid residents who are insensitive to price. Scanlon's empirical work suggested the Medicaid side of this market could be characterized nationally by an excess demand. CON and construction moratoria policies had constrained growth in the supply of nursing home beds, and nursing homes preferred to admit the higher paying private patients. As a result, when a bed shortage existed, it was the Medicaid patients who would be excluded.

At the time, many noneconomists thought that the problem of quality in nursing homes could only be solved by raising Medicaid reimbursement rates. By incorporating a quality variable into Scanlon's model, several early research papers showed that raising Medicaid rates in a market with excess demand would result in nursing homes facing a reduced incentive to use quality of care to compete for the private patients. The decline in nursing home occupancy rates, repeal of CON laws in certain states, and emergence of improved data over the past decades have all contributed to a renewed interest in the relationship between the Medicaid payment and nursing home quality. Unlike the earlier research on this issue, results from more recent studies have generally found a modest positive relationship between the state Medicaid payment rates and nursing home quality. Importantly, the more recent studies provide little support for a negative relationship between the Medicaid payment level and quality.

Asymmetric information may also be a potential explanation for low-quality nursing home care. Although nursing home care is fairly nontechnical in nature, monitoring of care can often be difficult, and the quality learning period may be nontrivial relative to the length-of-stay in some instances. The nursing home resident is often neither the decision maker nor able to easily evaluate quality or communicate concerns to family members and staff. Furthermore, elderly individuals who seek nursing home care are disproportionately the ones with no informal family support to help them with the decision process. Finally, relatively few transfers occur across nursing homes. Movement among homes may be impeded by tight markets due to supply constraints such as CON and construction moratorium laws and health concerns regarding relocation (termed 'transfer trauma' or 'transplantation shock'). Thus, consumers may not be able to 'vote with their feet' by taking their business elsewhere.

To address this perceived lack of consumer information, the US government publishes a web-based nursing home report card initiative called 'Nursing Home Compare' (www.medicare.gov/NHCompare), which contains information on nurse staffing, regulatory deficiencies, and MDS-based quality indicators. If consumers use this information to make informed decisions about nursing home entry, then public information may help to improve quality. The existing literature to date suggests that the Nursing Home Compare report card initiative has led to a modest (but inconsistent) positive effect on nursing home quality of care. Key factors that may impede the use and efficacy of nursing home report cards include the heterogeneity in the preferences of short-stay and long-stay consumers; potential difficulties in accessing report card information during times of crisis; potential difficulties in interpreting report card data when the measures conflict or fail to provide a clear signal; key role of hospital discharge planners in the selection process; and limited choice set many nursing home consumers face due to rural markets, price, high occupancy, or other extenuating circumstances.

Macroeconomic factors such as wage rates for nursing home staff may also be important toward explaining the level of quality. For example, one study measures the extent to which nursing homes substitute materials for labor when labor becomes relatively more expensive. From a quality perspective, factor substitution in this market is important because materials-intensive methods of care are associated with greater risks of morbidity and mortality among nursing home residents. Indeed, as the market wage rises, nursing homes are more likely to employ labor-saving practices such as the use of antipsychotics.

Pay-for-Performance in Long-Term Care

Through the Medicare and Medicaid programs, the US government purchases significant amounts of nursing home services. Moreover, an emerging literature suggests poor nursing home quality results in higher Medicare spending for acute care services. As such, the government seeks to obtain high-quality services for Medicare beneficiaries. However, administrative pricing arrangements mean that – for many residents – nursing homes cannot charge higher Medicare or Medicaid prices for better quality. Moreover, the government cannot simply ask for a level of quality for Medicare beneficiaries. This set of circumstances can be analyzed using a principal agent model. In this instance, the 'principal' is the government, whereas 'agent' is the nursing home.

This principal–agent model in economics is useful in analyzing circumstances in which providers, such as nursing homes, are not driven by market forces to the level of quality desired by the purchaser and, further, where the purchaser cannot contract directly for a given level of provider quality. One way to induce nursing homes to improve quality is to make payments at least partly contingent on an indicator of nursing home effort to deliver high-quality care. Such indicators of nursing home effort are embodied in various structural (e.g., staffing), process (e.g., physical restraint use), and outcome (e.g., pressure ulcers) measures. These indicators only measure a few of the many dimensions of quality that a health care purchaser (and consumers) might care about, and each of them may require separate, costly efforts to generate improvement. That is, the structures and processes that create improvements in pressure ulcers might be largely distinct from what is needed to raise performance in lowering resident pain or depression. Purchasers must decide which dimensions of quality to target and consider how outcomes on unrewarded dimensions of performance might be affected. Pay-for-performance schemes in this way introduce a form of price flexibility that rewards desirable performance. Theoretically, the effectiveness of payments contingent on quality measures depends principally on the relative magnitude of expected costs and benefits to the provider of improving quality. That is, do expected incremental nursing home payments exceed the costs to facilities of supplying the desired level of quality? Costs should be thought of broadly here and may include, for

example, the value of additional unreimbursed time spent with patients or investments in information technology.

Pay-for-performance arrangements also have the potential for unintended consequences. Critics of paying for quality in health care have identified a number of drawbacks that might arise from the introduction of such schemes. The principal category of unintended consequences that might result from pay-for-performance is generally termed gaming where participants find ways to maximize measured results without actually accomplishing the desired objective of improved quality of care. In the nursing home setting, providers or administrators might 'game' incentive systems by miscoding diagnoses or services or selecting patients on the basis of the likelihood of a positive outcome or compliance with treatment protocols rather than need. Selecting healthier patients for treatment may reduce aggregate health benefits; miscoding may also have longer run effects on quality because of missed opportunities to identify and improve low quality. A second major concern with paying for quality is known in the economics literature as the multitasking problem. If the goal of the payer is multidimensional and not all dimensions can be measured and 'paid on' (e.g., resident quality of life), compensation based on available measures will distort effort away from unmeasured objectives that may be important to patient well-being. Finally, concerns have been raised about the impact of paying for quality on intrinsic motivation, cooperation, and professionalism, particularly among physicians.

Recent concerns have also been raised about the impact of market-based approaches for quality on racial and ethnic disparities in health care. In the context of the nursing home market, published research has described its two-tiered nature, with the lower tier consisting mainly of residents with Medicaid-financed care and having fewer nurses, lower occupancy rates, and more health-related deficiencies. These low-performing facilities are disproportionately located in the poorest communities and are more likely to serve African-American residents than are other facilities. Even within markets, African-American and poorly educated patients have been found to enter the worst-quality nursing. Although pay-for-performance initiatives are typically aimed at improving quality of care broadly, it is important to monitor whether these initiatives further disadvantage poor performing providers and the individuals they serve.

Cost-Effectiveness of Home and Community-Based Services

Individuals generally prefer care in least restrictive setting possible, and for certain individuals with less intensive care needs, it may be possible to provide lower per capita cost care at home or community relative to a nursing facility. However, the historic institutional bias in long-term care coverage relates partially to a perceived moral hazard problem (or woodwork effect) whereby publicly financed noninstitutional services substitute for informal services previously provided by family members and friends. Program administrators have found it very difficult to structure coverage such that only individuals who otherwise would have entered nursing homes utilize noninstitutional services. States have employed targeting (or screening) mechanisms in an attempt to limit care to only those individuals who otherwise would have accessed nursing home care.

If targeting were perfect, then the noninstitutional treatment model would need to be only marginally less costly than the institutional model to generate savings. However, as targeting becomes less perfect, the aggregate savings from noninstitutional care needs to increase in order to cover the increased costs associated with the moral hazard effect. The empirical literature has generally supported the idea that spending from increased HCBS utilization typically exceeds the savings from decreased nursing home utilization. However, this type of cost analysis is distinct from a cost-effectiveness analysis, in which differences in costs are benchmarked against differences in outcomes. Even if rebalancing toward HCBS is associated with higher aggregate costs, the services may still be cost-effective due to an even greater increase in aggregate effectiveness. Toward the end, a number of research studies have supported the idea that psychosocial outcomes such as life satisfaction, social activity, social interaction, and informal caregiver satisfaction were higher under HCBS. Moreover, unmet needs have been shown to decrease under HCBS. To date, research has not formally balanced the costs and benefits of HCBS.

Effects on Informal Caregiving

Economic theory suggests a range of supply- and demand-side factors may influence the provision of informal caregiving. On the supply side, given the potential substitution of formal and informal care services, changes in the generosity of public payment for home health care services may influence the provision of informal care. Research suggests that older US adults with functional limitations who were exposed to more restrictive payment caps offset reductions in Medicare home health care with increased informal care, although this effect is only observed for lower income individuals. Direct public payment of family caregivers may also influence informal caregiving. The US 'Cash and Counseling' program, currently active in 18 states, provides Medicaid enrollees with a monthly cash allowance to purchase personal assistance and related goods and services. The majority of recipients purchase this care from family members. In a randomized three-state demonstration evaluating Cash and Counseling against the traditional agency-directed model of home care, the program was found to reduce some unmet needs and greatly enhance quality of life, but Cash and Counseling increased overall program spending.

Several economic analyses have considered the effect of demand-side factors on the provision of informal care services. Using US data, the availability of immediate family such as a spouse or adult children, being male, being a minority, and owning a home were all associated with a greater likelihood of informal care use. When income is treated as exogenous, studies have found that higher income is associated with a lower probability of informal care use. However, when the Social Security 'benefit notch' was used as instrument for income, higher permanent income is not found to have a statistically meaningful effect on the provision of informal care among older adults with lower education.

Integration of Long-Term Care with Other Health Services

Individuals who require long-term care services typically also require a mix of primary, acute, postacute, and palliative services at different times. The coordination of these different services has become a major issue within the US health care system. Importantly, the coordination of health care services at the delivery level relates directly to the financing and payment of those services.

At the financing level, the presence of multiple payers in health care is known to introduce conflicting incentives for providers, which may have negative implications for cost containment, service delivery, and quality of care. The fundamental issue is that the actions of one payer may affect the costs and outcomes of patients covered by other payers. These 'external' costs and benefits can occur both within and across health care settings, and little incentive exists for a payer to incorporate them into payment and coverage decisions. As a result, the behaviors of health care payers – even public payers – often deviate substantially from the social optimum.

This observation is particularly relevant in regards to the coverage of acute and long-term care services in the US The federally run Medicare program provides a set of insurance benefits for virtually all individuals age 65 years and older, regardless of income, and for younger people with disabilities 2 years after they qualify for Social Security's disability benefit. Medicaid, a state-run program jointly funded by the state and federal governments, provides coverage for its low-income enrollees that supplements Medicare coverage. Many individuals who are dually eligible for both Medicare and Medicaid require both extensive acute and long-term care services. However, given the bifurcated coverage of acute and long-term care under Medicare and Medicaid, neither program has an incentive to internalize the risks and benefits of its actions as they pertain to the other program. Each program has the narrow interest in limiting its share of costs, and neither program has an incentive to take responsibility for care management or quality of care. For example, under the traditional benefit structure for duals, little incentive exists for state Medicaid programs to enact policies to lower Medicare-financed hospitalizations because they do not accrue any of the potential savings. Indeed, state Medicaid programs often enact policies such as bed-hold payments that increase hospital and postacute expenditures for the Medicare program. A model that blends Medicare and Medicaid financing introduces a stronger incentive to minimize transitions for dually eligible beneficiaries from Medicaid-financed nursing home care, for example, to higher cost Medicare-financed hospital care.

Payment structure also has implications for the coordination of care. Cost-shifting occurs for reasons beyond the fragmentation of financing across programs. For example, the high rate of 30-day hospital readmissions from Medicare-financed skilled nursing facilities is an example of poor coordination within the Medicare program. Traditional fee-for-service payment creates little incentive for providers to manage the volume and intensity of services because providers are rewarded with greater revenue when they deliver more services. Indeed, hospitals are rewarded with higher revenue when beneficiaries are readmitted to the hospital.

Through risk-based capitation, managed care potentially encourages more efficient care delivery. Under this model, a single entity receives a fixed predetermined monthly payment (i.e., capitation rate), which provides the incentive to minimize wasteful care. Ideally, under capitation, hospitals would not be rewarded when individuals are readmitted. Similarly, other risk-based models such as accountable care organizations, bundled payment, global budgeting, and medical homes also provide similar incentives to coordinate care in ways that could reduce inefficient medical and long-term care service use.

With respect to care delivery, the coordination of financing and payment can be thought of as necessary, but not sufficient, conditions for the coordination of services. For example, at the delivery level, care coordination activities might include case management, team-based care models, patient education, management of care transitions, communication protocols for providers, and shared clinical and social information. However, without an alignment in payment and financing in which providers can internalize the costs and benefits of their actions, there is little reason to suspect any sustainable coordination in service delivery at the ground level.

See also: Home Health Services, Economics of. Long-Term Care Insurance

Further Reading

Brown, J. R. and Finkelstein, A. (2009). The private market for long-term care insurance in the US: A review of the evidence. *Journal of Risk and Insurance* **76**(1), 5–29.

Grabowski, D. C. (2006). The cost-effectiveness of noninstitutional long-term care services: Review and synthesis of the most recent evidence. *Medical Care Research and Review* **63**(1), 3–28.

Grabowski, D. C. (2007). Medicare and Medicaid: Conflicting incentives for long-term care. *Milbank Quarterly* **85**(4), 579–610.

Grabowski, D. C. (2008). The market for long-term care services. *Inquiry* **45**(1), 58–74.

Grabowski, D. C. and Norton, E. C. (2006). Nursing home quality of care. In Jones, A. M. (ed.) *The Elgar companion to health economics*, pp. 296–305. Cheltenham, UK: Edward Elgar Publishing, Inc.

Grabowski, D. C., Norton, E. C. and Van Houtven, C. H. (2012). 'Informal Care.' In Jones, A. M. (ed.) The Elgar Companion to Health Economics, pp. 318–328, vol. 2. Cheltenham, UK: Edward Elgar Publishing, Inc.

Konetzka, R. T. and Werner, R. M. (2010). Applying market-based reforms to long-term care. *Health Affairs* **29**(1), 74–80.

Norton, E. C. (2000). Long-term care. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, pp. 955–994. Amsterdam: Elsevier Science.

Scanlon, W. J. (1980). A theory of the nursing home market. *Inquiry* **17**(1), 25–41.

Long-Term Care Insurance

RT Konetzka, University of Chicago, Chicago, IL, USA

Glossary

Activities of daily living Self-care tasks required on a regular basis, such as bathing, dressing, toileting, transferring in and out of bed, and eating.
Formal long-term care Paid care.
Informal long-term care Unpaid care provided by family or friends.
Instrumental activities of daily living Tasks related to the ability to live independently, such as housekeeping, using a telephone, shopping, preparing meals, and money management.
Long-term care Assistance with functional and/or cognitive impairments on an ongoing basis.
Medicaid crowd-out Reduced demand for private insurance due to the availability of Medicaid as an alternative.
Policy lapse Intentionally or unintentionally allowing an insurance policy to expire or become invalid.

Introduction

Long-term care is a sector of the healthcare industry that is growing in importance with the aging of populations around the world. In the United States, according to the Congressional Budget Office, expenditures on long-term care totaled US$135 billion in 2004 and are expected to double in several decades. People with long-term care needs generally have chronic conditions and associated functional and/or cognitive limitations that require assistance with activities of daily living (bathing, dressing, toileting, transferring, eating) or instrumental activities of daily living (housekeeping, using a telephone, shopping, preparing meals, money management). These types of needs can be served in a variety of settings: in the home (formally by paid home care or informally by family and friends), in a nursing home, in an assisted living facility, or in an adult day care center, among others. Although the lines between acute care, postacute care, and long-term care have become blurred for long-term care recipients as more and more high-tech services formerly provided only in hospitals are now administered in a variety of settings, an ongoing need for assistance with functional or cognitive limitations remains the defining feature of an individual with long-term care needs. Although some recipients of long-term care are under the age of 65, the majority are elderly.

Long-term care is growing in importance not only due to demographic shifts but also due to the emergence of chronic conditions as a primary healthcare challenge. Much of the developed and parts of the developing world, having experienced both the eradication of many infectious diseases and the benefit of technological advances that lessen the mortality from the largest causes of death in earlier eras, are now struggling with a growing prevalence of chronic health conditions, sometimes exacerbated by poor health behaviors. These are often costly conditions that require ongoing care over many years, and failure to access appropriate chronic care can lead to a greater need for acute care. Nonetheless, payment systems have generally not adapted to the growing importance of chronic conditions and associated long-term care needs. In the United States, Medicare, the publicly run health insurance system for the elderly, was designed to cover temporary acute care and explicitly disavows responsibility for covering long-term care. Medicaid, designed to cover healthcare needs of the poor, has by default become the dominant public payer in long-term care (approximately two-thirds of nursing home residents at any given point in time are on Medicaid), but as this was not the original intent of the program, substantial gaps and inefficiencies remain. In addition to lack of recognition or foresight about the growing financial burden of long-term care, many societies express some ambivalence about who should be responsible for long-term care, as much of it is relatively low-tech and can potentially be provided by family members. The resulting lack of intentional and systematic financing is a key feature of the economics of the long-term care sector that distinguishes it from other healthcare sectors involving the elderly and sets the stage for a private long-term care insurance market.

Developed nations around the world face similar situations in terms of demographic change and a growing need for long-term care that was not entirely recognized or anticipated when coverage for acute care needs was evolving. Depending on their resources, cultural norms, ideology, and existing healthcare delivery and payment infrastructure, countries have followed a variety of approaches to covering long-term care. Several have opted for national long-term care insurance systems (e.g., Germany, Japan); others have incorporated some long-term care services, especially home- and community-based services, into existing social insurance programs (e.g., Denmark); and some rely on a combination of self-funding, private long-term care insurance, and a safety net of public funding as a payer of last resort (e.g., the United States). Countries that rely mainly on private financing – or would like to ease the burden on public coffers – have a stake and interest in the existence and survival of private long-term care insurance markets.

The risk of needing long-term care is, in theory, an appropriate risk to be insured against. The average probability among aging individuals of needing long-term care is not trivial and is associated with substantial and unevenly distributed cost, but which individuals will experience the highest costs is seemingly random when viewed at the typical age of insurance purchase (50–65). On average, individuals

 doi:10.1016/B978-0-12-375678-7.00921-4

turning 65 will need some type of long-term care for 3 years, but half will have no private out-of-pocket expenditures, due to lack of either need or the availability of informal care to meet low-level needs. However, more than 1 in 20 is projected to spend more than US$100 000 out of pocket in 2005 (Kemper *et al.*, 2005). According to MetLife Mature Market Institute (2008), nursing home care costs more than US$70 000 per year on average, which implies that only a small minority of individuals can finance an extended stay out of pocket. The skewed distribution of uncertain costs associated with long-term care is a feature that would normally bode well for a robust insurance market.

Despite the conceptual appropriateness of a robust long-term care insurance market, only approximately 13% of the elderly population in the United States reports having long-term care insurance. Most policies are purchased on the individual-payer market, as group long-term care insurance policies are relatively rare. Benefit eligibility is usually triggered with medical certification of a minimal level of functional dependence, defined as assistance needed with activities of daily living. The vast majority of policies cover home care as well as nursing home care for a given number of years, but benefits are generally paid as a set per diem amount to be applied toward a given service as opposed to covering the total cost of the service. Many policies adjust benefits for inflation over time. Policies typically cost several thousand dollars per year, but costs can range substantially depending on age and health status. Individuals who exhibit signs of existing or imminent long-term care need (e.g., those who already have mild cognitive or functional impairment) are generally ineligible for policies at any price. Long-term care insurers are generally not allowed to raise premiums over time for an individual whose health risk increases, but they can adjust for changes in risk for an entire class of policyholders if payouts are higher than expected. Individuals who fail to pay premiums (policy lapse) forfeit all benefits and all premiums paid previously; few policies to date have built in nonforfeiture benefits that would allow individuals to recoup some of the investment in a lapsed policy. Most state insurance regulations include safeguards that help to avoid unintentional lapse.

Theory of Demand for Long-Term Care Insurance

Economists have generally modeled the behavior of consumers in the decision to purchase insurance using a standard expected utility framework; i.e., insurance will be purchased if expected utility with insurance is greater than without insurance. The theoretical underpinnings of long-term care insurance differ from these standard theories of insurance purchase mainly due to the role of family and bequests. That is, when consumers consider purchasing insurance against the risk of long-term care costs, they consider not only the direct expected utility of smoothing consumption but also the utility elicited through the behavior of a spouse or children and the altruistic utility derived from leaving a bequest to heirs.

The prominent theoretical model in this area is Pauly (1990), with an extension by Zweifel and Struewe (1998). Assuming imperfect annuity markets, Pauly considers expected utility optimization under several scenarios: single elderly with no children and no bequest motive, with differential quality, and with adult children and a possible bequest motive. Expected utility in the presence of a spouse is also discussed briefly. The model is aimed at explaining purchase (or nonpurchase) of long-term care insurance among middle-income individuals, as nonpurchase is obvious among very poor individuals likely to qualify for Medicaid and among richer individuals who can easily self-insure. The expected lifetime utility function (EU) to be maximized is given as

$$\max_{C_t} \mathrm{EU} = \sum_{t=1}^{H} p_t^h U(C_t) + \sum_{t=1}^{H} p_t^s \overline{U}^s \, s.t. \, \overline{W} \geq \sum_{t=1}^{H-S} C_t + S\overline{X}$$

where H is the maximum length of life, p_t^h is the probability of surviving to period t in the healthy state and p_t^s in the sick state (in need of long-term care), C is dollars of consumption, and $\overline{U}^s$ is the level of utility if one is in need of long-term care and consuming $\overline{X}$ dollars worth of care, the only type of desired consumption in the sick state. $\overline{W}$ represents initial wealth, which is assumed to be substantially larger than $S\overline{X}$ such that the individual initially has enough wealth to pay his or her maximum long-term care costs and is unlikely to qualify for Medicaid. Individuals choose C to maximize expected utility.

The model under each scenario predicts that the low demand for long-term care insurance may be rational. The case in which single elderly individuals have no children and no bequest motive is straightforward given the assumptions of the model. Because Medicaid exists as a safety net when wealth is exhausted, the only benefit to purchase of long-term care insurance is to increase consumption in the sick state, the marginal benefit of which is defined to be zero. Although this assumption is restrictive, one might imagine that the marginal benefit of additional consumption while in need of nursing home care is at least low if not zero, in which case the same conclusion would result.

The case in which private insurance enables access to higher quality nursing home care than that obtained under Medicaid funding is a realistic one in that nursing homes with large Medicaid populations are generally considered to be of lower quality. Thus, one might expect middle-class individuals to purchase long-term care insurance if they value quality. However, Pauly argues that this would rarely be the case because one cannot pay an incremental premium to purchase an incremental quality supplement to Medicaid. Rather, the purchase of private insurance replaces Medicaid. Thus, a consumer would have to value the incremental quality of a privately financed nursing home over and above a Medicaid-financed nursing home stay enough to outweigh the additional cost of foregoing Medicaid completely and paying private long-term care insurance premiums. Relatively few individuals are likely to have this high valuation for incremental quality.

A similar argument applies to the case in which bequests are valued, i.e., the individual receives utility from leaving wealth to his or her heirs. One would expect that valuing bequests would lead to a higher propensity to purchase long-term care insurance, as private insurance for nursing home care allows an individual to retain wealth in the sick state in contrast to relying on Medicaid, which requires the exhaustion

of one's wealth before coverage begins. However, the value of bequests would have to be quite large in order to lead to purchase, for two reasons. First, purchase of insurance decreases consumption not only at time t but also in the future if the person remains in a healthy state, so additional savings may provide greater utility than insurance purchase when bequests are valued. Second, although insurance may be preferable to savings if the individual lives a long time with chronic illness, this scenario is unlikely because chronic illness is generally associated with earlier mortality. Thus, even if bequests are valued, purchase is unlikely unless the utility from bequests is high and does not decline sharply with age.

The bequest argument is more complicated when a spouse is involved rather than just adult children. In this case, Pauly argues that both household consumption and income may be affected if one spouse enters a nursing home, depending on the extent to which these are joint. Demand for long-term care insurance may be relatively high if consumption of the nonsick spouse is substantially affected by the nursing home stay of the sick spouse.

Finally, perhaps the most important contribution of Pauly's paper is the introduction of intrafamily bargaining into the conceptualization of demand for long-term care insurance, drawing on earlier work in the bequest literature, which posited that parents use bequests to elicit desired attention or caregiving from children. Pauly modifies this premise somewhat to argue that, once in the sick state, parents will have little control over consumption or bequests, such that parents choose whether to purchase insurance in the healthy state but that children control the level of care in the sick state. Parents may prefer care from children and may want to purchase long-term care insurance to preserve bequests that the parent values altruistically and with which to elicit caregiving behavior on the part of children. However, as children decide on the level of care in the sick state, children are subject to moral hazard associated with the presence of insurance. That is, children will choose more formal care (nursing home placement) in the presence of insurance than what the parent would prefer because the price they face is lower than in the absence of insurance. Anticipating this moral hazard effect on the caregiving behavior of children, the parent may be better-off not purchasing insurance and getting the higher level of care from children.

Zweifel and Struewe (1998) formalize this intrafamily bargaining argument using a principal-agent framework and a two-generation model that is independent of assumptions about altruism. The elderly parent chooses consumption and whether or not to purchase long-term care insurance to maximize expected utility, and the amount of care provided by children is an argument in the utility function in the sick state. The child maximizes his or her own expected utility, choosing consumption and the amount of care to provide if the parent enters the sick state. By providing care, the child is presumed to forego work in the labor force but also to expect a higher bequest, as less will be spent by the parent on formal long-term care. Zweifel and Struewe show that, under these circumstances, the child's response to purchase of long-term care insurance depends heavily on the child's wage rate. At low wages, where one might expect the most caregiving, the presence of insurance is most likely to produce a moral hazard effect. Anticipating this response, purchase of long-term care insurance is often not in the best interest of parents who desire caregiving by their low-wage children, for the same reasons that Pauly posited.

The Pauly model is general and intuitive in many respects, explaining its pervasive use and longevity in the study of long-term care insurance. However, it has some limitations and may be dated in some ways. First, it does not formally model joint decision-making with a spouse, one of the most common scenarios among potential purchasers. Second, it is assumed that parents prefer care from children, which may be an outdated notion for many families. Preferences may depend importantly on the severity and type of long-term care needs, on the relationship between the parent and the child, and on the extent to which a parent prefers to stay independent and not burden the child. Third, Pauly models long-term care insurance as only nursing home insurance, whereas the vast majority of long-term care insurance policies now cover home care and other community-based options as well as nursing home care. Home care is generally considered much more desirable than nursing home care, so more parents may prefer it to informal care, and it may entail a completely different set of family dynamics; for example, informal care may be a complement to formal home care rather than a substitute for it. Thus, there exists a need for updated theoretical models of the demand for long-term care insurance that consider these factors.

Theory of Supply of Long-Term Care Insurance

Research on long-term care insurance to date has focused largely on the demand side, with relatively little theoretical or empirical work on the supply side. As in other insurance markets, insurers consider the potential for adverse selection and moral hazard in deciding whether to offer a product, at what price, and with what attributes. One of the few papers to consider this perspective is Cutler (1993), which discusses both adverse selection and moral hazard as important potential market failures in long-term care insurance markets. Adverse selection may be a more serious concern in long-term care insurance than in acute health insurance because the elderly and near-elderly population is naturally more heterogeneous in health status than a younger population, and this heterogeneity may not be observable to insurers. Thus, there is greater potential in this population for the existence and use of private information leading to a sicker risk pool than anticipated by insurers when setting price. Cutler also raises the issue of long-term intertemporal risk. In other types of health insurance in which premiums are set annually, prices can be reconciled with unanticipated trends in claims and provider prices fairly quickly. In long-term care insurance, however, because the event being insured today may not occur for another 20 years, insurers face the risk of rising prices over time that cannot be diversified across a risk pool. Thus, insurers generally shift this risk to consumers. Fears about the extent of adverse selection and moral hazard, coupled with this intertemporal risk and a lack of claims experience for this relatively new product, has led to strict underwriting, indemnity policies, high administrative loads, and consequently 'expensive'

premiums that may not seem affordable or of good value to many potential purchasers.

Distinctive Features of the Long-Term Care Insurance Market and Related Empirical Evidence

Long-term care insurance has several key features distinguishing it from acute care insurance: the role of the family (especially adult children), low prevalence of insurance, and greater concern about adverse selection. A relatively small but growing body of empirical work on long-term care insurance reflects these features and can be broadly categorized into research on intrafamily decision-making, price and other determinants of purchase and nonpurchase (including Medicaid crowd-out), and adverse selection. As a whole, the evidence is consistent on a few aspects of this market – for example, that Medicaid crowd-out exists – but on other aspects, the evidence is often sparse, inconsistent, incomplete, and inconclusive. Each of these categories is discussed below, followed by a discussion of the remaining theoretical and empirical gaps.

Intrafamily Decision-Making

As established by Pauly, the role of families in decision-making, insurance purchase, and provision of long-term care is a feature of the long-term care insurance market that distinguishes it from other types of health insurance. However, empirical research has not been able to substantiate the contention in Pauly's model that children will be more likely to institutionalize parents in the presence of insurance, a key premise underlying the rational nonpurchase of long-term care insurance among parents with adult children. The main study to address this issue (Mellor, 2001) used Health and Retirement Study (HRS) data in a longitudinal study design; the HRS is one of the few national data sets to include questions on long-term care insurance and is used in the vast majority of studies noted in this article. Mellor found point estimates generally in the expected direction – institutional long-term care use was more likely in the presence of insurance – and with potentially meaningful magnitudes, but the results lacked statistical significance. However, the study used a measure of insurance from the early years of HRS that was later shown to be subject to measurement error and included a shorter panel of data than is now available, limiting power. Thus, current evidence cannot establish conclusively how the presence and preferences of adult children impact long-term care insurance purchase and subsequent long-term care provision, and the need for further research remains.

Evidence on the role of family in long-term care insurance and provision is also tied to the bequest literature, as the desire to leave a bequest to one's heirs has often been posited as a potential motivator for long-term care insurance purchase. If a bequest is desired, one might think of long-term care insurance as bequest insurance, because in the absence of insurance, one's saving may be needed for long-term care costs, thus reducing or eliminating the prospect of a bequest. Early empirical evidence using direct queries about the desire to leave a bequest found no support for such a bequest motive in insurance purchase decisions (Sloan and Norton, 1997). A recent working paper, however, looks indirectly at long-term care insurance purchase to distinguish precautionary savings motives from bequest motives in savings behavior late in life (Lockwood). The main premise is that a precautionary savings motive is consistent with purchase of long-term care insurance, as the underlying goal would be to ensure availability of resources for healthcare needs. Low levels of long-term care insurance purchase are therefore indicative of a strong bequest motive in savings behavior, as the precautionary savings motive is ruled out. The author reconciles the strong bequest motive with low levels of long-term care insurance by suggesting that insuring the bequest is not valuable enough to justify the purchase of currently available long-term care insurance policies. This may explain the apparent lack of support for the bequest motive found in earlier studies.

Determinants of Purchase and Nonpurchase

Compared with acute care health insurance, the demand for which is generally thought to be a function of health, income, price, and risk aversion, the demand for long-term care insurance appears to be more complicated. The low prevalence of long-term care insurance has engendered numerous studies of why people do or do not purchase it. Evidence on private long-term care insurance (LTCL) prevalence suggests that among the elderly and near-elderly, the younger, healthier, and more educated people are more likely to have LTCI, and that there is some relationship, most likely nonlinear, between purchase of LTCI and income and assets. Although those in the lowest income and asset groups are not likely to purchase LTCI because it is expensive (generally several thousand dollars per year) and because they face a lower 'price' of Medicaid in terms of spending down assets to qualify, those in the highest income and asset groups may also not purchase insurance because they can self-insure. Therefore, it is often the 'middle' income and asset groups that are the most likely purchasers. Earlier studies cited Medicaid crowd-out, underestimation of risk, and the presence of adult children or bequest motives as potential reasons for nonpurchase but found mixed or inconclusive empirical results. Norton (2000) provides a useful summary of these arguments and the earlier evidence on purchase and nonpurchase. Many of these studies were limited by reliance on cross-sectional analyses, which precludes the establishment of a causal link between the predictors and the outcome.

More recent studies have taken advantage of exogenous variation in Medicaid policy and state and federal tax policies to move toward causal inference in estimating the demand for long-term care insurance and specifically to examine the issue of Medicaid crowd-out, i.e., the substitution of private insurance for public insurance when public insurance exists. Because Medicaid has become a primary payer of long-term care services, both in nursing homes and in the community, crowd-out is a potential obstacle to any expansion of the private long-term care insurance market. These recent studies generally find that Medicaid crowd-out is substantial and suggest that even a tightening of Medicaid eligibility rules would not be effective in mitigating crowd-out. Brown and Finkelstein (2008) argue, using a utility-based model and simulation, that

Medicaid crowd-out can explain nonpurchase of long-term care insurance for at least two-thirds of the wealth distribution. The large crowd-out effect stems from the large 'implicit tax' that Medicaid imposes on private insurance benefits in that the majority of private insurance benefits go toward covering services that Medicaid would have paid in the absence of private insurance. Thus, consistent with Pauly's reasoning, the value of a private policy to consumers is incremental whereas the premium derived from the total package of benefits is not. Although one might argue with the extent of the income distribution that is potentially affected, the existence of some degree of crowd-out is a reasonable conclusion. As Medicaid is generally incomplete insurance relative to private coverage, Medicaid crowd-out of private long-term care insurance may increase the overall risk exposure of the population.

It has often been posited that supply-side market failures contribute to low demand for long-term care insurance because these market failures result in undesirable policy attributes and a perception by consumers that the policies are not of good value. Value of an insurance product may be perceived as low if the administrative load is high, i.e., if the discounted expected present value of premiums far exceeds the discounted expected present value of benefits. In turn, concerns about substantial adverse selection, moral hazard, and, in the case of long-term care insurance, undiversifiable intertemporal risk may contribute to high administrative loads; these are the market failures. Brown and Finkelstein (2007) calculate that the average administrative load on long-term care insurance is 51%, substantially higher than loads estimated in other private insurance markets. This estimate includes the probability of lapse, in which case consumers generally forfeit all benefits. However, the authors also argue that despite these high loads, supply-side factors cannot explain the majority of nonpurchase of long-term care insurance. The argument is based mainly on the fact that administrative loads vary substantially by gender, with women facing much lower loads, yet women still do not purchase long-term care insurance at much higher rates than men. Thus, it is demand rather than supply that drives the behavior. In particular, the effect of Medicaid crowd-out is possibly much stronger than the effect of supply-side attributes.

Several studies have attempted to estimate a price elasticity of demand for private long-term care insurance. Cramer and Jensen (2006) combined HRS data with estimated prices derived from published rate schedules of several major insurers to calculate an estimated price elasticity of −0.23 to −0.87, indicating that new purchase of long-term care insurance is relatively price inelastic. Courtemanche and He (2009) also used HRS data, but derived an exogenous change in price using a change in federal tax treatment of long-term care insurance (new eligibility of long-term care insurance premiums to be deductable as a medical expense under the Health Insurance Portability and Accountability Act of 1996) combined with marginal income tax rates. They found a price elasticity of −3.9, indicating that purchase of long-term care insurance is highly elastic. These disparate results may perhaps be explained by the fact that identification was derived from different parts of the income spectrum, but in any case, the need for further research in estimating the determinants and elasticities of demand remains.

Adverse Selection

The potential for adverse selection is a concern for insurers of any type of event. Under adverse selection, potential purchasers have more information about their own risk than what is available to insurers and use this private information to assess the value of a policy. Because premiums do not account for the private information, riskier individuals are more likely to find the policy of value than less risky individuals, with the result that the pool of actual purchasers is riskier than what insurers would expect given an actuarially fair premium – a situation that is not sustainable in the long run. The potential for adverse selection is arguably greater in long-term care insurance than in other types of healthcare insurance for several reasons. First, the typical purchaser of long-term care insurance is elderly or near-elderly, and health states become more heterogeneous with age. Thus, the potential for private information about one's health risk is greater in an elderly population than in younger populations. Second, the market for long-term care insurance is small and largely based on individual policies rather than group policies. Thus, the broad diversification that can be achieved through, for example, employer-based group health insurance is not currently possible in long-term care insurance. In the one rigorous and broad-based study of this issue, Finkelstein and McGarry (2006) find empirical evidence in the HRS for this type of adverse selection in that individuals with private information that they are at high risk are more likely to purchase long-term care insurance. However, they find that it is balanced by favorable selection into insurance by individuals who have private information that they are more risk averse (but healthier). Thus, although adverse selection exists in long-term care insurance, the overall insured pool is not sicker than what insurers expect when calculating premiums.

The emergence of personalized medicine and genetic testing has led to increasing interest in genetic adverse selection. The availability of genetic tests for several serious diseases associated with long-term care needs makes this an especially salient issue for the long-term care insurance market, and the small size and individual-payer nature of the market has proved to be useful in studying this type of adverse selection. Recent evidence finds that, not surprisingly, people found to be at genetic risk for Huntington's disease or Alzheimer's disease are much more likely to purchase or to plan to purchase long-term care insurance than others – 2.3 times as likely in the case of Alzheimer's disease (Taylor *et al.*, 2010) and five times as likely in the case of Huntington's disease (Oster *et al.*, 2010). Although the absolute prevalence of these genetic markers in the population is small, this evidence provides a challenge not only for insurers but also for policymakers interested in balancing privacy rights against the need for a sustainable insurance market.

A related issue to adverse selection at the time of purchase is that of dynamic adverse selection. Because long-term care insurers are generally not allowed to raise premiums over time for an individual whose health risk increases, one can conceptualize purchase of long-term care insurance as insurance against reclassification into a higher risk category, much as in life insurance markets. In theory, if premiums are actuarially fair when purchased but are paid over time, individuals may decide to drop insurance (lapse) if their risk ex post appears

lower in later years than when they bought the policy. Lapse thereby becomes a mechanism for ex post adverse selection, a dynamic inefficiency in the insurance market that puts upward pressure on premiums for those remaining in the risk pool. Finkelstein *et al.* (2005) examine this issue in long-term care insurance markets, conceptualizing lapse as a rational response to a reevaluation of health risk. Empirically, the authors find that respondents who have 'ever let a LTCI policy lapse' are less likely to have a nursing home stay within 5 years than similar respondents who bought and kept policies, providing support for their hypothesis that lapse represents *ex post* adverse selection. However, the results may also be explained by a moral hazard effect. Using more years of data and testing for a broader variety of covered services and health status measures not subject to moral hazard, Konetzka and Luo (2011) find that lapse is driven more by financial reasons than health-related reasons, resulting in a healthier insured pool remaining. Individuals who lapse are generally poorer, less educated, less healthy, and more likely to be racial and ethnic minorities than those who retain their policies. Thus, although *ex post* adverse selection may occur for some groups of purchasers, it is not a primary driver of lapse and lapse as a whole is unlikely to affect the risk pool adversely. In addition, lapse rates are generally considered low in long-term care insurance relative to other insurance markets.

Given the aging of the population and the associated need for solid theory and evidence to inform public policy, the need for further research on long-term care insurance markets is great. Because long-term care insurance is different in marked ways from acute care health insurance, lessons learned in those markets may not be transferrable. To date, however, the theoretical foundation and empirical evidence on purchase and retention of long-term care insurance policies is far from complete. Although a growing body of evidence supports the existence of some degree of Medicaid crowd-out, the other determinants of policy purchase and retention remain murky, and evidence on the extent and nature of adverse selection is sparse. Two areas in particular are in need of better theoretical understanding and empirical research. First, although the role of the spouse and extended family is central in long-term care issues, still very little is understood about how intrafamily decisions are made with respect to long-term care insurance purchase and long-term care utilization. More sophisticated modeling of joint decision-making about this issue is paramount. Second, the literature on private long-term care insurance largely ignores moral hazard. Clearly, insurance ownership is only one key attribute of the market. Equally important is how insured individuals behave once they become insured. The significance of moral hazard – the utilization of long-term care services that are due to the presence of insurance and that would not be purchased without insurance – parallels the significance of adverse selection. Both are important because they could alter the cost of the insurance and thus the amount of the payout relative to the premiums.

Public Policy and Long-Term Care Insurance

Because long-term care is arguably the largest uninsured healthcare risk facing the United States (and many other countries) but political support for additional public coverage has been weak, policymakers have long been interested in finding ways to expand the private long-term care insurance market. The most established and well-known program designed to encourage purchase is the Partnership for Long-Term Care program, a state-based program developed in the late 1980s and first implemented in California, Connecticut, Indiana, and New York. The Deficit Reduction Act of 1995 enabled expansion of the program to other states. Under this program, purchasers of private long-term care insurance policies that cover a given number of years of care are afforded some degree of asset protection if and when they turn to Medicaid after their private policy benefits are exhausted. (Normally, Medicaid requires that individuals 'spend down' the majority of their assets before qualifying for benefits.) The specific rules about the degree of asset protection vary from state to state, but the two main models include a dollar-for-dollar matching of the amount of maximum benefit purchased with the amount of assets protected and a total asset protection model, which requires the purchase of a fairly comprehensive policy in return for total asset protection under Medicaid. Although there have been no rigorous evaluations of the program in the economics literature, estimates of take-up and potential Medicaid savings have been fairly small, as most people who purchased policies would have bought them in the absence of the program. Policymakers therefore appear supportive of the program but do not expect large expansions of private long-term care insurance coverage as a result.

A second tactic employed by US policymakers in pursuit of expanded private coverage is tax breaks, both state and federal, designed to lower the effective purchase price of private long-term care insurance policies. The Health Insurance Portability and Accountability Act of 1996 allowed long-term care insurance premiums to be deductable as a medical expense in calculating federal income taxes, similar to treatment of other medical expenses and insurance. Courtemanche and He (2009) studied the effect of the federal tax change on purchase behavior and found that the tax deduction led to significantly higher probability of purchase, on the order of a 25% increase among those eligible for the tax break. However, that effect translates to only a small increase in coverage across all seniors. Furthermore, they found that the loss in revenue to the government exceeded the potential savings to Medicaid in long-term care costs, leading to a net revenue loss. Similarly, Goda (2011) examined the impact of state tax incentives on private long-term care insurance coverage and the resulting effect on Medicaid expenditures, finding that the average state tax subsidy raised coverage by 28% and that the lost tax revenue exceeded the savings to Medicaid. In both cases, the net revenue loss was attributable largely to the fact that the part of the wealth and income distribution that responds to the tax incentives is generally not the part that relies on Medicaid.

Perhaps the most significant US public policy on this issue to date was the Community Living Assistance Services and Supports (CLASS) Act, passed as part of the Patient Protection and Affordable Care Act of 2010 but subsequently repealed when it was found to be financially unviable. CLASS was intended to reduce the uninsured risk of substantial long-term care costs by establishing an entirely voluntary, private-premium-funded, but publicly administered long-term care

insurance program. By statute, individuals who paid premiums for a minimum of 5 years and were working for at least 3 of those years would have been potentially eligible for benefits if they stayed in the program and reached an appropriate level of need, levels that would be similar to eligibility triggers used in private long-term care insurance. It was designed to be an 'opt-out' system such that employers could choose to participate or not, but if they chose to participate, employees would be automatically enrolled with the option to drop out if they chose. The benefits would be worth at least US$50 per day and would be available for a variety of long-term care services, not just nursing homes, but the benefit would be tied in some way to long-term service use (as opposed to a pure cash benefit). Most of the details of the design of the program were left to the 'discretion of the Secretary' (of Health and Human Services), but key restrictive attributes were written into the statute, including a requirement that the program be financially self-sustaining with no taxpayer subsidies for 75 years. Given the minimal requirements for eligibility and voluntary nature of the program, serious concerns about the potential for adverse selection made it impossible to design a premium structure that would meet the sustainability requirement. It was also unclear how the existence of a program like CLASS would affect the private long-term care insurance market as it stands today, but the rise and demise of CLASS underscores the need to better understand the private long-term care insurance market and the role that it can play as public policy toward long-term care financing evolves.

Conclusion

Theories and empirical evidence drawn from other types of health insurance may not apply to private long-term care insurance, as long-term care is distinct in several key ways. Family members, especially spouses and adult children, are hypothesized to play significant roles in decisions about long-term care insurance, yet the empirical evidence on the role of family is remarkably inconsistent and sparse. The market is small relative to other types of health insurance, with only 12% of the elderly population in the US holding policies. However, other than Medicaid crowd-out, the evidence on why people purchase or do not purchase policies is fairly weak, and policy efforts to expand the market have not been very successful. Given the existence of Medicaid as a safety net payer and the ability of the upper end of the wealth and income distribution to self-insure, it may be that the current size of the private long-term care insurance market is somewhat of a steady state. If that is the case, then the potential for market failures such as adverse selection and moral hazard – already of more concern in long-term care insurance markets than in other health insurance markets – becomes more of a threat to the stability of the market. Increases in adverse selection through advances in technology such as genetic testing, for example, could have serious implications for the existence of the market if it remains small.

Economists have identified and focused on these distinct features of long-term care insurance in a growing body of work. But despite the importance for public policy of economic theory and empirical evidence on long-term care insurance, significant gaps remain in the understanding of this market. As states and nations struggle with strategies to reduce the substantial individual and public risk of long-term care costs associated with aging populations, it will become increasingly important to fill these gaps.

See also: Access and Health Insurance. Aging: Health at Advanced Ages. Health Insurance and Health. Health Status in the Developing World, Determinants of. Healthcare Safety Net in the US. Long-Term Care. Mandatory Systems, Issues of. Moral Hazard. Performance of Private Health Insurers in the Commercial Market. Private Insurance System Concerns. Risk Selection and Risk Adjustment. Supplementary Private Insurance in National Systems and the USA

References

Brown, J. R. and Finkelstein, A. (2007). Why is the market for long-term care insurance so small? *Journal of Public Economics* **91**(10), 1967–1991.

Brown, J. R. and Finkelstein, A. (2008). The interaction of public and private insurance: Medicaid and the long-term care insurance market. *American Economic Review* **98**(3), 1083–1102.

Courtemanche, C. and He, D. F. (2009). Tax incentives and the decision to purchase long-term care insurance. *Journal of Public Economics* **93**(1–2), 296–310.

Cramer, A. T. and Jensen, G. A. (2006). Why don't people buy long-term-care insurance? *Journals of Gerontology, Series B: Psychological Sciences* **61**(4), S185–S193.

Cutler, D. M. (1993). Why doesn't the market fully insure long-term care? *NBER Working Paper Series #4301*. Available at: http://www.nber.org/papers/w4301 (accessed 28.08.13).

Finkelstein, A. and McGarry, K. (2006). Multiple dimensions of private information: Evidence from the long-term care insurance market. *American Economic Review* **96**(4), 938–958.

Finkelstein, A., McGarry, K. and Sufi, A. (2005). Dynamic inefficiencies in insurance markets: Evidence from long-term care insurance. *American Economic Review* **95**(2), 224–228.

Goda, G. S. (2011). The impact of state tax subsidies for private long-term care insurance on coverage and Medicaid expenditures. *Journal of Public Economics* **95**(7–8), 744–757.

Kemper, K., Komisar, H. L. and Alecxih, L. (2005). Long-term care over an uncertain future: What can current retirees expect? *Inquiry* **42**(4), 335–350.

Konetzka, R. T. and Luo, Y. (2011). Explaining lapse in long-term care insurance markets. *Health Economics* **20**(10), 1169–1183.

Lockwood, L. (2010). The importance of bequest motives: Evidence from long-term care insurance and the pattern of saving. *Working Paper*. Available at: http://www.aria.org/rts/proceedings/2011/Lockwood-BequestMotives.pdf (accessed 28.08.13).

Mellor, J. M. (2001). Long-term care and nursing home coverage: Are adult children substitutes for insurance policies? *Journal of Health Economics* **20**, 527–547.

MetLife Mature Market Institute (2008). *The MetLife market survey of nursing home & assisted living costs*. Westport, CT: MetLife Mature Market Institute.

Norton, E. C. (2000). Long-term care. In Cuyler, A. and Newhouse, J. (eds.) *Handbook of health economics*, vol. 1A, pp. 955–994. Amsterdam: Elsevier Science.

Oster, S., Shoulson, I., Quaid, K. and Dorsey, E. R. (2010). Genetic adverse selection: Evidence from long-term care insurance and Huntington disease. *Journal of Public Economics* **94**(11–12), 1041–1050.

Pauly, M. V. (1990). The rational nonpurchase of long-term care insurance. *Journal of Economic Perspectives* **6**(3), 3–21.

Sloan, F. A. and Norton, E. C. (1997). Adverse selection, bequests, crowding out, and private demand for insurance: Evidence from the long-term care insurance market. *Journal of Risk and Uncertainty* **15**, 201–219.

Taylor, Jr., D. H., Cook-Deegan, R. M., Hiraki, S., et al. (2010). Genetic testing for Alzhemer's and long-term care insurance. *Health Affairs (Millwood)* **29**(1), 102–108.

Zweifel, P. and Struewe, W. (1998). Long-term care insurance in a two-generation model. *Journal of Risk and Insurance* **65**(1), 13–32.

Further Reading

Brown, J. R., Coe N. B. and Finkelstein, A. (2007). Medicaid crowd-out of private long-term care insurance demand: evidence from the health and retirement survey. *NBER Working Paper Series #10989*. Available at: http://www.nber.org/papers/w12536 (accessed 30.08.13).

Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity

B Shankar, Leverhulme Centre for Integrative Research on Agriculture and Health, London, UK, and University of London, London, UK
M Mazzocchi, Università di Bologna, Bologna, Italy
WB Traill, University of Reading, Reading, UK

Glossary

Asymmetric information A situation in which the parties to a transaction have different amounts or kinds of information as when, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances.
Externality An externality is a consequence of an action by one individual or group for others. There may be external costs and benefits. Some are pecuniary, affecting only the value of other resources (as when a new innovation makes a previously valuable resource obsolete); some are technological, physically affecting other people (communicable disease is a classic example of this type of negative externality); some are utility effects that impinge on the subjective values of others (as when, for example, one person feels distress at the sickness of another, or relief at their recovery).
Market failure Markets in healthcare are notable for 'failing' on a number of grounds, including asymmetry of information between producers (medical professionals of all kinds) and consumers (patients actual and potential); distorted agency relationships, failure of patients to behave in accordance with the axioms of rational choice theory; incomplete markets, especially those for risk; monopoly; externalities and the presence of public goods.
Obese Individuals are classified as obese when their body mass index (weight in kilograms divided by squared height in meters) exceeds 30.
Oligopoly A departure from competitive markets, where the number of sellers is small, so that each adopts strategic behaviour by taking into account the behaviour of others.
Overweight Individuals are classified as overweight when their body mass index is between 25 and 30.
Productivity The amount of output or effect per unit of input in a period of time.
Utility Variously defined in the history of economics. Two dominant interpretations are hedonistic utility, which equates utility with pleasure, desire fulfilment, or satisfaction; and preference-based utility, which defines utility as a real-valued function that represents a person's preference ordering.

Introduction

Rapid increases in overweight and obesity prevalence rates over the last few decades, accompanied (and caused) by widespread dietary imbalances, are imposing huge burdens on health care systems and reducing the quality of life of populations around the world. These trends are not limited to the developed world alone, where there is talk of an 'obesity epidemic,' but also apply to several developing, transition, and middle-income countries. Between 1991 and 2008, the obesity prevalence rate in the UK grew from 14% to 25.4%, whereas in the US the percentage of obese individuals rose from 23.3% to 35.4%. Several countries undergoing economic transition have also witnessed a parallel 'nutrition transition,' characterized by significant increases in energy density, fat, sugar and salt content of local diets, and spiraling rates of overweight and obesity prevalence and associated disease costs. For example, 77% of Mexican men and 66% of women are now overweight, and Mexico is in the top tier of countries in obesity league tables.

In this article, the authors discuss the main macroeconomic causes and consequences of poor diets, obesity, and associated noncommunicable disease. The counterfactual implications of a movement toward better diets, and policy measures available to governments to improve diets are also discussed. Attention is restricted to the aggregate level – sector, economy, or population-wide issues – with microeconomic, individual/household level issues discussed only when relevant to the aggregate picture (government policies, applicable at the population level, are considered macro even if they work by affecting incentives at an individual level).

Causes

The debate about the attribution of obesity to economic factors has grown along with obesity rates in developed countries. A variety of factors – including genetic, psychological, and social drivers – have been put forward as potential causes. These are all relevant in explaining heterogeneity in weight within a cross section of people, but are consistent only to a limited extent with the speed of observed rise in the overall proportion of overweight and obese individuals over the last two decades. Because rapid changes are more likely to be rooted in socioeconomic factors, the role of economics in explaining the so-called obesity epidemic has gained prominence. Weight change is a function of the difference between calorie intake change and energy expenditure (physical activity) change (although some researchers also attribute a role to diet quality, for example, proportion of energy sourced from fat). Estimates of average daily per capita calorie intakes over the period 1991–2007 show a 7.8% increase

 doi:10.1016/B978-0-12-375678-7.00611-8

for the UK, a 6.8% increase for the US (and a 9.9% increase for developing countries, although in their case a proportion of this is a welcome improvement to the calorie intakes of the undernourished). The US has also experienced a substantial rise in fat intake (+14.1%). Economic drivers are seen as fundamental to these changes.

Technological Change and Commodity Prices

The argument that has gained most consensus in explaining the growth in obesity rates is related to the impact of technological change. Technical progress has rapidly increased agricultural productivity and lowered the cost of food. Furthermore, this trend has been uneven across foods, as the relative price of industrial and processed foods (and raw inputs like sugar) has declined at a faster pace compared to raw foods like fruits and vegetables. Figure 1 shows how productivity in the cereal sector has sharply risen over the last two decades, even in countries like the UK (+15%) and the US (+47%), where yields were already very high and much higher than the world average (+29%).

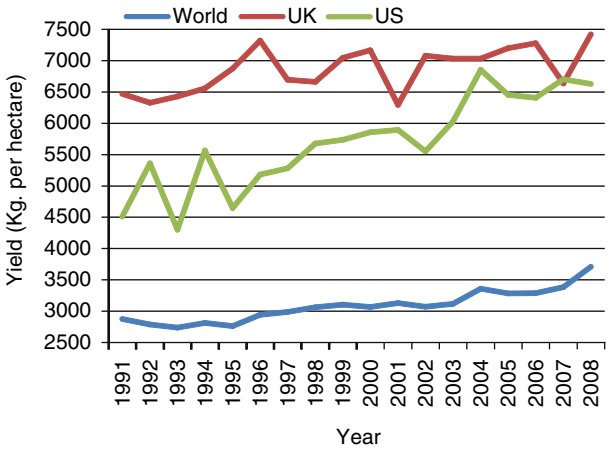


Figure 1 Cereal yields. Based on data from World Bank, World Development Indicators (2011).

The sharp decline in real commodity prices shown in Figure 2, together with the relatively small but regular increase in incomes, has contributed to the rise in calorie intakes observed in developed countries. In developing countries – especially in transition countries where income growth has been much more substantial – the effect on calorie intakes is even stronger, which explains why many of these countries are now experiencing rapidly rising obesity rates while still being affected by food insecurity. Technological change also matters to the other term in the weight change equation, calorie expenditure, and this effect is possibly even more influential than commodity price decline or income growth. There is strong evidence that jobs have become much more sedentary, and that physical activity has been transferred from paid working time to costly leisure time.

In summary, technological change has made calorie consumption progressively cheaper, whereas raising the costs (including time costs) of calorie expenditure.

Food Availability and Globalization

On the supply-side, particularly in developed countries, the increased availability of 'junk food,' defined as calorie-dense foods high in fats, sugar and salt has been also blamed (in developing and transition countries, increased livestock product consumption has similarly been blamed). From an economist's perspective, unless one accepts the asymmetric information assumption or some sort of oligopoly due to market segmentation, an increase in production of junk foods can be explained either by higher profitability for the industry or by the increased consumer demand. The former explanation can be traced back to the technological change hypothesis, as processing of commodities into energy-dense packaged foods has become cheaper over time. Growing demand (and consumption) would reflect changing preferences toward these foods compared to healthier dietary options, but no data exists to test this hypothesis over time.

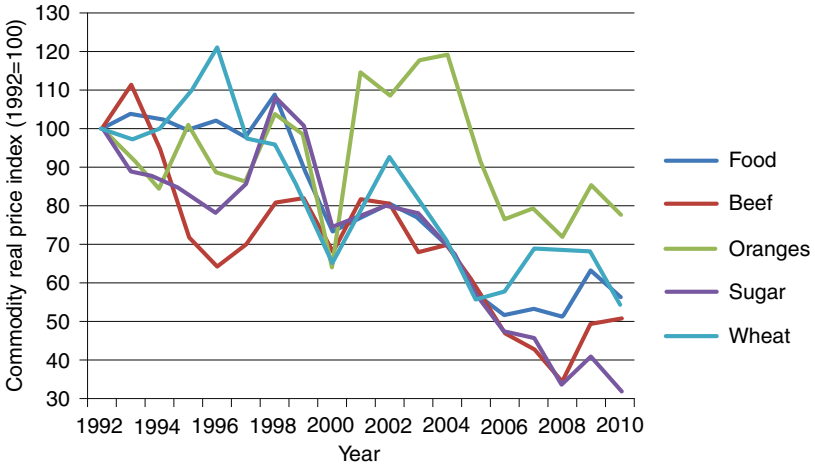


Figure 2 Real food commodity prices (relative to the price of all primary commodities). Based on data from International Monetary Fund, World Commodity Prices (2011), our processing.

Increased globalization and trade openness have played an important part, particularly in many developing countries. This is partly considered a cultural (demand-side) effect, sometimes called 'coca-colonization,' wherein unhealthy dietary patterns first established in developed countries are emulated by developing country consumers, with increased globalization and trade openness facilitating availability. On the supply-side, returns to scale afforded by new developing country markets can be exploited by large food manufacturers and multinationals, enabling even cheaper production of processed food already benefiting from lowered cost of production as a result of technological change.

Other Factors

A strong association between obesity rates and income disparities, which has been observed based on geographical comparison, may well hold across time considering that inequalities have been increasing in several countries over the years. Increase in female labor participation has been also proposed as a potential explanation for the decline in dietary quality and the consequent weight increase, especially for the younger generations. However, the link is not as clear as for technological change. Although there are studies showing that such an effect exists, it only explains a small portion of the observed growth in weight. Rapid urbanization is a further factor that has been associated with increasing obesity rates in developed countries mainly because life is thought to be more sedentary in cities, although it is difficult to define the causal direction of this relationship. Other economics-related explanations lie in the dramatic progress in medical treatment for obesity-related conditions, with a consequent decline in perceived risks which may work as a rational disincentive to conducting a healthy lifestyle.

Insufficient or biased (asymmetric) information (e.g., through advertising), often invoked as a driver for unhealthy behaviors at the individual level, is an unlikely determinant at the aggregate level, unless one assumes that the quality of nutrition and health information has worsened over the last two decades, despite most obesity policies having been targeted at public communication. The same argument holds for the role of education, an important explanatory factor for micro-level heterogeneity, but not particularly relevant (or even beneficial) when looking at the time series of obesity rates.

More recently, the focus of economists has turned to individual behavioral factors – especially behavioral failures – such as inconsistent time preferences, addiction, and lack of self-control. As in the case of genetics or other biological factors, it is quite difficult to bring conclusive evidence on the role played by these individual-level factors. To do so, one has to once more accept that behavior at the population level (or its distribution) has rapidly changed over time, for which there is insufficient evidence, given available data.

Effects

Unhealthy diets in combination with lower physical activity levels and obesity have been linked to a range of noncommunicable diseases (NCDs), including several types of cancer, coronary heart disease, stroke, type II diabetes, osteoporosis, and osteoarthritis. There are a number of pathways from diets and physical activity to these diseases. Primarily through calorie balance (although there is some evidence that diet quality matters too), there is an impact on overweight and obesity, which are directly linked to many of these diseases. Overweight and obesity may also operate through intermediary conditions, such as hypertension and dyslipidemia to raise the risk of contracting some of these diseases. In addition, diets and physical activity may directly (rather than operating through an effect on risk of obesity) affect NCD risk, or through intermediary conditions as noted above. These effects impose a range of costs on the macroeconomy, classified as direct (medical) and indirect (productivity), as described below. Available estimates are largely for developed countries (see, e.g., http://www.youtube.com/watchv=mfnwZrLKfoo), and there is a significant paucity of developing country cost estimates.

Direct Costs

The bulk of cost estimates relating to unhealthy diets and obesity relates to direct costs arising from increased medical expenditure on diagnosis, treatment, and management. A range of methods have been used in estimating these, ranging from cohort studies, where medical costs arising among groups of subjects varying by body mass index (BMI) ranges are examined over several years, to regression models, to studies based on dynamic simulation models of the relationship between BMI and NCD risks. These studies frequently extrapolate from study samples to the national population. Costs accruing to the national economy have been found to be substantial, as can be seen from **Table 1**, although it is worth noting the comparability across studies is complicated by

Table 1 A selection of estimates of obesity costs

Country	*Direct costs*	*Indirect costs*	*Notes*
UK	$3 billion	$10.5 billion	2001 costs of elevated BMI
China	$5.8 billion	$43.5 billion	2000 costs. Includes separate diet, activity, and obesity pathways
USA*	$147 billion		2008 estimate

*Indirect costs for US not included here because available estimates are dated and/or partial in coverage.

Source: Reproduced from McPherson, K., Marsh, T. and Brown, M. (2007). *Tackling obesities: Future choices: Modeling future trends in obesity and the impact on health*. London: Government Office for Science; Popkin, B. M., Kim, S., Rusev, E. R., Du, S. and Zizza, C. (2006). Measuring the full economic costs of diet, physical activity and obesity-related chronic diseases. *Obesity Reviews* **7**: 271–293, and Finkelstein, E., Trogdon, J., Cohen, J. and Dietz, W. (2009). Annual medical spending attributable to obesity: Payer and service-specific estimates. *Health Affairs* **28**: w822–w831.

differing protocols, methods and pathways and components taken into account (e.g., consideration of costs arising from obesity alone vs. costs arising from diet quality as well as obesity). A key issue in cost estimation relates to 'lifetime costs' – whether medical cost savings due to early mortality caused by obesity offsets the increased medical costs accrued during the lifetime of overweight and obese individuals. The limited research available on this issue is inconclusive, and this remains an area for future research.

Indirect Costs

Indirect costs of obesity estimated in the literature encompass a range of nonmedical costs relating to productivity loss. These include absenteeism, disability, premature mortality, and presenteeism. Absenteeism and disability costs arise from time taken out of work due to obesity-related conditions. Premature mortality costs arise when workers die before retirement age due to obesity-related disorders. Presenteeism captures lowered efficiency at work arising from obesity-related disorders. There is debate about the extent to which lost time at work equates to lost productivity, because harder work from those present at work may compensate for time–input loss arising from obesity. As in the case of direct costs, available studies differ in terms of what they cover under indirect costs, and there are numerous measurement problems, prominent among these being distinguishing correlation from causation in measuring the effect of obesity on indirect costs. Table 1 shows that indirect costs can be very substantial, and can exceed direct costs by a significant margin in some countries.

Available estimates of the total burden of overweight and obesity, including direct as well as indirect costs, range from 0.2% to 0.6% of GDP in the developed west. For China, the estimate is as high as 4% of GDP.

Contemplating Hypothetical Scenarios: What Would the Implications of Improved Diets Be?

The flip side of the earlier discussion on how changes in food consumption patterns have contributed to unhealthy dietary outcomes and NCDs, are the questions: (1) what would the larger sectoral/economy-wide implications of improved diets be and (2) what policies are needed to get there?

A sparse literature exists that estimates (simulates) the implications of moving toward recommended dietary norms, such as the World Health Organization (WHO) guidelines, at the population level. These show that the biggest negative consumption impacts would be on the animal products (meat, animal fats, and dairy products), vegetable oil, and feed cereal sectors. In Organization for Economic Cooperation and Development (OECD) countries, for example, consumption of animal products would shrink by between 15% and 30%, if WHO norms are to be met. However, the largest global effects would be generated by lowering meat consumption in rapidly growing economies such as China rather than in OECD countries.

Health benefits from such adherence to norms are likely to be substantially higher in developed and transitioning countries, where overnutrition is a more pressing concern, than in developing countries. However, patterns of international trade in agricultural products and general equilibrium effects imply that the effects of consumption changes in any large country or sets of countries are likely to be felt in other parts of the world, particularly developing countries. For example, a significant reduction in meat consumption in major markets, such as the EU, US, Canada, and Japan would have a substantial effect, notably a sharp increase in short-run unemployment in a large meat-exporting country such as Brazil. There is little evidence available to indicate that a global movement toward healthier diets can do much to enhance food and nutrition security in developing areas. The key implications of such movements are for meat consumption and for cereals used as feed. Although meat-exporting developing countries may suffer from reduced export opportunities, wheat and rice, the main staples used in developing countries have been shown to be little affected by such changes. However, it must be noted that the potential supply response from developing countries of a movement toward increased fruit and vegetable consumption is an area that has not been investigated thoroughly.

Policy for Better Diets

If overweight and obesity are accepted to be the outcome of individual utility maximizing decisions, then the economic rationale for public policy intervention has to be market failure. Foremost among these is externalities; people who choose to be overweight do not bear the full social cost of their actions, they are partially borne by others to the extent that health care is subsidized and employment law guarantees wages are paid when obesity-related ill health forces time off work. A second market failure occurs if information is imperfect, perfect information being a precondition for the informed choice underpinning utility maximization. Finally, food markets may display imperfect competition, specifically resulting in competition centered around advertising; food is an advertising-intensive industry, particularly fast food, confectionary, savory snacks, and soft drinks, largely viewed as principal culprits of the growth in obesity, particularly in the developed world.

In reality, governments also justify intervention for noneconomic reasons, notable among these being the correction of health inequalities (the socially deprived show a higher prevalence of obesity). Acting to change social norms has been used as a further justification for action; essentially this means changing people's utility functions so that they choose to weigh less, comparison being made with the successes in changing attitudes to drunk-driving, smoking in public places, and wearing seat belts when driving. Children are often seen as a special case for whom more overt intervention to control is justified. More recently, behavioral economists have focussed attention on widespread systematic divergence from the rational behavior assumed by neoclassical economic models, arguing that such 'behavioral failures' have been exploited by the food industry to encourage higher consumption

Table 2 Classification of policy actions

Information measures
Public information campaigns
Advertising controls
Nutrition education
Nutritional labeling
Nutritional information on menus
Measures to change the market environment
Fiscal measures
Tax/subsidies on foods to the population at large
Subsidies to disadvantaged consumers
Regulate meals
School meals (including vending machine bans and provision of free fruits and vegetables)
Workplace canteen meals
Nutrition-related standards
Government action to encourage private sector action
Availability measures for disadvantaged consumers

of processed foods; benevolent paternalism, it is argued can similarly exploit such behavioral failures to nudge people toward choosing healthier lifestyles they themselves would prefer (helping them to maximize their individual utilities).

The policy responses can be usefully grouped into two main categories, those actions centered around information and those which more directly intervene in markets. The actions which have been taken are shown in Table 2.

Of the information actions, public information campaigns exploit media communication and other social marketing tools to improve individual and social knowledge about health issues connected to food habits, and may be directed at any kind of target population. It is by far the most common healthy eating policy, together with education interventions. The aims may be simply to better inform people (e.g., about the health risks of obesity), or to change social norms. Advertising controls (bans) could in principle be used to limit advertising to adults and children if it was believed that all ages were encouraged to overeat by commercial advertising, though in practice, the measures have only been applied to children, presumably because it would be considered overly paternalistic to take such measures for adults. Nutrition education actions could likewise be used for adults or children, but have in practice only been used for children in schools. Nutrition labeling is essential to informed choices because the nutritional composition, notably number of calories, in foods, particularly processed foods, cannot be easily assessed, even by food scientists. Some form of labeling of the nutritional content of processed foods is compulsory in many developed countries and common even when not compulsory; the debate now is over the most effective form of communication using simplified messages, such as traffic lights to represent high, medium, or low levels of the major nutrients. There is a move toward extending labeling to food eaten outside the home in restaurant chains selling standardized products such as hamburgers.

Market intervention measures are less common. Fiscal measures have been proposed and widely assessed (simulated) by economists. On the positive side, taxes on unhealthy foods could be used to make people pay the full social cost of the food they eat, including the health care and economic productivity loss costs. Subsidies for healthy foods such as fruits and vegetables could be similarly justified as aligning social and private costs. The counterargument is that taxes would be regressive (the poor spend a higher share of their incomes on food), though there is some debate as to whether a fiscally neutral system where the subsidy cost exactly matches the tax revenue would suffer in this way. In any case, the measure would be highly unpopular with the food industry, and governments have not gone down this route, though small taxes, especially on soda, are widespread in the United States. One fiscal measure that has been employed, albeit to a limited extent, is subsidies, in the form of vouchers, to low-income households for the purchase of specific healthy foods. This is a promising area as it also addresses the issue of health inequalities, but may be deemed too expensive to apply in anything other than a very limited manner.

The other measures are all designed to influence the availability of foods, or rather nutrients. These tend to be targeted at diet quality more often than obesity per se, particularly measures to encourage food reformulation to reduce levels of salt, saturated fat, and trans fat in processed food; and measures to promote convenience stores in low-income areas to carry fruits and vegetables (the premise being that people in these areas without cars cannot access healthy food). The school food environment is commonly regulated to control the availability of junk foods (in canteens or vending machines) and the menus of meals (less chips, sausages, chicken nuggets, and hamburgers; and more salad, fruits, and vegetables). Menu control in public sector workplaces has been considered, but not widely applied.

See also: Macroeconomy and Health

Further Reading

Hawkes, C. (2006). Uneven dietary development: Linking the policies and processes of globalization with the nutrition transition, obesity and diet-related chronic diseases. *Globalization and Health* **2**, 4.

Lakdawalla, D., Philipson, T. and Bhattacharya, J. (2005). Welfare-enhancing technological change and the growth of obesity. *American Economic Review* **95**, 253–257.

Lock, K., Smith, R. D., Dangour, A. D., *et al.* (2010). Health, agricultural, and economic effects of adoption of healthy diet recommendations. *Lancet.* doi:10.1016/S0140-6736(10)61352-9.

Mazzocchi, M., Traill, W. B. and Shogren, J. (2009). *Fat economics: Nutrition, health and economic policy.* Oxford: Oxford University Press.

Msangi, S. and Rosegrant, M. (2011). *Feeding the future's changing diets: Implications for agriculture, markets, nutrition and policy.* IFPRI 2020 conference paper. Washington: International Food Policy Research Institute.

Popkin, B. M. (2003). The nutrition transition in the developing world. *Development Policy Review* **21**, 581–597.

Popkin, B. M., Kim, S., Rusev, E. R., Du, S. and Zizza, C. (2006). Measuring the full economic costs of diet, physical activity and obesity-related chronic diseases. *Obesity Reviews* **7**, 271–293.

Rosin, O. (2008). The economic causes of obesity: A survey. *Journal of Economic Surveys* **22**, 617–647.

Srinivasan, C. S., Irz, X. T. and Shankar, B. (2005). An assessment of the potential consumption impacts of WHO dietary norms in OECD countries. *Food Policy* **31**, 53–77.

Trogdon, J. G., Finkelstein, E. A., Hylands, T., Dellea, P. S. and Kamal-Bahl, S. J. (2008). Indirect costs of obesity: A review of the current literature. *Obesity Reviews* **9**(5), 489–500.

Macroeconomic Dynamics of Health: Lags and Variability in Mortality, Employment, and Spending

TE Getzen, International Health Economics Association, Philadelphia, PA, USA

Prologue: Lags Tend to Obscure What is Known

The best known 'facts' about the macroeconomics of health are that rich nations are healthier and spend more on medical care than poor nations, but that additional wealth or spending may not add much to life expectancy after some threshold level has been exceeded (Figure 1(a)–(c)). A fact that receives insufficient attention is that any major macroeconomic change takes time, often quite a long time. An often repeated but generally incorrect 'fact' is that population aging and health risks (obesity and cancer) are major drivers of aggregate spending growth.

Macroeconomists focus on large-scale issues at the national or global level – growth, distribution, business cycles, money, and finances – rather than the micro individual rational choice decisions examined by most health economists. Macroeconomists tend to use time series methods and address dynamics rather than the cross-sectional methods and comparative statics of micro studies. Analyzing when and how change occurs forces more explicit consideration of lags, heterogeneity, and variance – and of the differences between micro and macro processes that might superficially appear to be the same.

Some notable disparities addressed in this article are the contrast between the quick, anticipatory movements of financial markets and the slow inertial flow of complex health care systems (smoothing that renders regular business cycles almost invisible); discrepancies in the determinants of spending between the individual micro level (illness) and the national macro level (per capita gross domestic product (GDP) – with a lag); and divergences in sustainable rates of growth.

Mortality and GDP

During the past 200 years, many parts of the world experienced unprecedented growth in material well being and human health. In the UK, real income per capita rose 10-fold while life expectancy doubled. Demographic transition and the industrial revolution brought similar improvement in the US, France, Germany, Sweden, Japan, and most developed nations. The massive effect of modern economic development on human conditions is well known and beyond dispute. The timing and uneven distribution of such gains is less well recognized. What has become increasingly evident in recent research is that the relationship between 'GDP' and 'Health,' although quite strong and clearly causal, is far from simple.

Long Lags

Any major social change takes time and rests on many preconditions, making a precise dating of a 'starting point' at best imprecise, and possibly misleading. That said, a reasonable consensus among the economic historians and macroeconomists who study growth is that the industrial revolution began around 1775 ($\pm$ 75 years) and was well established by 1850, although wider diffusion and follow-on benefits continued through much of the twentieth century. Therein lies the rub. Although the surge of innovation and economic development was manifestly widespread in nineteenth century Dickensian England, the industrial revolution in 1850 – and for a long time thereafter – is associated with widespread misery and substantial declines in life expectancy. The data presented by Angus Maddison are consistent with the following rather loose and lengthy causal chain: A burst of productivity-enhancing innovations (steam engine and factory work) starting around 1780 allowed rapid growth in population and trade, which eventually (20–50 years later) led to rising average incomes and material well-being of individuals, which in turn (after another 20–50 years) led to a rise in human life expectancy. Some details of timing, paths, and dynamics of this process are discussed in section Growth, Business Cycles, and the Long Run below.

Business Cycles and Employment

Figure 2 compares 'total' and 'health' employment in the US 1990–2010 and reveals two major macro conclusions: The health sector is growing much faster than the rest of the economy (rising share), and that growth is much steadier (lower variance). The jagged seasonal variation very evident in total employment is almost nonexistent in health care. The significant deviations from trend due to recessions in 1990–91, 2001, and 2007–09 readily discernable in total employment are also missing. Instead, health employment shows an almost steady upward incline throughout this 20-year period and for earlier decades as well.

The health sector's lack of response to recession is evident in Figure 2(b). The 'great recession' officially dated as beginning in the fourth quarter of 2007 appears here as a slowdown in rate of job growth starting after a peak (2.1%) in March 2006, which then went below the long-run sustainable rate of increase (0.9%) in November 2007 and turned negative in May 2008, finally reaching a trough in August 2009 when jobs were disappearing at a 5% annual rate. Only after June 2010 did job growth turn positive, and it will still be a number of years before overall US employment again reaches the previous level (139 million) and even longer (perhaps 5–7 years) to compensate for the intervening population growth. In contrast, growth in health employment continued to increase throughout 2007 and decelerated moderately after that. The great recession, to the extent that is visible at all in health care, shows up as a slight dampening in a continuing high rate of growth 2 years after the most massive economic downturn since the depression.

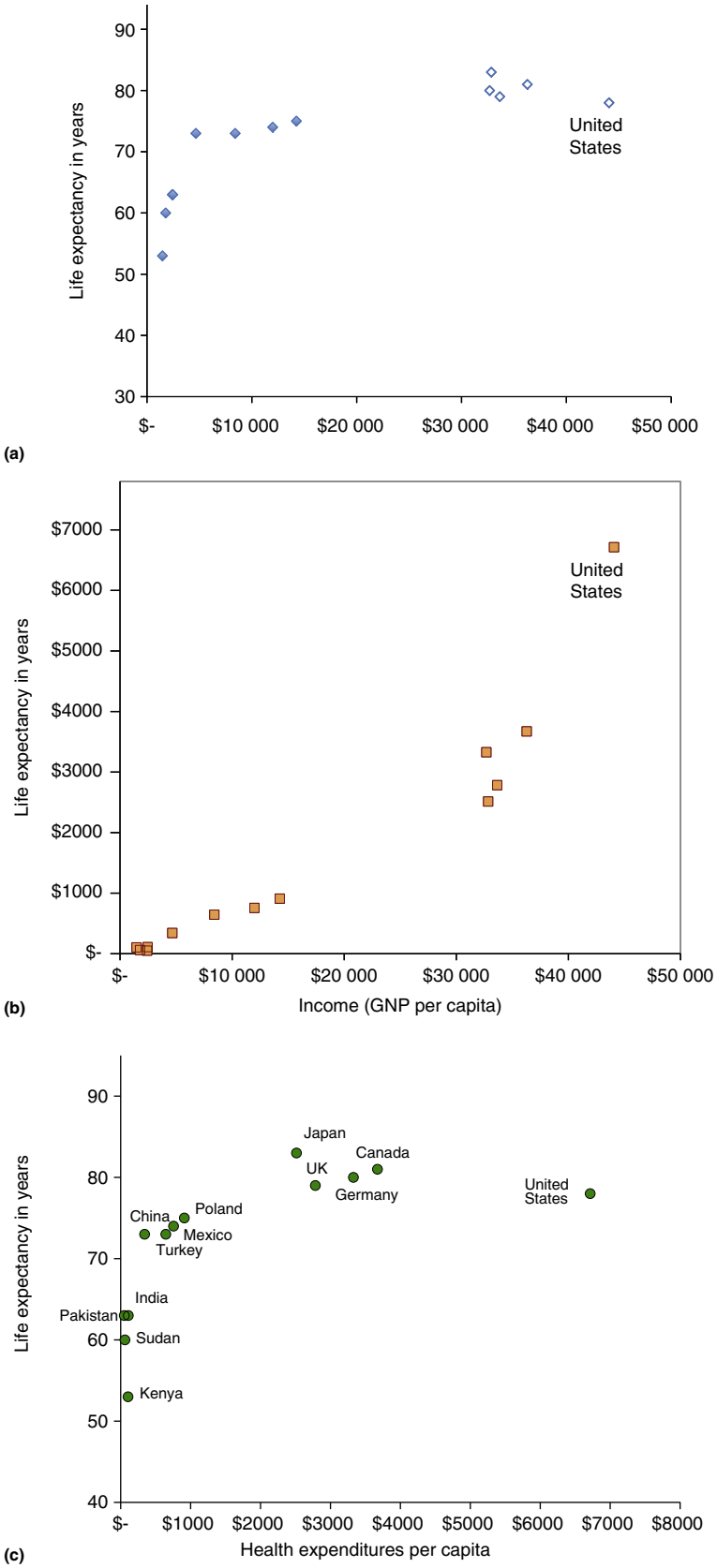


Figure 1 (a) Life expectancy and GDP per capita. (b) Per capita health expenditures and income. (c) Health expenditures and life expectancy across countries.

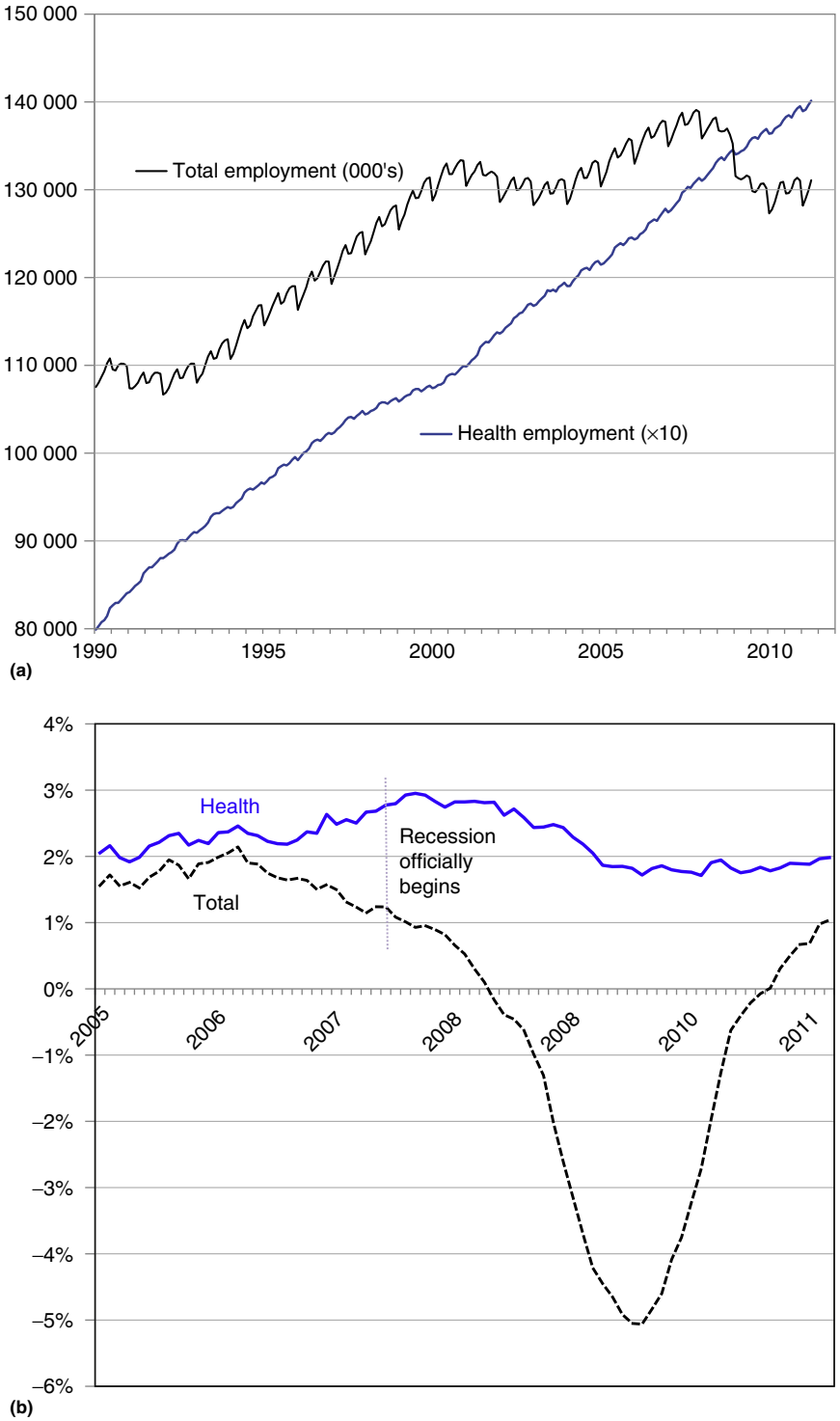


Figure 2 (a) US employment: 1990–2011. (b) Annual % change US in employment: 2005–11.

Unemployment and Mortality

Employment (and its obverse, unemployment) is a main indicator of economic growth. Hence, it seems reasonable that more employment (unemployment) should be associated with higher survival (mortality). Indeed economic historians traced the path of medieval economic fluctuations correlating the price of grain with mortality rates. Twentieth century policy makers often pointed to the adverse effects of unemployment on population health as a justification for countercyclical monetary and fiscal interventions. The research and legislative testimony of M. Harvey Brenner quantifying the

expected number of lives lost for each additional percent of unemployment became so well known that the association of unemployment with mortality was widely referred to as the 'Brenner Hypothesis.' The strong long-run and cross-sectional connection between GDP and mortality made it seem like 'common sense' that a similar short-run relationship should hold. However, Jose Granados, Hugh Gravelle, Audrey Laporte, Jes Sogaard, Adam Wagstaff, and others attempting to empirically verify the Brenner hypothesis reported great difficulty in doing so. In a seminal paper in 2000, Christopher Ruhm reported compelling evidence that recessions were in fact associated with less, rather than greater, mortality – and was able to explain why. Briefly and incompletely put, Ruhm and others have shown that unemployment and the concomitant reduction in general economic activity is associated with changes in behavior and consumption (less driving, more exercise, etc.) that reduce contemporaneous mortality without affecting long-run mortality very much. This is especially true for deaths due to accidents, cardiovascular disease, births, and some other medical conditions, whereas the converse holds for suicide and some other causes of death where acute stress may play a greater role. The conclusion that unemployment lowers mortality rates, although considered counterintuitive 20 years ago, has been so frequently confirmed empirically that most informed researchers would now consider it conventional – even though much of the public still thinks unemployment causes mortality rather than the reverse.

Some of the public confusion arises because these macro results apply to aggregate population mortality rates rather than the typical individual results that people 'see for themselves.' A negative macro correlation between unemployment and mortality does not imply that unemployment is healthy for the individual who loses a job. Indeed, there is compelling research showing that unemployment is highly damaging to the individual who is laid off. Daniel Sullivan and Til von Wachter report that involuntarily unemployed workers suffer a 10–15% increase in annual mortality rates that persists for at least 20 years, reducing average life expectancy by 1–1.5 years. Jason Lindo reports that parental job loss substantially reduces birthweight and child health, while Gerard Van den Berg, Maarten Lindeboom, and France Portrait show that infants born during economic crises in the nineteenth century had reduced life expectancies. These results make it clear that the impact of job loss on individual health (micro effect) is quite different from the macro effect on population rates.

Spending

Expenditures on health care have increased rapidly in all developed (Organization for Economic Co-operation and Development (OECD)) countries over the past five decades, with total spending rising more than 1000% in most countries due to inflation, demography, technology, income, and other factors. However, the relative contribution of each factor is often uncertain, variable over time and across countries, as well as being subject to inertia and lags of varying lengths.

Inflation and 'Real' Expenditures

Differences in the nominal value of money over time and across countries cause large yet presumably unimportant differences in measured spending. If medical transactions were simple spot exchanges and price indexes were perfect, adjustment using deflators and exchange rates would not be an econometric problem. Instead, medical transactions are usually complex, involving group contracts and institutional interactions extending over years or decades. In such a context, inflation and purchasing power parity discrepancies will often distort measures of 'real' health expenditures.

To sidestep real versus nominal issues quantifying resource use within a country, region, township, or household by share of GDP (or of consumption, income, employment, etc.) may sometimes be preferable. However, the inertial response of health care systems to macroeconomic forces means that short-run shifts in shares are more apt to come from delays and measurement errors than substantial changes in real resource use. This is shown below with data from Canada during a spike of inflation in 1974. The measured health share of GDP fell from 0.073 to 0.069, whereas the share of employment in health increased. It is most likely that the real share of economic activity devoted to health was rising rather than falling in 1974.

Year	*1972*	*1973*	*1974*	*1975*
Inflation	5.6%	8.9%	14.4%	9.8%
Nurses	1 52 005	1 59 274	1 68 530	1 77 182
Health share of GDP	7.3	7.0	6.9	7.4
Fraction of employment	0.0180	0.0180	0.0183	0.0189

One way such systematic errors are generated is by delaying wage increases. In a study by Getzen and Kendix, wages of health care workers were estimated to have a secular increase of 0.6% above that of other workers and respond essentially 1:1 to inflation – but with a lag. When inflation goes up (or down), less than half of that change in the rate of inflation is reflected in health care wages in the current year, and even after 2 years, about one-fourth of any shift is still waiting to trickle into the health sector.

$$\text{Wages} = 0.6\% + 1.02\ \text{CPI} - 0.61\Delta\text{CPI}_{0-1} - 0.11\Delta\text{CPI}_{1-2}$$

If it takes 18 or 24 months for changes in the general level of prices to be reflected in the wages of health care workers, significantly longer than for most employees, then measured labor will appear to be significantly below(above) real employment whenever inflation is rising(falling), even though the long-run effect of general price inflation is neutral. Similar distortions arise when purchasing power parities (PPPs) deviate widely from exchange rates. Internationally traded items, such as pharmaceuticals, are priced in international currency units, whereas wages and physician services reflect domestic (PPP) equivalents.

Income Effects

Measurement and estimation of income effects are even more affected by lags and inertial response. In **Figure 3(a)**, the

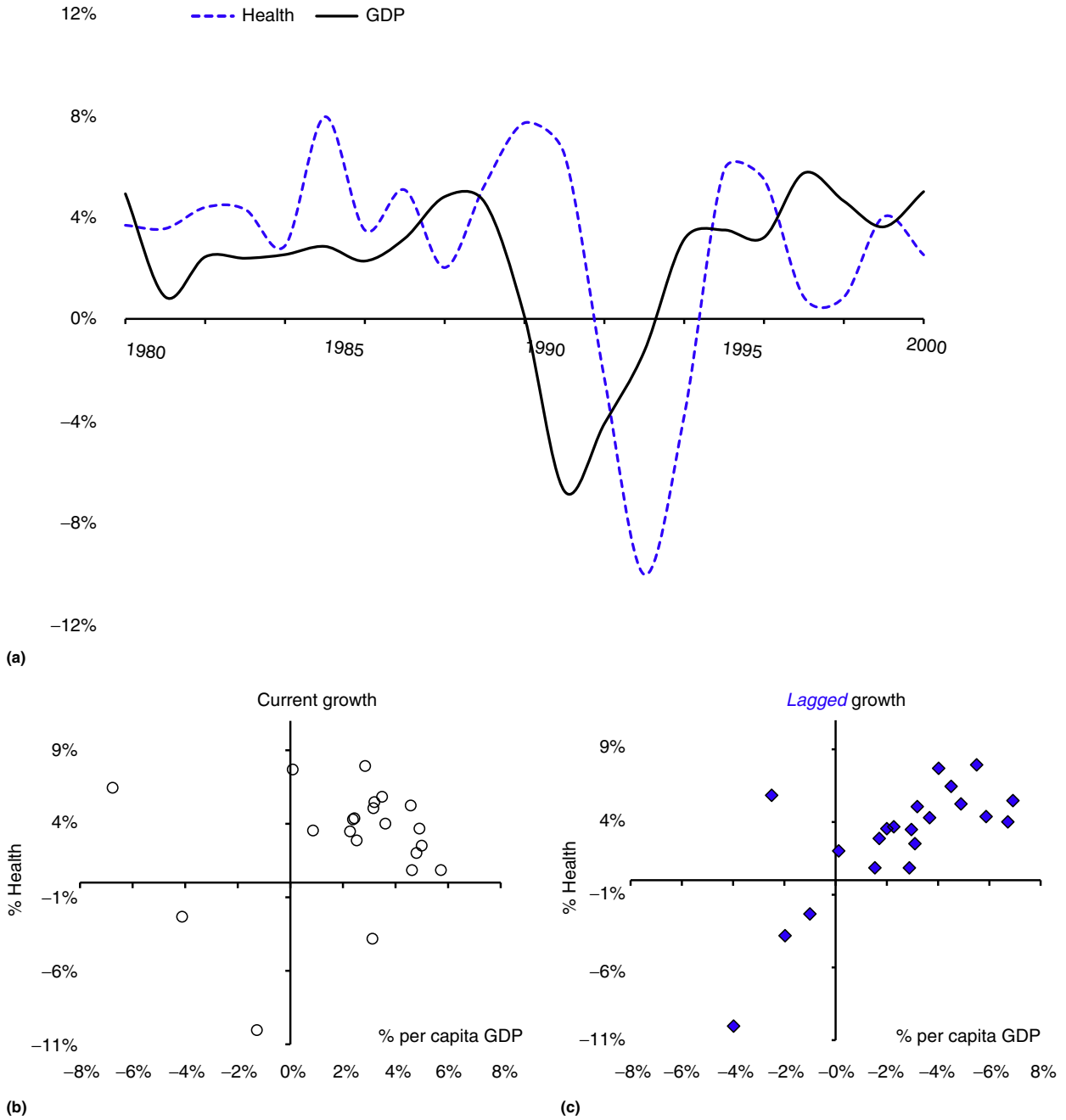


Figure 3 (a) Finland 1980–2000 annual % growth in Income and Health. (b) Current growth. (c) Lagged growth.

1990–93 recession in Finland and the subsequent decline in national health expenditures occurring after a lag of 2 years is clearly visible. Yet when the same data are presented as a scattergram in Figure 3(b), the correlation between GDP growth and health expenditures growth is almost entirely obscured. Only once allowances are made for delayed response spread out over several years does the correlation again become clear, as in Figure 3(c), which plots annual expenditure increases against a lagging 3-year moving average of GDP growth.

Health care spending depends on permanent income, which changes slowly over time. Even after a decision to spend more (or less) has been made, the rigidity of budgets and licensed professions delays implementation. With slowly evolving expectations regarding permanent income and complex institutional inertia, the impact of current changes in GDP are barely apparent in contemporaneous spending. Estimation across a panel of OECD countries from 1961 to 2008 indicates an average lag of 2 or more years before changes in per capita GDP affect health care spending.

$$\%\mathrm{HE} = 0.035 + 0.13\mathrm{GDP}_0 + 0.32\mathrm{GDP}_{-1} + 0.33\mathrm{GDP}_{-2} + 0.15\mathrm{GDP}_{-3} + 0.10\mathrm{GDP}_{-4} + 0.13\mathrm{GDP}_{-5} - 0.26\Delta\mathrm{CPI}_{0-1} - 0.16\Delta\mathrm{CPI}_{1-2}$$

Table 1 Growth in real health expenditures (%) as function of lagged GDP growth: US 1960–2009

	Constant	*rGDP-0*	*rGDP-1*	*rGDP-2*	*rGDP-3*	*rGDP-4*	*rGDP-5*	*Deflator-0*	*Deflator-1*	*Time*	*R2*
US – Total NHE	0.046	0.17	0.07	0.04	0.19	0.29	0.23	−0.28	−0.12	−0.0006	0.702
Hospital	0.068	−0.17	0.06	0.06	0.14	0.39	0.24	−0.15	0.25	−0.0011	0.705
Physician	0.044	0.03	0.36	0.14	0.13	0.37	0.03	−0.52	−0.69	−0.0006	0.312
Dental	0.009	0.36	0.16	0.22	0.18	0.22	0.32	−0.41	0.07	−0.0002	0.311
Pharmaceutical	−0.071	0.73	0.34	0.70	0.71	0.09	0.15	−1.04	−0.95	0.0016	0.457
LTC	0.058	0.22	0.65	0.45	0.18	0.54	0.35	−0.67	0.03	−0.0013	0.662
Insurance administration	0.064	0.17	0.44	0.26	0.34	0.57	0.49	−0.57	0.12	−0.0013	0.470
Out of pocket	−0.006	0.43	0.26	0.13	0.11	0.26	0.00	−0.57	−0.53	−0.0001	0.245

Abbreviations: LTC, long term care; NHE, national health expenditures; R2, r-squared.
Source: Author's regressions from data at www.cms.gov/nationalhealthexpenddata (accessed 22.05.11).

What is not so apparent in this single equation estimate on panel data encompassing 17 countries and 46 years is that the lags between measured current income and changes in spending vary substantially from country to country, by the particular type of health spending (research, physician, hospital, and dental) and even from one time period to the next. Table 1 provides coefficients estimated for several categories of health expenditures within the US for 1960–2009. The average lag for all categories of spending combined is a little over 3 years, but varies from less than 2 years for personal out-of-pocket spending and drugs to more than 4 years for hospitals. Presumably, more detailed accounting would reveal an even greater range, perhaps just a month or two for bandages and over-the-counter medicines but close to a decade or more for construction of new buildings.

500 Observations can robustly establish that lags occur and that they vary, but is hardly able to specify the range and shape of those variations or to identify the many institutional features that cause responses to be delayed. Kenneth Arrow's classic 1963 paper focuses on uncertainty with regard to incidence (risk) and quality (effectiveness) as the cause of special characteristics in the health care market. The first risk, is dealt with primarily through pooled financing. Third party insurance, whether government or private, builds a structural lag into the link between income and expenditure. Premiums are set well in advance and based on expectations – that is, on a form of permanent rather than temporary income.

Although demand side buffering causes some lags, the more significant institutional rigidities that Arrow identifies occur on the supply side – licensure, cost shifting, nonprofit community organization, and other barriers. Adjustment of physician supply is enmeshed in traditional educational institutions that are resistant to change, so much so that any equilibrium must be considered heavily punctuated if not ossified. From 1980 to 2005, US population grew 23% and real health expenditures per capita grew 187%, but just one new medical school was built and the number of US medical graduates rose by only 2% (from 15 632 to 15 962). The shift to produce more graduates carried out in the early 1960s reverberated for more than 30 years (the average length of professional practice), but the grudging and belated accommodation of growth through the professional supply chain indicates how inertial the medical care system can be.

Expectations prevalent during creation tend to get built in, embedded in the processes and organizations by which a medical financing system operates. The Medicare and Medicaid programs in the US were conceived during a period of endless growth and bright technological promise and thus were designed to increase the wages of health workers, to subsidize the construction of hospitals, and to support experimental treatments through generous funding. Enactment in 1965 promoted the rise of Academic Medical Centers, sophisticated subspecialty practice and rapid increases in health spending. Only after the Oil Producing Economies Oil Crisis and recession of 1974 dimmed, the once rosy economic outlook was a serious attempt made to control (rather than expand) the growth of medical spending. Yet grafting cost controls onto an expansionary system has proven difficult. Decades later these two government payment programs, originally just 2% of GDP, are projected to rise above 10% and threaten the entire budget process. Conversely, the UK National Health System was established in a context of postwar austerity in 1948. Although growth in UK spending has been substantial, sometimes more than desired, it has usually been below the OECD average and certainly well below the excessive rates in the US.

Institutional forces are hard to quantify. Empirical estimates of long-run trends (or curvature in trends) are difficult to make and seldom compelling. That said, economic historians and theorists such as Daron Acemoglu, Philippe Aghion, David Landes, Joel Mokyr, Douglass North, Mancur Olson, Dani Rodrik, James Robinson, Paul Romer, Oliver Williamson, and others have concluded that institutions are a primary factor in economic growth and development. In the case of health care, it seems apparent that macroeconomic factors prevalent when the foundations are laid can continue to exert an influence on spending for at least as long as the doctors and politicians then present continue to exist and perhaps as long as the defining institutional structures (licensure, voluntary nonprofit hospitals, and insurance pools) endure.

Population Demographics and Aging

Population can be a neutral denominator by which costs or mortality are scaled. There is little evidence to contradict the simple notion that a group or nation two (or twenty) times the size of another differs in costs or mortality *per person* (or per thousand or per million, holding other factors constant). Growth (or decline) makes the situation more complex, as the

dynamics of changing dependency ratios, disability, aging, and time-to-death come into play. Births and deaths are the basic building blocks of demographics, and both events are expensive. Although now discredited, the impression that population aging itself was the important factor accounting for rising national health expenditures probably arose because: (1) health care spending on the elderly was rising rapidly, (2) health care spending was higher in nations with a more elderly population, and (3) growing fiscal concern regarding how governments could pay for expected increases in pensions and health care services. During the 1970s and 1980s, a number of 'demographic models were constructed that projected future health expenditures using a linear matrix that mimicked the format used for projections of future pension payouts (the '*i*' are 'age–sex' categories or 'age–sex-disease-disability' categories if more detail is desired).

$$\text{Total Cost} = \sum \text{Cost}_i \times \text{pop}_i \times (\text{Total Population Growth} \times \text{Excess Cost Growth})$$

Empirical investigation quickly showed that change in the percentage of population aged 65 years or more or number of elderly accounted for only a small portion of total cost increases, with most attributable to increased cost per person (holding age and sex constant). As government and employer financing of health care expanded, the personal budget constraints that had prevented many people, especially the elderly, from spending considerably on medical care in the 1950s were largely removed during subsequent decades.

The main reason for a rising health share of GDP is secular 'excess cost growth' per person (i.e., medical costs for every age–sex category has grown more rapidly than per capita income). A secondary factor is the extra 'excess' among the elderly (again, holding age and sex constant). In the US, the ratio of spending over:under age of 65 years has moved from 168% in 1953 to 345% in 1970 and above 500% in the 1980s before falling back below 400% after 2001, with similar changes in relative spending ratios occurring in most OECD countries. Governments were spending more, a lot more, on elderly people who had been significantly relieved of the financial burden of doing so. A false impression of causality was created as economic development led to concurrent rises in both average age and per capita spending for most nations. A panel study of OECD data by Getzen demonstrated that the cross-sectional association between age (%65+) and expenditures at a point in time tends to disappear once income effects are accounted for, and also that more rapid growth in the elderly population of a country during the decades 1960–1990 was not correlated with that country's rate of growth in real health spending per capita (illustrated below in **Figures 4(a)** and **(b)**).

Most health economists now agree (even when arguing details, estimation procedures, and causes) that it is more (excess) spending per person, and not population aging, that threatens the fiscal health of nations. Why then were commentators so convinced three and four decades ago that 'aging causes higher health care costs,' – and why was that mistaken impression so persistent when it could so easily be overturned by empirical investigation? Confusion arose from a failure to distinguish between micro and macro phenomena as well as the facile but misleading association of concurrently rising trends. At the individual micro level, older persons do spend more than younger persons because older people are usually sicker and stand to benefit more from therapy. Pooled government financing strengthens the connection between an individual's age and medical expenditures by removing the personal budget constraint. However, the system also disconnects total (and hence average per capita) financing from need. At the national macro level, spending decisions (total funds available) are driven by budgets, not by need or illness. A nation populated only by poor old people suffering from diabetes, dementia, and other illnesses would have to spend less, not more, on health.

Macro (National) and Micro (Individual) Expenditures: Budgets and Allocation

Asked why health care spending is so much higher in Germany than in Ghana, most respondents quickly offer the answer that Germany is much richer. When pressed, they acknowledge that need and potential benefits from medical care are likely to be much higher in Ghana, but 'the funds are not available there.' The connection between purchasing power and spending, so obvious at the national level, is often obscured in microeconomic analyses of aging, disability, or time-to-death. Clarification comes from recognizing that on one hand the use (allocation) of available medical resources is determined by clinicians on the spot immediately responding to the health of the patient, while on the other hand the total amount of national medical resources available (budget) to treat patients is determined through the political process, shaped most strongly by fiscal policies that respond slowly, and with a lag, to changes in GDP (national permanent income).

Finland, like most European countries, is steadily aging. From **Figure 3(a)** above, it is evident that spending on health care was severely restricted in Finland after the deep recession of 1992–94, and also the response was slow, delayed for 2 or 3 years. One searches in vain for evidence that per capita spending for Finland, or any other country rose or fell in response to changes in health status – or that differences in the rates of death, disability, or aging were matched by differences in the rate of growth in spending. Pooling of funds through insurance and tax financing removes the budget constraint from the individual, so that personal income is no longer a major factor determining the amount of care used. However, the budget constraint still applies for the pool as whole (in the case of Finland, the nation), so in aggregate the sum of spending on all individuals is constrained by the average contribution paid in (which, in turn, is usually strongly correlated to per capita GDP). Of course, some medical spending is made by patients from their own budgets or by subnational entities (kin, employee groups, neighborhoods, counties, and provinces) constrained by their own budgets – hence more related to per capita income of that particular group than the nation as a whole.

Spending depends on who makes the decision and how. For food, housing, transportation, and most other consumption, total spending is the sum of many individual decisions. Medical

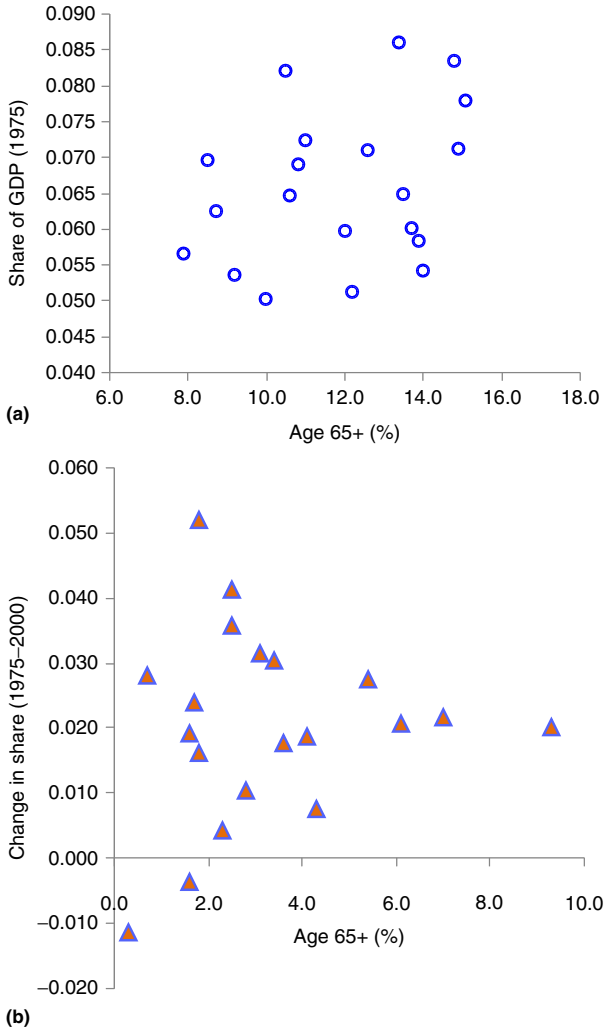


Figure 4 (a) Health share of GDP/ % population of age more than 65 years (twenty OECD countries, 1975). (b) Change in share/change in % population of age more than 65 years (1975–2000).

decisions, conversely, are made by professional agents (physicians) operating within a highly structured system dominated by third-party insurance or tax financing, divorcing spending on an individual from that individual's budget.

Growth, Business Cycles, and the Long Run

A Tale of Two Necessities

Housing and healthcare are both generally considered 'necessities,' although neither conforms to 'Engels law' or meets the technical definition used by most economists (income elasticity <1.0). What they have in common is that both are considered vital and are sufficiently expensive as to require external financing for mass consumption. Housing needs are financed through the mortgage and rental markets, which pool the resources of investors, banks, and other intermediaries. Health care financing needs are met through broad-based taxation and employer insurance pools.

What sharply distinguishes the two sectors is dynamics – different, almost opposing, responses to macroeconomic fluctuations. Housing swings wildly with the business cycle, anticipating and amplifying the ebb and flow of money, employment, and interest rates. Healthcare plods along, an inertial stabilizer that muffles shocks, only belatedly registering the effects of booms and busts, and then with such long and variable lags as to smooth business cycles into near invisibility (Figures 2–4). Differences in financing mechanisms account for much of the differential in dynamics. As pointed out by Robert E. Hall and legions of other macroeconomists, it is the flow of money, which links regions, sectors, and countries – and puts the economy at risk of business cycles, with interest rates as the key transmitter. Slack and contraction make adjustment to financial frictions problematic, sometimes sufficiently so that the equilibrating mechanisms are seriously compromised. The use of money to facilitate transactions, allocate capital, provide credit, and spread risks comes at a cost that rises exponentially during a systemic crisis, with savings, interest rates, and employment unhinged. Housing is

highly leveraged with debt financing and bears the brunt of adjustment. Healthcare is not. It is, in contrast, routinely financed within a pay-as-you-go framework by government or employers. Medicine proceeds with blithe indifference to financial markets. Doctors, nurses, administrators, and even pharmaceutical companies are often unaware of and relatively unaffected by interest rates. Stock markets soar and crash with little more effect on the operation of hospitals than sunspots or tidal waves. The only oscillation that seems to be generated within the healthcare financing system is the 'underwriting cycle' of alternating hard and soft markets for private insurance premiums, pushing quoted rate increases slightly above or below the rise in medical costs. The private health insurance underwriting cycle, however, has a little power and is often offset by countervailing trends in government tax financing. Probably the only way to get a real and significant financial disruption of the medical sector would be to put corporate and government financing under such stress that the entire structure was threatened. Fortunately, this has not yet happened, or at least not with sufficient force as to be evident in the modern national economies and health systems characteristic of most OECD countries since 1960 (although continuing fallout from the 2008 to 2009 recession and subsequent bank collapses in Iceland, Ireland, and Portugal may put that to the test).

Limits to Growth and 'The Great Inflection(s)'

No matter how vital or necessary, there is a limit to the amount of spending on any sector. Expenditures cannot logically exceed 100% of income (at least, not for any extended period of time). Currently, most wealthy OECD economies seem resistant to spending much more than 10% of GDP on health care, with the US at 16% a notable exception. The rise in health spending was very rapid during the 1960s and 1970s, moderated a bit during the 1980s and 1990s before bumping up around the turn of the millennium and then becoming somewhat restrained over the past 5 years (an uptick in share for 2008 and 2009 is mainly because of countries having a temporary decline in GDP rather than a more rapid rise in real health expenditure). Figure 5 was constructed using historical data from the UK and US, but the shape would be similar for other OECD economies – and is remarkably like the typical inflected logistic *S*-curve that characterizes most growth processes. A long period of slow growth (incubation) builds toward an explosive spurt (exponential growth) that is bent back (inflected) as it approaches some constraint that limits growth in the long run (upper bound/stability).

Many aspects of health and economics appear to follow a typical growth process during the nineteenth and twentieth centuries: Medical costs, life expectancy, population, urbanization, industrialization, trade, workforce participation, and GDP all trace recognizable *S-curves*, although each differs somewhat in shape and timing. The growth of per capita income slid along at a very low rate for millennia before rising abruptly after 1850, soaring for a century, and appearing to stabilize (?) near an upper bound of 1–2% within the next century. Life expectancy fluctuated from 20 to 40 years before

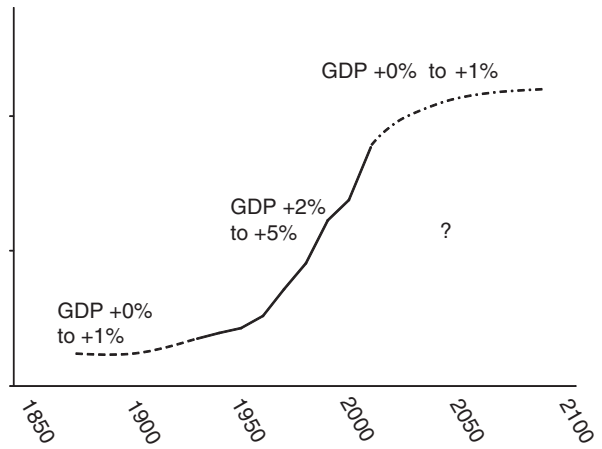


Figure 5 Growth curve: Health share of GDP 1850–2100?

making tremendous gains during the twentieth century and is expected to face diminishing returns as genetic and social factors impose an upper limit (110, 125, or ?). Global population took hundreds or thousands of years for each doubling from prehistoric eons through the middle ages before reaching half a billion around 1550, quickly climbed to one billion by 1820, two billion by 1925, three billion in 1960, six billion in 2000 – and is projected to stabilize after reaching a peak of 10 billion within the foreseeable future. The process of 'demographic transition' traces a similar curve, reversed and displaced in time.

The coincidence of so many dramatic changes in human society could hardly be attributed to chance, yet the causality and order is much debated – and has only recently (and partially) been illuminated with empirical data through the efforts of cliometric economic historians such Gregory Clark, Dora Costa, Robert Fogel, Angus Maddison, and others and placed within a conceptual framework by development theorists and macroeconomists such as Daron Acemoglu, Oded Galor, Chad Jones, Michael Kremer, Rodrigo Soares, and others. The tentative consensus among these scholars is that the gradual increase in knowledge and technology over millennia brought about an end to the Malthusian era, appearing first as a dramatic increase in total population and urbanization, a shift from agriculture to industry, a decline in mortality, and a steady increase in income per capita – transformations that were well underway toward the end of the nineteenth century and clearly before the rise of modern (expensive) scientific medicine or the modern increase in life expectancy beyond the biblical three score and ten. This cluster of transformations by which humanity escaped the era of Malthusian constraints in a burst of exponential growth is variously known as 'development,' 'demographic transition,' 'the industrial revolution,' the 'modern era', or most misleadingly, as 'normal times.'

In 2001, macroeconomists were discussing 'the great moderation,' showcasing compelling results from rational expectation models and financial forecasts based on recent time series data. By 2011, such discussions were supplemented or supplanted by observations on the great depression and financial panic of 1873. The postwar 'normal' should be seen as

an aberration – calm at the eye of a storm that transformed human society and is not yet finished.

Complexities of Measurement and Specification

The reason that business cycles are mostly invisible in health care spending and employment is not that inflation and GDP growth have no effect, but that the relationships are misspecified. To estimate the effect of one variable on another, the units of observation must match the span of action in both time and space. One hundred or even one thousand observations cannot capture the effect of a change in GDP on life expectancy or health expenditures if each observation is one minute long. A minute-by-minute time frame is, however, quite useful for determining the effect of a 10.00 a.m. announcement of clinical results on the price of an exchange-traded biotech stock. Using minute-by-minute, hourly, or even daily measures will tend to increase the signal-to-noise ratio and obscure the long-term low-frequency effects of a recession on health employment or mortality rates. Note also, that observations on the price of a specific biotech stock are neither likely to reveal the broad effects of macroeconomic factors on the market as a whole nor is investigating the determinants of one individual's medical costs during an illness episode likely to reveal the factors that cause national health spending to rise over years or decades.

Inequality and nonlinearity

Not long after Samuel Preston published his analysis in 1975 showing that the positive effect of income on longevity became progressively smaller at higher levels of income, GB Rodgers used a multivariate regression to show that for any given level of per capita GDP, greater income inequality (Gini coefficient) was associated with lower life expectancy. This led to suggestions that inequality and social stress could be an underlying cause of disparities in mortality by ethnicity and occupation status. However, work by Gravelle and others subsequently made it clear that what appeared to be an independent factor was instead an artifact due to nonlinearity: any mean-preserving spread would necessarily cause the estimated coefficient of 'inequality' to be negative – the mortality reduction obtained by the higher income group from gaining $1000 would (diminishing returns) be smaller than the mortality increase imposed on the lower income group losing $1000. Subsequent studies have supported this explanation. An extensive review of the literature by Deaton in 2003 concluded that there is still no compelling evidence that inequality in itself is a major cause of population mortality rates once sufficient care is taken to consider the effects of nonlinearity and other contributing factors.

Income, education, wealth, or socio-economic status (SES)?

Why spending depends on broader measures such as 'permanent' or 'shared' income rather than current individual earnings is fairly evident. Categories and concepts, like temporal boundaries, may also be indistinct. Demographers, sociologists, epidemiologists, and public health researchers examining the connection between income and health at the individual micro level are apt to use a broad concept of resource availability such as 'socioeconomic-status' for which 'household income' is just one aspect or indicator. For macroeconomists, the catchall term is 'level of development' or technology. Occupation, assets, poverty, and malnutrition are all associated with income levels – and with mortality. Ethnicity, education, urbanization, and social status may not be so directly related to income but are rarely independent of it – and are sometimes even stronger predictors of morbidity. The black/white differential appears to be larger in the US than the UK, but UK occupational status disparities seem to be greater. The strength and relative importance of factors varies so much across places and periods that it is unlikely that the determinants of health are constant or fixed, even though almost every region has ethnic (Inuit, Sami, Maori, Romani, etc.) or other groups (widows, orphans, albinos, and refugees) for which health outcomes are persistently worse than average.

Macro models

Measuring income, SES, or economic development is difficult but less problematic than quantifying 'health.' Longevity and mortality rates are clear but crude measures, and neither is applicable to individuals. More detailed, specific, or nuanced assessments (activities of daily living, Euro QoL quality of life measurement-36, quality of life, diabetes prevalence, anti-depressant drug expenditures, disability days, cancer survival, hospital utilization, psychiatric visits, etc.) are all sufficiently incomplete or ambiguous that none can be satisfactorily aggregated to macro measures of 'real' health outcomes. Analysis of system effects is further complicated by reverse causality between health and income – and also by interactions between marital status and occupation, education, and family size, or almost any set of contributory factors. Although each variable has a distinct connotation that is important in certain contexts, they are almost always acting together in related ways that make it difficult, if not impossible, to decompose a compound total network effect into shares, or to reliably estimate an independent coefficient for each variable.

Empirical analysis of macro determinants is often quite limited by the time frame and number of large-scale long-run observations available to discern diffuse and low-frequency responses. Temporal, spatial, and organizational boundaries must be carefully specified to distinguish and reveal micro and macro effects. Changes in coefficients as the unit of observation expands or contracts can be a key for understanding the underlying structure of the process – opening up the institutional black box of a firm, a hospital, the medical profession, or pharmaceutical discovery. The fact that health care employment adjusts quite slowly to inflation tells us something about wage formation within this industry; a mismatch between price indexes and expenditure patterns suggests that little significance should be attached to publicly listed prices; the fact that pharmaceutical research and development is more strongly correlated with prior firm profits than future prospects suggests something about capital allocation within the industry; disparity between individual cross-sectional expenditure estimates and national time series

results may be a useful indicator of the likelihood that a specific policy will be able to 'bend the (national) cost curve.'

Conclusion: Structure and Lags in the Macroeconomics of Health

The health sector is technologically dynamic but fiscally inertial. Major change often takes decades rather than months or years. Responses to macroeconomic shocks are delayed and damped by organizational rigidity so that ordinary business cycles are mostly smoothed away. Price changes, whether physician fees, hospital charges, medical care price index, consumer price index, inflation, or interest rates, can make appropriate measurement difficult but appear to have little effect on aggregate real health resources or outcomes. The process is subject to highly variable lags and complicated by interactions and feedback among variables to such an extent that almost any broadly correct generalization has one or more counter examples that can be named. Coefficients are difficult to estimate with precision and parsing total network effects into a linear combinations or shares for each factor is not very meaningful.

With regard to the current state of the literature, it may be said that since 1960 the development of national health accounting and a host of econometric studies have allowed us to become more precise about what we know and do not know, and considerably more humble about how easy it is to decompose and discern the relative contributions of various factors. Despite the humbling lack of progress in specifying many of the mechanisms and magnitudes involved, several popular hypotheses (aging, unemployment, and inequalities) have been rejected by repeated empirical tests. Some tentative conclusions are probably justified by the extensive research to date:

- The relationship of national GDP to mortality and health expenditures is strong, but not simple or constant.
- Responses are usually delayed, subject to long and variable lags. The inertial smoothing means that most effects of ordinary business cycles are rendered nearly invisible.
- The spatial and temporal boundaries of observations must be matched to the decision process of the phenomena to be estimated. Often the long-run effects are not the same as short run and may even have the opposite sign: for example, unemployment is associated with decreases in short-run mortality but increases decades later. Macro effects on national outcomes and measures are not the same as micro effects on individuals: for example, getting older greatly increases personal risk and individual medical spending, but population aging has little, if any, effect on average per-capita expenditures.
- The main determinants of individual medical costs (illness) have almost no effect on national health expenditures, which are largely shaped by budget and political pressures. Institutional factors (licensure, nonprofit hospitals, and government financing schemes) seem to dominate with prices, including interest rates, playing a much smaller role in health than other sectors.
- Income is intertwined with social organization, ethnicity, education, and other factors in a complex way that precludes any clear decomposition or reliable estimates of independent or relative importance. The magnitudes and interactions of these effects are demonstrably different for different causes of death, for different countries, and for different time periods.
- Nonlinear flattening of the income-mortality curve at the upper end implies that a mean-preserving spread will help the poor more than it harms the rich, thence reducing average mortality, but income inequality *per se* does not have much, if any, independent effect on aggregate mortality rates.
- Demographic transition, industrialization, urbanization, education, life expectancy, increases in health expenditure growth, and other aspects of modern development all appear as typical logistic *S*-shaped growth curves during the twentieth century. This suggests that the postwar span of rapid growth, rather than being a new normal equilibrium, was more like the inflection point in a centuries-long turbulent process of global development that has not yet achieved a long-run steady state.

See also: Aging: Health at Advanced Ages. Dynamic Models: Econometric Considerations of Time. Education and Health in Developing Economies. Education and Health. Global Public Goods and Health. Macroeconomy and Health. Nutrition, Health, and Economic Performance. Panel Data and Difference-in-Differences Estimation

Further Reading

Abel-Smith, B. (1967). *An international study of health expenditure.* Public Health Papers No. 32, Geneva: WHO.
Cutler, D., Deaton, A. and Lleras-Muney, A. (2006). The determinants of mortality. *Journal of Economic Perspectives* **20**, 97–120.
Fogel, R. (2004). *The escape from hunger and premature death, 1700–2100.* New York: Cambridge University Press.
Galor, O. (2011). *Unified growth theory.* Princeton, NJ: Princeton University Press.
Getzen, T. E. (2000a). Health care is an individual necessity and a national luxury: Applying multilevel decision models to analysis of health care expenditures. *Journal of Health Economics* **19**, 259–270.
Getzen, T. E. (2000b). Forecasting health expenditures: Short, medium, and long (long) term. *Journal of Health Care Finance* **26**, 56–72.
Hall, R. E. (2010). Why does the economy fall to pieces after a financial crisis? *Journal of Economic Perspectives* **24**, 3–20.
Newhouse, J. P. (1977). Medical care expenditure: A cross-national survey. *Journal of Human Resources* **12**, 115–125.
Porter, R. (1999). *The greatest benefit to mankind: A medical history of humanity.* New York: Norton.
Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies* **29**, 231–248.
Smith, J. P. (1999). Healthy bodies and thick wallets: The dual relation between health and economic status. *Journal of Economic Perspectives* **13**, 145–166.
Swift, R. (2011). The relationship between health and GDP in OECD countries in the very long run. *Health Economics* **20**, 306–322.
Weil, D. N. (2007). Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* **122**, 1265–1305.

Relevant Websites

http://www.ggdc.net/maddison/
Angus Maddison Project (World Population and GDP 0 to 2010 CE).

http://www.euro.who.int/en/home/projects/observatory
European Observatory on Health Systems.
http://www.cms.gov/nationalhealthexpenddata
OACT-NHE (US Spending and Projections).
http://www.oecd.org/
OECD Health Data.
http://www.soa.org/research/research-projects/health/research-hlthcare-trends.aspx
SOA Society of Actuaries (Long Run Medical Cost Trends Model).
http://www.who.int/whr/en/index.html
WHO (World Health Report).
http://data.worldbank.org/
World Bank Data.

Macroeconomic Effect of Infectious Disease Outbreaks

MR Keogh-Brown, London School of Hygiene and Tropical Medicine, London, UK

Introduction

Infectious disease outbreaks such as pandemic influenza or severe acute respiratory syndrome (SARS) in 2003 are, thankfully, rare events, but they do occur with some degree of regularity and impose a significant public health burden over a short period of time. For instance, there were three influenza pandemics in the twentieth century: in 1918, 1957, and 1968–69. Each was characterized by the rapid global spread of influenza. In the UK (and many other countries), there were three distinct waves of the 1918 pandemic, each lasting 10–15 weeks, with the largest occurring in the autumn of 1918. The 1957 pandemic occurred in the autumn of that year and comprised a single wave of approximately 15 weeks. The 1968–69 pandemic affected the UK somewhat late in the normal influenza season, resulting in a small first wave in March 1969 and a main wave in mid-winter of 1969–70. Clinical attack rates for the 1918 pandemic are thought to have been approximately 25%, and approximately 2.5% of those infected died. The 1957 and 1968–69 pandemics had higher clinical attack rates, ($>$30% and reaching 45% in some places), but with lower rates of mortality (case fatality ratio approximately 0.04%). The duration of illness is estimated at approximately 5–7 working days per case. In addition, the Swine Flu pandemic of 2009 was a reminder that pandemics seem to occur approximately every 30–40 years, but due to the mildness of the strain, greater population immunity, and the unusual timing of the pandemic there were fewer cases and deaths.

The structure of this article is as follows. First an overview of the link between outbreaks, productive labor supply, and economic effects is provided with reference to morbidity, mortality, and additional absenteeism. A brief discussion of healthcare expenditure related to outbreaks follows this before considering the more weighty issue of behavioral change in response to outbreaks. In the Section Health-Related Expenditure a discussion of the two main methods for macroeconomic assessment, retrospective analysis, and prospective modeling are discussed and selected results are presented and contrasted before conclusions are drawn.

Labor Supply Effects: Morbidity and Mortality

Without entering into a formal analysis of the effects of these outbreaks on productive labor supply, it is evident that, based on the observed 35% clinical attack rates for influenza pandemics, an additional period of absence in a given quarter year by one-third of an economy's labor supply, will have a notable effect on its productive capacity. Simple economic theory teaches us that labor needs to be combined with capital and natural resources in order to produce goods. Although some degree of substitution of labor with capital may mitigate the impact of a reduction in labor supply, such substitution is unlikely to be able to counter the loss of, perhaps, 3–5% of an economy's labor supply in a given quarter. Also, although the value of labor lost from the wage that is assigned to that specific quantity of labor can be estimated, this does not necessarily reflect the loss of productive capacity and the knock-on effects as the decline in one sector reduces the supply of intermediate goods to another and hence just-in-time deliveries are no longer able to meet their tight time targets. Although many of these economic effects are difficult to quantify, it should be evident that productive labor supply is just one element of the full economic cost of an outbreak. Clearly, the morbidity and mortality effects of an infectious disease outbreak vary greatly from perhaps 40 million worldwide deaths (aside from morbidity effects) from Spanish flu to 800 worldwide deaths from SARS. However, research suggests that the economic impacts of SARS were much greater than previous pandemics and much of this may be attributable to globalization and indirect health effects. Therefore, there is some evidence to suggest that direct health effects are not necessarily the only or even the main determinants of economic impact, even if they can be correctly estimated.

Labor Supply Effects: Additional Absenteeism

The loss of productive labor supply through illness and death is not the only factor which could reduce labor inputs to production. During school closure, government policies to mitigate an infectious disease outbreak may also be imposed. In extreme cases, these policies could include advising workers to avoid attending their place of work and, where possible, to work from home. However, a much more likely policy, (and one which the UK implemented in the mild swine flu pandemic in 2009), is a policy of either blanket or reactive school closure. Because children have lower immunity to influenza, mostly attributable to their lack of exposure to previous pandemics, higher clinical attack rates tend to be exhibited in schools than in the population at large. By closing schools, it is intended that the rate of infection amongst children will be reduced and thus will decrease or slow the burden of illness in the population.

However, the closure of schools, particularly primary schools, has an impact on working parents, some of whom may have to take leave from work in order to care for their children. Labor force estimates from the UK suggest that an average of 4.8% of working days will be lost in the quarter of the pandemic due to school closure that lasts 4 weeks, or 15%, if they close for the duration of the outbreak. Some of these estimates may be reduced when informal care arrangements and the ability of some parents to work from home are accounted for. However, the potential for a policy of school closures to result in a greater labor supply loss than the direct health-related effects is evident, and such costs occur in all sectors, not just health. As with the morbidity and mortality effects previously mentioned, these labor force losses will have

ripple effects throughout the economy, which need to be captured from the whole-economy perspective.

Health-Related Expenditure

In addition, many of those who are unwell and absent from work will not visit a hospital or primary-care facility, choosing rather to self-medicate, thus creating another health-related consumption change which may be hidden from healthcare sector expenditure. Some of this consumption may be captured by pharmacies in terms of increased purchase of, for example, pain medication and cold/flu remedies, but other purchases such as face masks and antibacterial hand gels will extend beyond the usual domain of treatment/prevention costs.

Externalities: Behavioral Change

Perhaps the largest potential contributor to the economic cost of infectious disease outbreaks is the externality of behavioral change. Many of the externalities which could potentially affect the economy as a result of an infectious disease outbreak are fear driven and difficult to predict, yet there is evidence to suggest that they do occur. Mention has already been made of the potential changes in shopping behavior which is linked to communicable disease, such as the purchase of self-medication, face masks, and alcohol gel, but more extreme changes in behavior may occur resulting in greater economic effects.

A survey was conducted in the follow up to SARS in eight countries (five European and three Asian), with a sample size of approximately 3500 individuals, to estimate the potential extent of precautionary behavior in order to avoid a pandemic. Although preferences elicited in this way may not reflect real behavior during an outbreak, conducting the survey shortly after the SARS outbreak may be of assistance in improving the validity of the theoretical responses in estimating true practice. The survey results suggest that 70–80% of Europeans would avoid using public transport, avoid entertainment events, and limit their shopping to the essentials. The percentages were similar but slightly smaller for Asian respondents, although Asian respondents were less likely to avoid entertainment events. In response to other questions, approximately one-quarter to one-half of respondents indicated that they would consider taking work absence, remove their children from school, limit social contact, avoid trips to the doctor, and remain indoors.

The evidence of whether such behavioral change is likely to take place in practice will shortly be examined. However, it is anticipated that a significant economic effect would result from any event imposing a substantial change in shopping patterns, attendance at work, and patterns of travel by the public at large, almost all of which would be manifested outside the health sector.

Several potential economic effects of communicable disease have been suggested which cannot be fully (or partly) captured from a partial equilibrium approach focused on the health sector and societal cost, which brings us to consider the evidence that such effects occur and present an alternative approach for their estimation.

Macroeconomic Evidence

In general, two approaches have been used: (1) retrospective estimation from economic statistics and (2) prospective macroeconomic modeling. Owing to their retrospective/prospective directionality, these provide complementary rather than competing evidence.

Retrospective Estimation

Using national economic statistics, it is possible to retrospectively estimate the impact of a significant economic event. Economic series are notoriously variable and therefore the isolation of an event's impact assumes that all other factors remain relatively predictable or consistent. The analysis can take various forms, from a simple comparison of average statistics with those relating to an event of interest to more complicated statistical methods, and such analyses have been performed for infectious disease outbreaks.

The relatively few number of cases and deaths recorded during the SARS outbreak has already been mentioned. These low-level impacts on the productive labor supply would be expected to have little economic effect. However, the economic impacts of SARS have been estimated to be significant.

To capture the economic effect of the SARS outbreak retrospectively, a study was published in 2008 to estimate the economic impacts of SARS from national statistics. Results from that study suggest that Hong Kong suffered an approximate US$3.7 billion loss to gross domestic product (GDP) and China's GDP growth was reduced by approximately 3%. As less than 0.03% of Hong Kong and approximately 0.0004% of China's population were infected with SARS, it seems unlikely that these economic impacts are greatly influenced by healthcare costs and losses of productive labor supply due to illness.

Further retrospective examination of sector-specific effects revealed losses to tourism-related sectors (hotels, restaurants, etc) for several countries amounting to, in particular, approximately US$4.3 billion for Canada and US$3.5 billion for China. In Canada, for example, there were declines in the output of the air transportation and accommodation industry of 14% between March and May 2003 and accommodation output fell by 8%.

These effects present compelling evidence that reasonably large-scale population behavior changes took place at the time of SARS. Some of this behavioral change may have been fear driven in order to avoid infection, and other changes may have been in response to the World Health Organization directive cautioning against travel to infected regions. It is also possible that some effects were attributable to an increased fear of travel at the time of the Gulf War, which highlights the potential uncertainties of retrospective macroeconomic analysis. The pros and cons of this approach are discussed in the following paragraph.

The advantage of retrospective macroeconomic estimation is that it is based on real data and is, therefore, not limited by assumptions as are modeling studies. However, there are three main limitations with this approach. The first, as has already been mentioned, is the confounding influence of other

significant sectoral or macroeconomic effects occurring at the time of the event being analyzed. The second is the limitation imposed by data availability. National statistics data can take time to reach the public domain, often in excess of 3 years, and this imposes considerable delays on effect estimation. Finally, and perhaps most obviously, such analysis cannot be used for prospective estimation and policy analysis, which brings us to consideration of an alternative tool.

Prospective Macroeconomic Modeling

Prospective modeling is very different from retrospective estimation. Macroeconomic models are usually based on real economic data and parameterized using either econometric estimation or calibration. Modeling scenarios are an essential element of macroeconomic modeling. These scenarios are designed to reflect the policy under analysis, including any investment required to accomplish an intervention or policy change, instruments (such as tax changes) to accomplish the policy goal, and, perhaps most importantly from our perspective, the health effects implemented through changes in labor supply to the economy.

Macroeconomic modeling is strong on the issues which are not well addressed by retrospective analysis. It is used for predictive purposes and is able to isolate the specific effects of the policy under analysis. Conversely, it is limited by the scenario design and, as with any modeling exercise, is limited by the validity of the assumptions underlying the scenarios and the model itself. However, most importantly, macroeconomic modeling is able to capture the wider whole-economy effects of communicable disease, particularly those properties of infectious disease outbreaks previously mentioned, which cannot be captured from the microeconomic perspective.

Several macroeconomic studies have been conducted to estimate the cost of infectious disease outbreaks. It is neither possible nor necessary to mention all these results in this brief article, but some results which highlight the importance of the macroeconomic approach to infectious disease impact evaluation will briefly be presented.

Labor Supply Effects

The computable general equilibrium (CGE) method is an important approach to macroeconomic modeling. It consists of a system of equations which specify the behavior of economic 'agents' in an economy and calibrates them on the basis of real economic data for a given country or region. For example, the agents include firms (who combine resource inputs to maximize profits), consumers (who consume and save to maximize their welfare), government, and foreign agents. Using this approach, it is possible to compare the economic impact of counterfactual (do nothing) scenarios with scenarios which reflect changes in health and policy. Several studies have designed scenarios which consider the labor supply effects of pandemic illness alone. In particular, the UK studies use two different models: the COMPACT model of the UK and the CGE approach. The models' scenario designs differ slightly, but estimates of the GDP loss from labor supply reductions due to morbidity and mortality vary between approximately 0.2% for mild disease and up to 1% for severe disease. The scenario designs differ in their assumptions concerning school closure duration and the effect of that school closure in mitigating the outbreak, but all studies highlight that the economic impact of school closure alone is likely to impose equivalent or greater additional economic loss than the disease only effects.

Behavioral Change Effects

There may be many ways in which behavioral change can be mirrored using macroeconomic models. Two examples of this in published studies are work avoidance due to fear of infection relating to the labor supply and changes in consumption. In one article, prophylactic absence from work was modeled as an effect triggered in an individual by the knowledge that someone in their social network has died from the disease. The authors estimated the size of the average social network to be approximately 300 people, and by modeling disease scenarios of differing severity and interventions (vaccination) of differing efficacy, they were able to highlight the potentially much greater economic impact of a fear-induced response to avoid work compared with the disease only effect. The scenarios modeled showed that by avoiding a behavioral response to an outbreak, the potential value of interventions to prevent this harmful economic response might be greater than the value of the health effects alone, and had fear been the driver of behavioral change, the mortality rate of an outbreak might have a more significant effect than the number of people infected.

Changes in consumption based on the survey mentioned earlier have been captured in a macroeconomic modelling study. The modeling scenarios mirrored the postponement of purchasing luxury items: a 50% postponement of clothing purchases and an 80% postponement of goods and services. Some additional purchases were lost rather than postponed: 50% of car and service use and 30% of recreation and culture purchases. These consumption impacts contributed a first-year GDP loss increase of approximately 2% of GDP, which was 10 times the impact of mild disease alone. Although the degree to which this consumption change may take place is questionable, the ability to capture these macroeconomic effects and contrast them with the health effects demonstrates the strengths of macroeconomic modeling in the context of communicable disease.

Accuracy of Macro Models

As has been previously mentioned, the accuracy of macroeconomic models depends crucially on the modeling assumptions used. Furthermore, because macroeconomic models are designed for predictive purposes to isolate the economic effect of an event assuming all other things remain the same, it can be difficult to assess prospectively the validity of such models. However, immediately following the SARS outbreak, two macroeconomic modeling studies were published. One used the results of a CGE model designed to predict the impact of the SARS outbreak based on a 6-month

duration and capturing the changes to consumer demand and confidence in the future (investment implications). This model predicts a 2.63% GDP loss for Hong Kong and a 1.05% loss for China. The China loss, in particular, would be difficult to distinguish in such a rapidly growing economy, but the predicted GDP loss was approximately US$4.15 billion, which is similar to the approximate US$3.7 billion obtained by retrospective estimation from national statistics. Similarly, the other post-SARS study estimated the impact of SARS to vary, depending on its duration, to be between 0.2% and 0.5% for China, which would, again, be difficult to distinguish and which agrees with the retrospective study's suggestion of 'no evidence of a loss.' The estimate for Hong Kong was between 1.8% and 4% or US$3–6.6 billion, which again contains the retrospective estimate. Although this is not proof of the accuracy of macroeconomic models, it provides some evidence of their usefulness in the context of communicable disease modeling.

Conclusion

Evidence suggests that the economic cost of communicable disease, particularly infectious disease outbreaks, is more than the sum of its direct health effects. Interactions between various sectors of the economy, and the processes of combining factors of production and the externalities asociated with communicable disease, indicates that a whole economy, or a macroeconomic approach to economic analysis, is of great importance and is unlikely to equate to the 'societal' cost, which is estimated by scaling up microeconomic data. Therefore, although the detailed health sector or microeconomic approach remains very important for cost-effectiveness and cost-benefit analysis in general, it is important to remember that the health sector and its patients are inextricably linked to the wider economy and those wider economic effects must, therefore, be captured using appropriate tools such as macroeconomic analysis and modeling. By doing so, the wider implications of communicable disease and related policies can be assessed beyond the health sector at a population and economy-wide level.

See also: Infectious Disease Modeling. Macroeconomic Dynamics of Health: Lags and Variability in Mortality, Employment, and Spending. Macroeconomy and Health. Peer Effects in Health Behaviors. Health and Health Care, Macroeconomics of

Further Reading

Cooper, B. S., Pitman, R. J., Edmunds, W. J. and Gay, N. J. (2006). Delaying the international spread of pandemic influenza. *PLoS Medicine* **3**(6), e212.

Department of Health (2007). *Pandemic flu: A national framework for responding to an influenza pandemic*. London, UK. Available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_080734

Fan, E. X. (2003). SARS: Economic impacts and implications. *Asian development bank policy brief 15*. Manila: Asian Development Bank.

James, S. and Sargent, T. (2006). The economic impact of an influenza pandemic. *Canada Working Paper 2007-04*. Department of Finance. Available at: http://www.fin.gc.ca/pub/pdfs/wp2007-04e.pdf (accessed 27.07.12).

Keogh-Brown, M. R. and Smith, R. D. (2008). The economic impact of SARS: How does the reality match the predictions? *Health Policy* **88**(1), 110–120.

Keogh-Brown, M. R., Smith, R. D., Edmunds, J. W. and Beutels, P. (2010). The macroeconomic impact of pandemic influenza: Estimates from models of the United Kingdom, France, Belgium and The Netherlands. *European Journal of Health Economics* **11**(6), 543–554.

Keogh-Brown, M. R., Wren-Lewis, S., Edmunds, W. J., Beutels, P. and Smith, R. D. (2010). The possible macroeconomic impact on the UK of an influenza pandemic. *Health Economics* **19**(11), 1345–1360.

Lee, J. W. and McKibbin, W. J. (2003). Globalization and disease: The case of SARS. *Asian Economic Papers* **3**(1), 113–131.

Ministry of Health (1920). Great Britain Ministry of Health Reports on public health and medical subjects, no. 4. *Report on the Pandemic of Influenza, 1918–19*. London: His Majesty's Stationery Office.

Sadique, M., Edmunds, W. J., Smith, R. D., et al. (2007). Precautionary behavior in response to perceived threat of pandemic influenza. *Emerging Infectious Diseases* **13**(9), 1307–1313.

Smith, R. D., Keogh-Brown, M. R., Barnett, T. and Tait, J. (2009). The economy-wide impact of pandemic influenza on the UK: A computable general equilibrium modeling experiment. *British Medical Journal* **339**, b4571.

Smith, R. D., Keogh-Brown, M. R. and Barnett, T. (2011). Estimating the economic impact of pandemic influenza: An application of the computable general equilibrium model to the UK. *Social Science and Medicine* **73**, 235–244.

Macroeconomy and Health

CJ Ruhm, University of Virginia, Charlottesville, VA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

Glossary

Attenuation It is the reduction in the absolute value of a regression parameter estimate when adding variable to (or otherwise changing) the model.
Bias It is a systematic error in the estimates of an econometric or statistical model.
Confounding factors These are outside factors that are not controlled for but influence the dependent variable.
Dynamics It is the adjustment process when moving from one equilibrium value to another.
Elasticity It is the percentage change in one variable expected due to a 1% change in another variable.
Fixed effect estimate It is a method of estimating parameters in longitudinal data that focuses on deviations from within-group means.
Gross domestic product The total monetary value of all finished goods and services produced in a county in a given year.
Health capital It is the level of health as conceptualized from an investment process resulting from previous flows of health investment and depreciation.
Human capital These are the skills embodied in an individual resulting from training, education, and experience.
Morbidity It is an illness or health condition.
Neonatal mortality These are the deaths within the first 28 days of life.
Nonstationary It is an economic series that has a systematic change (usually over time) in the mean or variance.
Procyclical A condition of moving in the same direction as the overall state of the economy.
Regressors (controls) It is a right-hand side variable in a regression model.
Time price The value of the time required to obtain a good or services.
Time series It is a sequence of data points, measured typically at successive times.

Introduction

The first evidence of mortality being procyclical had been provided by Ogburn and Thomas during the 1920s – procyclical means increasing in good economic times and falling during periods of decline. Additional confirmatory analysis was supplied by Eyer during the 1970s. Nevertheless, until the preceding decade, the conventional wisdom was that health and macroeconomic conditions were positively related. A variety of analyses had been conducted by the strongest adherent of this view, Brenner (1979), who suggested that overall mortality, infant deaths, and fatalities from a variety of sources (including cardiovascular disease, suicide, and homicide) increased during economic downturns, and that morbidity, alcoholism, and admissions to mental hospitals also grew during such periods.

The view that health and economic conditions must be positively related probably rests more on strongly held prior beliefs than convincing evidence. Even a cursory look at the data raises doubts about whether this is necessarily the case. For instance, Figure 1 shows the relationship between detrended age-adjusted total mortality and unemployment rates in the US, from 1980 to 2007 (both transformed to have a mean of zero and a standard deviation of one). The two data series are close to being mirror images of each other. For instance, normalized unemployment rose rapidly during 1980–82, 1989–92, and 2000–04, whereas mortality was declining faster than its long-term trend. Conversely, improvements in economic conditions during 1983–89 and 1992–2000 were accompanied by smaller than usual declines in mortality (or even increases in some years). Such relationships need not be causal but they do suggest that skepticism is warranted with regard to the conventional belief that health improves during good economic times.

Time-Series Analyses

Research conducted before the beginning of the twenty-first century for examining the relationship between macroeconomic conditions and health, typically used a lengthy time series of data aggregated over an entire country. For instance, Brenner's influential research had utilized data from the US or the UK, covering a four-decade period beginning in the 1930s.

The typical model estimated in these types of analyses is some variation of:

$$H_t = \alpha + X_t\beta + E_t\gamma + \varepsilon_t \quad [1]$$

where H is the health or mortality outcome, E is the proxy for macroeconomic conditions, X is a set of supplementary controls, and ε is an error term. More complicated specifications are often estimated including, for example, lags of the macroeconomic variables or detrended values of the dependent and some independent variables. However, this does not change the basic nature of these estimates. The coefficient of key interest $\hat{\gamma}$ will be biased if $\text{cov}(E_t,\varepsilon_t) \neq 0$, which occurs if there are important uncontrolled for confounding factors. This will frequently be a significant problem because any long time series is likely to have omitted factors that affect health and may be spuriously correlated with economic

Figure 1 Unemployment and mortality rates, 1980–2007. Mortality and unemployment rates are detrended (using a linear time trend) and normalized to have a zero mean and standard deviation of one.

conditions. For instance, unemployment declined dramatically after the 1930s, when the Great Depression ended, but mortality decreased at the same time due to improvements in nutrition and in the availability of antibiotics. Failure to control for these causes of better health leads to an overestimate of the detrimental effects of poor macroeconomic conditions.

Presumably because of these issues, time-series studies have arrived at mixed conclusions, with the results being sensitive to the countries, time periods, and proxies for health analyzed. Recent time-series analyses attempt to correct for some inherent in earlier studies, for instance, using statistical rather than ad hoc procedures to model the effect of lags in economic conditions, and correcting for nonstationarity in the data. These innovations do not, however, resolve the basic shortcoming of using a single time series and the results remain ambiguous, although most frequently suggesting that economic downturns are associated with lower mortality.

Estimation Using Pooled Data

One solution to the problem of omitted time-varying confounding factors is to estimate models using pooled data containing time-series information for multiple geographic areas. A key advantage is that, if economic conditions evolve at least somewhat independently (across locations), this geographic heterogeneity can be utilized to control for time-varying confounding factors that have a common influence on health (across locations) at a point in time. An example is the development of widely disseminated new medical technologies for the improvement of health.

These analyses may use aggregate data (such as total or cause-specific mortality rates) or individual-level information, but with the macroeconomic proxies referring to the area and not the person. In the first case, the typical estimation model is some modification of:

$$Y_{jt} = \alpha_j + X_{jt}\beta + E_{jt}\gamma + \lambda_t + \varepsilon_{jt} \quad [2]$$

where Y_{jt} is a health outcome or input in location j at time t, E is the proxy for macroeconomic conditions, X indicates supplementary controls, α is a geographic area-specific fixed effect, λ a general time effect, and ε is the regression error term. The corresponding specification being used with individual data is:

$$Y_{ijt} = \alpha_j + X_{ijt}\beta + E_{jt}\gamma + \lambda_t + \varepsilon_{ijt} \quad [3]$$

where i indexes the individual and some of the X variables may be at the person rather locality level.

In eqns [2] and [3], the location-specific 'fixed effects' (α_j) account for all health determinants that vary across geographic areas but are stable over time. For instance, this could include persistent differences in health behaviors (Victor Fuchs' provides the classic example of disparities in lifestyles between residents of Nevada and Utah), road conditions (that affect traffic fatalities), or medical facilities (e.g., the presence of tertiary-care hospitals). The time effects (λ_t) control for health determinants varying over time uniformly across locations. This includes many innovations in medical technologies, as already mentioned, and also other factors such as national trends in eating habits. Factors that vary within locations over time are not accounted for, but this is often, at least partially, remedied by including controls for location-specific time trends.

The macroeconomic effects are then identified by comparing changes in within-locality health, behaviors, or mortality outcomes, as a function of within-locality changes in macroeconomic conditions (controlling for general time effects). This procedure exploits the fact that local economies are less than perfectly correlated. For example, California's unemployment rate rose much more rapidly from January

of 2007 to January of 2010 (from 5.4% to 13.2%) than that of either Texas (from 4.8% to 8.6%) or New York (from 5% to 9.4%).

A potential shortcoming of this procedure is that national changes in macroeconomic conditions are absorbed in the vector of time variables. Thus, the effects of localized rather than national variations in economic performance are identified and the two need not be exactly the same. Some researchers have addressed this issue by using similar estimation techniques but with data pooled across countries (rather than regions within countries), although this raises questions about generalizability of the results because institutions exhibit substantial cross-national variation.

Researchers have most frequently used unemployment rates as the macroeconomic indicator, although other measures (such as deviations of gross domestic product from trend or the percentage of the prime-age population employed) have sometimes been utilized. However, it is important to realize that these estimates do not measure the effects of an individual becoming unemployed or changing labor market status per se – which is often the focus of epidemiological studies – but instead, these rates are used as a broader marker of economic conditions. It is possible for average health to improve during economic downturns, even when there are negative health effects on those who lose jobs.

Supplementary controls vary but frequently include age, education, and race/ethnicity, with more detailed sets of regressors being generally incorporated into models that are estimated using individual-level (rather than aggregated) data. Incomes are often also included as right-hand side variables but the results must be interpreted with care, because a portion of the macroeconomic effects may operate through changes in incomes. Similar issues arise when controlling for health behaviors (like smoking or physical activity) or medical care utilization, because these may be correlated with health, but partially determined by economic conditions. A variety of methods have been used to examine dynamics of the adjustment process such as, for example, adding lags of the macroeconomic proxies to the model and simulating the effects of either temporary or lasting changes in economic conditions.

Mortality is Procyclical

Research using the longitudinal methods just described in Section Estimation Using Pooled Data has most commonly examined mortality rates. Deaths are of obvious importance because they constitute the most severe negative health shock. They are also objective and well-measured indicators of health that do not require access to the medical system for diagnosis. However, the cause of death may be measured with error, and fatalities do not capture the effects of some health problems (e.g., arthritis) that are either unrelated or only weakly related to mortality.

In a particularly influential study published in the May 2000 issue of the *Quarterly Journal of Economics*, Ruhm had examined how total, age-specific, and cause-specific mortality varied with economic conditions (primarily proxied by unemployment rates) for the 50 US states and District of Columbia over the period 1972–91. Key results, summarized in Table 1, indicate that a one percentage point increase in state unemployment rates was predicted to reduce the total fatality rate by 0.5%, corresponding an unemployment elasticity of mortality equal to −0.04. The strongest responses were for traffic deaths, other accidents, and homicides – declining by 3.0%, 1.7%, and 1.9%, respectively – but significant reduction are also estimated to occur for deaths from cardiovascular disease (0.5%), influenza or pneumonia (0.7%), and liver ailments (0.4%). Infant and neonatal mortality were also expected to fall but there was no change found for cancer deaths, whereas suicides were estimated to increase. Interestingly, although the strongest effects had occurred for relatively young adults (where mortality is predicted to fall by 2.0%), substantial reductions were also predicted for senior citizens, who rarely worked.

Following the publication of Ruhm's article, researchers have used similar methods to examine how economic conditions are related to mortality in various countries and regions of the world. These analyses include studies of 16 German states between 1980 and 2000, 50 Spanish provinces from 1980 to 1997, 96 French departments from 1982 to 2002, 13 EU nations from 1977 to 1996, and 23 Organization for Economic Co-operation and Development (OECD) countries from 1960 to 1997. Virtually, in all of these studies, it has been found that total mortality and motor vehicle fatalities decline when economic conditions worsen, with the estimated elasticities being generally similar in size or larger than those found in the US.

Deaths from cardiovascular disease are also found to fall as the macroeconomy weakens, in most studies examining them, and a procyclical pattern of deaths from influenza or pneumonia is also generally obtained. In contrast, as in the US, cancer fatalities are generally (but not always) unrelated to the state of the economy. These results are plausible. For example,

Table 1 Predicted effect of 1% point increase in state unemployment rate

Type of mortality	*Predicted change (%)*	*Standard error (%)*	*Type of mortality*	*Predicted change (%)*	*Standard error (%)*
All deaths	−0.5	0.1	Heart disease	−0.5	0.1
20–44 year olds	−2.0	0.2	Cancer	0.0	0.1
45–64 year olds	0.0	0.1	Flu/pneumonia	−0.7	0.2
≥65 year olds	−0.3	0.1	Liver disease	−0.4	0.2
Vehicle accidents	−3.0	0.2	Infant deaths	−0.6	0.2
Other accidents	−1.7	0.2	Neonatal deaths	−0.6	0.2
Homicide	−1.9	0.4	Suicide	1.3	0.2

Source: Estimated provided in Ruhm, C. J. (2000). Are recessions good for your health. *Quarterly Journal of Economics* **115**(2), 617–650.

it seems likely that deaths from coronary heart disease will induce more responsive changes in modifiable health behaviors and environmental risks than cancer fatalities. Results have been more mixed when considering mortality due to liver disease, suicide, or homicide – with predicted increases when the economy strengthens in some analyses and decreases in others.

There is some indication that macroeconomic conditions have weaker effects on mortality in countries with strong social safety nets. The results for infant and neonatal mortality also appear to differ across institutional environments, with evidence of strong procyclical variations being obtained for the US, but not for Germany or when OECD countries are the unit of analysis.

Although most research has been for the US or Western European nations, this is starting to change. Recent studies have examined data from eight Pacific Asian nations during 1976–2003 and from 32 Mexican states between 1993 and 2004. The results from Asia largely mimic those obtained for the US, with the prediction of a substantial procyclical variation for total mortality and deaths from traffic accidents or cardiovascular disease, but with (insignificant) countercyclical variation for suicides. The results for Mexico are particularly interesting. The overall findings again indicate that deaths from all causes and most specific causes of mortality (including cancer deaths but not suicides) decline when the economy weakens. However, these patterns pertain to wealthy states only, with mortality in poor states exhibiting a countercyclical fluctuation. Given the wide income disparities between rich and poor Mexican states, such results are consistent with temporary improvements in macroeconomic conditions worsening average health in wealthy areas but improving it in poor ones. The latter finding is anticipated because the marginal benefits of income are likely to be exceedingly high when incomes are very low.

Other Measures of Health

There has been less study of how macroeconomic conditions are related to other measures of health, largely because data useful for examining this issue are harder to come by. Using information from the 1972 to 1981 waves of the National Health Interview Survey (NHIS), one study had found that adult morbidity declined when economic conditions weakened, with larger reductions in acute than in chronic medical conditions. Restricted-activity and bed-days also became less common and there were relatively large reductions in the prevalence of ischemic heart disease and certain back problems. However, this study has provided evidence that nonpsychotic mental disorders increased during such periods which, when combined with prior findings of a procyclical variation in suicides, suggests that mental health may decline during periods of economic deterioration despite the improvement of physical health.

Consistent with the possibility that individuals would become (physically) 'healthier but not happier' during downturns, a study of more recent (1997–2001) NHIS data revealed that the mental health of African-American and less-educated males declined when the economy weakened. Another analyses of 10 years of data (1984–93) from the Panel Study of Income Dynamics had revealed that average self-assessed overall health status fell when local unemployment rates increased and that these effects were largely driven by psychological rather than physical factors.

Changes in Behaviors and Use of Medical Care

There is improvement of physical health during bad economic times because healthier lifestyles are adopted by individuals. Alcohol sales and drunk driving vary procyclically and most research has also indicated that alcohol consumption, dependence, and heavy drinking decline when the economy weakens. However, the evidence from individual-level data is more ambiguous, with one study obtaining the contradictory result that binge drinking increases, whereas overall and heavy drinking fall; another finds an increase in alcohol use among teenagers during such periods. Finally, data for Finland provides some evidence of a countercyclical variation in certain categories of alcohol-related deaths between 1975 and 2001; however, the reverse pattern is observed for the period surrounding the extreme downturn of the 1990s and there is again evidence obtained for a procyclical pattern of overall drinking.

Other behaviors also become healthier when the economy weakens. Analysis of data of the Behavior Risk Factor Surveillance System (BRFSS) from 1987 to 2000 has indicated that severe obesity, tobacco use, and multiple behavioral risk factors decline in bad economic times, whereas physical exercise increases. Further evidence of a procyclical variation of obesity has been obtained from an analysis of the BRFSS during 1984–2002 and in smoking and physical inactivity from a study of 1976–2001 data from the NHIS. There is also an indication that diets become healthier during bad times, although relevant data are inadequate to state this with confidence. Also, less alcohol is consumed by pregnant women during such periods and their sleeping span (which has beneficial impacts on health) increases. However, the lifestyle changes need not be uniform across countries or population groups. For instance, there is some evidence of a countercyclical variation in obesity for African-American men and possibly for Finnish adults.

Better health during downturns is not the result of greater use of medical care – the utilization of most (but not all) types of medical services declines in such periods. Specifically, there is a reduction in routine medical checkups and doctor visits, screening tests, and hospital episodes. This is probably partially due to reductions in employer-provided health insurance, but may also reflect improvements in health itself. Nor are these effects uniform. For instance, there is evidence that advanced treatments for heart disease (like coronary bypass and angioplasty) become more common in bad times and that pregnant women receive earlier and more frequent prenatal care in such periods.

Sources of Countercyclical Variations in Health

As already been mentioned, one reason for health improvement during bad economic times is the adoption of healthier

lifestyles. Some of this change probably occurs because of increased availability of nonwork time during such periods, which is important because activities such as exercising and preparing meals at home are relatively time intensive. Consistent with this is the evidence that higher time prices are correlated with increases in tobacco use and reductions in exercise and socializing. However, there are other reasons why health is being countercyclical. For instance, hazardous working conditions, physical exertion of employment, and job-related stress may all increase during economic expansions, as working hours and pace of jobs rise. Moreover, employment growth during such periods is particularly large in the construction and manufacturing sectors, which have relatively high rates of work-related accidents, and these risks are amplified by the relatively higher presence of inexperienced workers. Incomes also rise during economic booms, which help to explain the rise in risky activities such as drinking and smoking. However, the direct effect of income as estimated for mortality and other health behaviors is often mixed, with a protective impact being often observed for morbidity and functional limitations.

Health may also decline when the economy improves because the former is an input for temporary increases in the output of the latter. As already been mentioned, many individuals will be required to work harder or longer in expansions, and joint products of economic activity – like pollution, driving, and traffic congestion – present further health risks. These latter effects are not limited to persons directly involved in the labor market conditions, but instead, they may frequently be concentrated among those with health vulnerabilities, like senior citizens or infants. Such groups may also be strongly though indirectly affected when care-giving behavior among prime age individuals is modified by increases in the work-hours of their employment or by their geographic migration in search of better employment opportunities.

Relatively strong procyclical fluctuations in mortality for senior citizens were documented and discussed in Section Mortality is Procyclical (**Table 1**), providing evidence of such indirect effects. An in-depth analysis of this same issue has recently been conducted by Miller *et al.* (2009) using data from the Centers for Disease Control and Prevention Multiple Cause of Death Files covering 1978–2004. They have confirmed that there is a strong pattern of procyclical mortality for young adults (18–35 year olds), but have also shown that death rates rise strongly in good times for children (0–17 year olds) and senior citizens, particularly those aged 80 years and above. In contrast, the fatality rates of 35–54 year olds are little affected by macroeconomic conditions. They emphasize the role of factors other than 'own work behavior' (like changes in pollution or the quantity, quality, and nature of health care) as potential mechanisms for explaining these results.

Caveats and Uncertainties

Two important caveats should be kept in mind when interpreting the preceding discussion. First, that the macroeconomic fluctuations so discussed refer to transitory rather than permanent changes in economic conditions. Evidence of physical health improving during transitory downturns should not be taken to imply that permanent economic progress has negative effects. A key distinction is that temporary increases in output can only be obtained by using inputs (including health) more intensively given existing technologies. In contrast, permanent growth results from a combination of technological improvements and expansions in the capital stock (including human capital) that would generally result in higher levels of both economic output and health. For example, there is clear evidence that economic development among previously impoverished countries yields health improvements (although there is less indication of corresponding effects among already industrialized nations).

That said, additional study is needed to determine how long economic growth must be sustained before the initial negative consequences for health turn positive. Previous research permitting such dynamics has generally found that the effects of sustained changes in economic conditions usually accumulate for at least 1 or 2 years, consistent with models where flows of health capital gradually affect overall levels of health, resulting in larger increases in the medium term than initially. Attenuation in the predicted health effects of longer lasting changes in the macroeconomy is observed for some outcomes or studies, but not for others and further investigation of this topic is needed.

Several other uncertainties could be resolved by further research. Generally, one has more understanding of how the macroeconomy affects health than of the mechanisms for these effects. It is particularly important to obtain estimates of the role of environmental risks and other factors (like care giving) that are not directly related to an individual's own labor market experience but may influence health. Data limitations also make it harder to study consequences for mental health and morbidity than it does for mortality, although progress in these areas is being made. How the health effects of macroeconomic conditions vary across institutional environments and levels of economic development are also begun to be learnt, but additional study is required.

See also: Health and Health Care, Macroeconomics of. Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity. Macroeconomic Dynamics of Health: Lags and Variability in Mortality, Employment, and Spending. Noncommunicable Disease: The Case of Mental Health, Macroeconomic Effect of. Pollution and Health. Public Health: Overview

References

Brenner, M. H. (1979). Mortality and the national economy. *Lancet* **314**, 568–573.
Miller, D. L., Page, M. E., Stevens, A. H. and Filipski, M. (2009). Why are recessions good for health. *American Economic Review* **99**, 122–127.

Further Reading

Charles, K. K. and DeCicca, P. (2008). Local labor market fluctuations and health: Is there a connection and for whom? *Journal of Health Economics* **27**, 1532–1550.

Dehejia, R. and Lleras-Muney, A. (2004). Booms, busts, and babies' health. *Quarterly Journal of Economics* **119**, 1091–1130.

Eyer, J. (1977). Prosperity as a cause of death. *International Journal of Health Services* **7**, 125–150.

Fuchs, V. (2011). *Who shall live? Health, economics and social choice (expanded second edition)*. Singapore: World Scientific Publishing.

Gerdtham, U. G. and Ruhm, C. J. (2006). Deaths rise in good economic times: Evidence from the OECD. *Economics and Human Biology* **43**, 298–316.

Granados, J. A. T. (2005). Increasing mortality during the expansions of the US economy, 1900–1996. *International Journal of Epidemiology* **34**, 1194–1202.

Gravelle, H. S. E., Hutchinson, G. and Stern, J. (1981). Mortality and unemployment: A critique of Brenner's time-series analysis. *Lancet* **318**, 675–679.

Ogburn, W. F. and Thomas, D. S. (1922). The influence of the business cycle on certain social conditions. *Journal of the American Statistical Association* **18**, 324–340.

Ruhm, C. J. (2000). Are recessions good for your health? *Quarterly Journal of Economics* **115**(2), 617–650.

Ruhm, C. J. (2005). Healthy living in hard times. *Journal of Health Economics* **24**, 341–363.

Ruhm, C. J. (2007). A healthy economy can break your heart. *Demography* **44**, 829–848.

Ruhm, C. J. (2008). Macroeconomic conditions, health and government policy. In Schoeni, R. F., House, J. S., Kaplan, G. A. and Pollack, H. (eds.) *Making Americans healthier: Social and economic policy as health policy: Rethinking America's approach to improving health*, pp 173–200. New York: Russell Sage Foundation.

Stuckler, D., Basu, S., Suhrcke, M., Coutts, A. and McKee, M. (2009). The public health effect of economic crises and alternative policy responses in Europe: An empirical analysis. *Lancet* **374**, 315–323.

Managed Care

JB Christianson, University of Minnesota School of Public Health, Minneapolis, MN, USA

Introduction

This article addresses the general topic of 'managed care,' which Kongstvedt, author of the standard reference on the topic, has characterized as "…..regrettably nebulous" but "… at the very least, … is a system of health care delivery that tries to manage the cost of health care, the quality of that care, and access to care. Common denominators include a panel of contracted providers that is less than the entire universe of available providers, some type of limitations on benefits to subscribers who use non-contracted providers (unless authorized to do so), and some type of authorization or precertification system" (p. 807). He further observes that "Managed health care is actually a spectrum of systems…" (p. 807). To complicate matters, the structure of managed care organizations (MCOs) has evolved over time, reflecting the efforts of MCOs to respond to the demands of employers and public programs that offer health benefits to their employees and participants.

This article describes the evolution of managed care over the past 35 years. To provide a context for that description, it begins with a review of basic findings from agency theory as they apply to MCOs. It then describes the way in which managed care and MCOs have evolved over time, focusing on three different managed care 'eras'. In each era, it reviews the empirical evidence regarding the effect of the financial mechanisms and utilization control techniques being used by MCOs to control costs, as well as the evidence of a 'competitive impact' of MCOs. The article concludes by discussing the current state of managed care. Although elements of managed care are evident in the health care systems of many different countries, this article focuses on the managed care experience in the United States.

Managed Care Organizations: A 'Nexus of Contracts'

In the United States, MCOs contract with private sector employers and government programs to manage the health benefits of their employees or program enrollees. To a lesser degree, MCOs also contract directly with individuals to provide health insurance coverage. As organizations, the revenues of MCOs depend on their ability to satisfy the demands of their purchaser-customers. As long as there are alternative MCOs, unhappy customers can decide not to renew their existing contracts, by seeking alternative MCOs that better meet their demands. In this article, the term 'purchasers' is used to refer to employers and government programs.

Typically, the core services that purchasers contract with MCOs to deliver include (1) establishing and managing a 'provider network' through contracts with providers that specify payment arrangements and provider participation in utilization management activities, (2) paying provider bills for their services, and (3) enforcing coverage limitations. In their contracts with employers, MCOs may assume risk for medical care costs (and purchasers pay 'premiums' to MCOs) or purchasers may retain 'medical risk' (in which case 'self-insured' purchasers pay administrative fees to MCOs for obtaining services). In addition to these 'basic' services, MCOs typically offer other programs to purchasers (e.g. related to utilization management or healthy lifestyle management), with program costs being incorporated in premiums or put up for separate payment by purchasers.

Historically, MCOs have responded to purchaser desires to control their health care costs by (1) applying utilization management techniques that cause network providers to substitute less expensive services or sites of care for more expensive ones, (2) negotiating payment arrangements that contain incentives for network providers to control their costs, and (3) simply using their negotiating power to hold down unit prices in provider contracts. At the market level, purchasers also have expected, or at least hoped, that competition among MCOs for their business, or competition among providers for inclusion in MCO networks, would place downward pressure on medical care costs. Some policy analysts have urged purchasers to adopt specific 'managed competition' strategies to encourage this 'cost conscious' competition on the part of MCOs. However, aggressive pursuit of cost control by MCOs has implications for private sector purchasers. Specifically, depending on how they are carried out, MCOs' cost control efforts have the potential to reduce the value that employees place on their health benefits.

Most employers believe that health benefits are an important part of overall employee compensation, and thus more attractive health benefits can help in employee recruitment and retention, much the same as higher wages. Therefore, in their health benefits strategies, employers attempt to balance the potential benefits of aggressive MCO actions to control costs with the benefits of offering attractive compensation to employees. (Economists generally agree that employees care about overall compensation and thus if employer cost control efforts reduce the value employees place on health benefits, labor market pressures will cause wages to adjust upward in order to compensate for this, with little or no overall gain for employers.) Obviously, conditions in the market for labor affect the importance that employers place on benefit attractiveness versus cost containment; when labor markets are tight, offering attractive benefits becomes more important than when they are soft. In the latter case, employers can support aggressive pursuit of cost containment by MCOs with less risk that any associated reductions in the value of their compensation will cause employees to seek other job opportunities or that the firm will fail to attract new employees. Over time, the types of activities pursued by MCOs, and the aggressiveness with which they are pursued, will reflect the weight that employers place on these two goals for their health benefits programs. And, the most successful MCOs will be ones that can modify their organizational

structures, activities, and products effectively in response to changing purchaser demands.

Agency theory provides one conceptual framework for understanding the pressures faced by MCOs and their options for responding to them. MCOs must contract with multiple providers for the delivery of services to MCO members, besides negotiating contracts with purchasers for management of their health benefits. Historically, many MCO contracts with providers have been of the 'contingent claims' nature in that the MCO agrees to pay the network provider a specified dollar amount for the delivery of an uncertain amount and mix of services in the future. This uncertainty can relate to the types and number of people who will seek services in some future period and the nature of their medical needs. It is very difficult to arrive at contingent contracts that are satisfactory as it is impossible to anticipate all possible future events, and one party to the contract (e.g. the provider in MCO/provider contracts) may be able to characterize the state of the world, for contract purposes, in a manner that serves its interests. For example, providers may argue that patients require extensive courses of treatment if paid on a fee-for-service basis, or very limited treatment if paid on a price per person per time period (capitated) basis. The MCO may not be able to determine if the provider did the right thing given the condition of the patient, especially if there is no consensus regarding the appropriateness and efficacy of different treatment options.

Using the language of agency theory, in negotiations with network providers the MCO (the 'principal') attempts to design contracts with financial incentives that reward the provider (the 'agent') for acting in the principal's best interests. However, typical payment approaches in provider contracts (fee for service, capitation) contain relatively strong incentives for behavior that could, at the extreme, be detrimental to the MCO's interests. For example, fee-for-service reimbursement rewards providers for delivering both necessary and unnecessary services to their patients. This could increase costs unnecessarily as well as expose patients to unwarranted medical risks. This being the case, and depending on the information at their disposal, purchasers might seek out MCOs that are able to negotiate provider contracts that minimize these undesirable outcomes. And, MCOs may attempt to incorporate rules and monitoring mechanisms in the contracts with providers that reduce the likelihood of an overaggressive response of the latter to financial incentives. Also, MCOs may seek to mitigate the incentives in 'pure' payment approaches such as fee for service by employing other financial rewards in contracts, such as payments for meeting care process goals (e.g. periodic testing for blood sugar among diabetic patients, not prescribing antibiotics for treatment of upper respiratory infections or reducing use of magnetic resonance imaging (MRI) in first visits by patients with lower back pain). In practice, there are many different so-called 'blended payment' arrangements accompanying the rules and monitoring mechanisms in the contracts between MCOs and providers.

The type of MCO/provider contract that emerges in any specific situation will depend in part on the competitiveness of the provider market. Where there is relatively little competition among providers (e.g. where provider concentration in a given geographic area is high), they could be expected to negotiate more favorable contractual terms. These could include higher levels of payment for services, an assignment of financial risk that more closely conforms with provider preferences, and/or less obtrusive or objectionable MCO monitoring and oversight of provider activities. In contrast, contracts with terms more favorable to MCOs are more likely where provider markets are competitive, and when excess provider capacity exists. Variations in the contracting environment such as these are likely to lead to a variety of contractual arrangements between MCOs and network providers within the same market as well as across geographic markets. And, contractual arrangements are likely to vary over time as well, being influenced by changes in the structure of the markets for provider services of specific types, the competitiveness of the MCO market, and the preferences of purchasers regarding employee health benefits.

Although MCOs are principals in their contracts with providers, they are agents in their contracts with purchasers. That is, the goal of purchasers is to negotiate contracts with MCOs that lead MCOs to act in the purchaser's best interests. If the actions of MCOs do not promote the interests of purchasers, the MCO risks incurring financial penalties (e.g. the MCO pays for medical care costs above a contract-determined amount) and/or may not have the contract renewed. Different preferences on the part of purchasers in different markets for specific outcomes (e.g. containment of specialist expenditures, avoidance of provider 'never events,' managing care for people with specific chronic illnesses) are likely to be reflected in different terms in MCO contracts with providers. Changes in purchaser preferences are likely to precipitate changes in MCO/provider contracts over time.

Evolution of Managed Care Organizations

MCOs have evolved over four decades from distinct organizations offering a single product is characterized primarily by (1) a restricted, relatively narrow, network of providers with severe penalties for out-of-network use, (2) financial arrangements that shared substantial risk with contracting providers, and (3) aggressive efforts to control utilization, to organizations that offer purchasers a choice of benefit designs for employees, most of which have (1) extensive provider networks and weaker financial incentives discouraging out-of-network use, (2) less financial risk shared with contracting providers, and (3) much more limited, targeted efforts to control utilization. This section describes this evolution of managed care in the United States, focusing on three different periods. In each case, it summarizes evidence on the use and impact of incentives and rules in MCO/provider contracts, and the market-level effects of managed care. The recurring theme in this narrative is how the changing demands of employers and their desires regarding MCO performance have shaped the evolution of managed care. Essentially, this evolution, being wedded to changes in the provider environment, has reduced the potential for MCOs to control purchaser costs through aggressive utilization management and price negotiation with providers. As a result, the role that MCOs are asked to play as agents of purchasers has changed in fundamental ways. Despite still being referred to as MCOs, many of these

organizations arguably no longer conform even to relatively broad definitions of managed care.

Early Stages of Managed Care Organization Development

Before World War II, there were a small number of organizations available to purchasers in some geographic areas that fit the definition of managed care. In particular, these organizations offered limited networks of providers at a lower cost to purchasers than conventional indemnity insurance. MCOs of this type (e.g. consumer cooperative prepaid group practices) had remained a relatively minor, but growing, component of the health insurance market in the United States until the early 1970s, when Congress passed the HMO Act. In addition to introducing the term 'Health Maintenance Organization (HMO)' into the health insurance lexicon, the Act focused employer attention on HMOs as alternatives that offered better benefit coverage at a potentially lower cost than traditional insurance.

The number of MCOs that met the legislative definition of an HMO grew steadily through the 1970s, so that by 1980 there were 236 HMOs with a total enrollment of approximately 9 million people. Over the next 6 years, however, enrollment grew dramatically to 25.7 million members in 626 HMOs. In particular, MCOs with more extensive but less integrated provider panels (IPAs), and often sponsored by local or state medical societies or Blue Cross/Blue Shield plans, emerged as competitive responses to HMOs with more restrictive provider panels. Most new HMOs during the early 1980s were IPA model plans that national HMO firms had established in local markets.

Large employers typically offered one or more HMOs as health benefits options alongside traditional plans, hoping to benefit from HMO presence in two ways: (1) some employees might choose to enroll in lower cost HMOs, accepting a more limited selection of providers and some restrictions on unfettered access in return for better coverage, and (2) the loss of enrollees to HMOs might stimulate other health insurers to more aggressively control their costs. Some policy analysts encouraged purchasers to leverage this new situation by contributing an amount equal to (or proportionate to) the cost of the lowest option toward whatever option the employee chose, with the employee paying the balance. The premise of this 'managed competition' model was that at least some employees would switch to the lower cost options, low cost plans would be rewarded with more revenue, aggressive price competition among HMOs and traditional insurers would ensue, and both employer and employee benefit costs, or at least cost increases, would be moderated.

The argument that HMOs would have lower costs than traditional insurance options rested on three premises. First, they were expected to be able to influence provider use of services because, with relatively limited plan networks, network providers received a substantial portion of their revenues from HMO contracts. Second, again because of the greater reliance of providers on specific HMOs for revenue, HMOs would be able to negotiate contracts that placed providers at risk (to some degree) for costs exceeding expectations, creating incentives for providers themselves to more effectively manage costs. And third, HMOs would be able to exercise their negotiating leverage to hold down provider unit prices.

From the beginning, there was disagreement among policymakers and in published research findings concerning whether lower costs reported for HMO enrollees were entirely the result of more effective utilization management and/or the negotiating power of MCOs. Some studies found that, when employers offered employees a multiple choice of benefit options, relatively healthy employees were more likely to choose HMO options. When offered a choice from among HMOs of different types, healthier enrollees were more likely to choose HMOs with more restrictive networks. Even so, research before 1980 did suggest that HMOs reduced the use of high cost treatment settings, especially hospitals, although more loosely organized HMOs (IPAs) were less effective in doing so. In a widely cited study by the RAND Corporation, hospital admissions in a single HMO were 40% less than in traditional fee-for-service insurance, and costs were 25% less. Research on the impact of specific utilization management techniques used by HMOs during this time period was relatively limited. However, one study reported that utilization review in hospitals reduced hospital expenditures by 12% for a sample of employer groups from 1983 to 1985. Others found similar results for use of inpatient review by BCBS plans.

Not all of the early research evidence supported the ability of MCOs to reduce costs of care or costs incurred by purchasers. For instance, a study of a single HMO found evidence that lower utilization of resources for some procedures was not always reflected in lower overall costs. Other research suggested that how physicians were paid was a key factor in explaining differences in findings for different types of HMOs. An analysis of Illinois HMOs between 1985 and 1987 concluded that providers reimbursed by HMOs using fee for service had higher rates of use of inpatient care and physician visits than those reimbursed by HMOs using other methods, except that the use of individual physician bonus payments resulted in lower utilization. Similarly, other research reported that physicians paid on a capitated basis in IPA type HMOs had service utilization rates which were comparable or lower than in group or staff model plans. This is consistent with general findings from multiple studies, indicating that reimbursement arrangements such as those placing providers at some degree of financial risk can reduce utilization of services.

There was also evidence that competition among HMOs for purchaser contracts occurred during this early period, with several studies describing competitive market dynamics that were stimulated by the development of HMOs in some geographic markets. Other studies sought evidence of an empirical relationship between HMO market presence and premiums of competing insurers, but these efforts were handicapped by the relatively low market penetration of HMOs in most communities.

Notwithstanding evidence of competitive behavior on the part of HMOs, the degree to which any real 'savings' generated by HMOs were passed on to purchasers in the form of lower premiums (for employers that were not self-insured) became a matter of dispute during this early stage of MCO development. HMOs were accused of 'shadow-pricing' traditional insurers, generating profits that were used for expansion. It was argued

that this was possible because most employers did not adopt a 'managed competition' model, choosing instead to cover the entire cost of whichever option the employee chose, or to employ a contribution strategy that substantially subsidized higher cost options. This weakened incentives for HMOs in to compete by offering lower prices to purchasers. Some purchasers may not have adapted a managed competition model because it would result in substantially higher contributions for those opting to retain their traditional insurance, thus resulting in dissatisfaction on the part of these employees with their health benefits.

The Golden Years for Managed Care

By the mid-1980s, HMOs (IPA and closed panel plans) had grown dramatically in number and enrollment. This growth continued from 1985 to 1995, with total HMO enrollment (including point of service (POS) HMOs, see below) increasing from 18 million to 58 million, and the number of HMOs from 381 to 571, peaking at 695 in 1987. From 1985 to 1992, 155 HMO mergers occurred, as well as 152 failures. In an attempt to better understand the changing HMO landscape, several studies examined the causes and impacts of HMO mergers. They found that profit-seeking HMOs seldom absorbed nonprofit HMOs in mergers, and premiums were relatively unaffected by mergers except in very competitive HMO markets, where they were higher, yet only for 1 year postmerger. Mergers did not generally allow HMOs to reach greater scale economies without improved efficiency levels.

Throughout this period, HMOs were offered as options by most large employers and as the only health benefit plan by many smaller employers. The early to mid-1990s marked a period of very low health insurance premium increases; some analysts saw this as the phase of a predictable insurance premium cycle, while others attributed this to the growing enrollment in HMOs and other types of MCOs, as well as their ability to control costs. This generated a significant body of new research on the factors that explained the lower cost of care in HMOs. For instance, a utilization review program instituted by a large national insurer was found to reduce spending on hospital care after 1 year by 8% and total expenditures by 4%. In a study that compared the treatment of heart disease in HMOs and traditional insurance plans from 1993 to 1995, HMOs had 30–40% lower expenditures, with little difference in treatments or health outcomes; the authors attributed the lower expenditures to the lower unit prices paid by HMOs. Trends in the use of outpatient versus inpatient care showed a decline in hospital days per thousand enrollees in HMOs from 1985 to 1995, whereas ambulatory visits per enrollee increased, suggesting that HMOs substituted less expensive for more expensive treatment settings. A review of studies of the use of diagnostic tests in HMOs found that HMO enrollees received fewer diagnostic tests during their inpatient stays than patients enrolled in traditional insurance plans, and did not receive any more tests on an outpatient basis. And, another study found that increases in market share of HMOs were associated with lower MRI availability between 1983 and 1993.

Research conducted during this period found that differences in payment arrangements and practice settings continued to be important in explaining differences in utilization in HMOs. For instance, one study estimated that patients in solo or single specialty group practices, where physicians were reimbursed on a fee-for-service basis, were 41% more likely to be hospitalized than when the group practice received a capitated payment.

A major factor in the growth in MCO enrollment overall (not just HMO enrollment) from 1985 to the mid-1990s was a decision by most large employers to offer Preferred Provider Organizations (PPOs) to their employees. Under this type of MCO, the penalty for seeing a provider outside of the limited network was much less severe than under the traditional HMO (where consumers bore 100% of the cost for services received 'out of network'). Typically, in the PPO model, consumers paid all costs up to a specified deductible level, then continued to pay a share of costs above that level until a specified maximum for consumer expenditures was reached. This design differed from traditional insurance in that the deductible and coinsurance rates were lower if enrollees used 'preferred' providers who agreed in their contracts to be paid set fees and also to participate in the plan's utilization management programs. Providers sought preferred status because they hoped to attract more patients and thereby generate more revenues. Alternatively, they viewed it as a means of protecting themselves against the loss of patients to providers who held preferred provider status. A key to the popularity of PPOs was that consumers could choose between seeing a preferred provider or some other provider at the point of service. By 1995, almost 35 million employees were enrolled in PPOs. HMOs responded to PPO development by devising a plan with similar provider and consumer incentives (the POS HMO), utilizing the HMO network as the preferred providers.

Skeptics doubted the ability of PPOs to effectively control health care costs because they typically reimbursed physicians using a fee-for-service approach, which rewarded provision of more services, and their preferred provider panels were large, presumably making the effective application of MCO utilization management techniques more difficult. However, the relatively modest premium increases of the mid-1990s, which were coincidental with growth in PPO enrollment, seemed to belie those concerns.

The rapid growth during this period in the number of MCOs, the number of national MCOs, and the enrollment in MCOs generated a large body of research addressing the competitive impacts of HMOs. Regarding the relationship between degree of HMO competition and level of HMO premiums, one study found lower premium revenue per HMO enrollee in markets that contained larger numbers of HMOs in combination with a relatively high percentage of the population enrolled in HMOs. Another study found that HMOs had a constraining effect on the premiums of other health insurers at low levels of HMO market penetration despite that premium levels for other insurers were higher at greater levels of HMO penetration. The authors speculated that this could reflect shadow-pricing strategies by HMOs as soon as they had established their market presence.

The impact of HMOs on quality of care was also an important topic of research during this period, that stimulated in part by concerns HMO utilization management policies and payment arrangements shifting risk to providers could have a

negative impact on quality. In general, review articles concluded that there was little support for the concern that HMOs reduced quality. For example, although one study found a negative effect of HMO competition on quality of care indicators relating to treatment of acute myocardial infarction, others found mixed or somewhat positive relationships between measures of HMO competition and quality of care.

As HMO presence grew in some markets, so did the degree of consolidation among hospitals and physician groups, raising concerns of whether HMOs could continue to contain costs by negotiating lower prices for inpatient care for their members. Quantitative analyses found that the increased presence of MCOs in local markets was not a major factor causing hospital mergers, but qualitative evidence suggested that the threat of managed care could have encouraged mergers. Irrespective of the role managed care played in stimulating mergers, quantitative studies found that hospital prices were higher in more consolidated hospital markets. Hospitals in more competitive HMO markets had slower rates of cost growth, but this HMO effect was not significant in highly concentrated hospital markets, suggesting diminished HMO negotiating leverage in consolidated hospital markets.

The Postbacklash Era: Rethinking Managed Care

By the mid-1990s, many large employers had begun to restructure their approaches to health benefits in a way that, arguably, subsequently shaped not just the trajectory of managed care, but the structure of the US health care system as a whole. First, influenced by relatively low premium increases that they attributed to the effective use of financial incentives and utilization controls by MCOs, along with their own savvy health benefits decisions, these employers eliminated their traditional health insurance options, replacing them first by MCOs and, subsequently, by consolidating the number of MCO options offered to employees toward one or two plans. By limiting the number of MCO options, employers hoped that they could reduce their health plan administration costs besides concentrating their purchasing power to achieve more favorable contractual terms with MCOs.

These employer decisions limited employee choice of health benefit options and, in effect, pushed many employees who had valued the flexibility and wide range of provider options offered by traditional health insurance into the more restricted MCO environment featuring both preauthorization for hospital admissions and limitations on referrals to specialists. New MCO members, unfamiliar with these restrictions, had their requests for reimbursement for care from out-of-network providers denied and experienced seemingly arbitrary decisions on the part of MCOs regarding access to care within MCO networks. Their unhappiness was reinforced by growing provider discontent with MCO payments, utilization review and other practice restrictions. The result was 'managed care backlash' that varied in its severity across different markets – less in areas where HMOs were well entrenched with a large market share, and more intense in markets where a large proportion of the population was affected by employer elimination of traditional insurance options.

In effect, purchaser attempts to capture a larger share of the presumed cost savings from enrolling employees in MCOs have resulted in a devaluation of health benefits for some employees. Although much of the anger of consumers was directed at MCOs, the decisions of employers to drop non-MCO plans too were resented by employees. This backlash came at an exceedingly inopportune time for employers, as the mid-to-late 1990s saw significant economic growth and competition to attract and retain employees. In this environment, employers turned to plans with broad provider networks and freedom for employees to access providers of their choice. MCOs responded by expanding their preferred provider networks, seeking to enroll as many providers as feasible in any given community, and by consolidating nationally. Blue Cross/Blue Shield plans held an advantage in this respect, as they already possessed expansive networks, and their enrollment grew, whereas enrollment in HMOs with limited networks declined or remained stagnant.

These changes had important consequences for the structure of MCOs as well as the subsequent shape of the health care delivery system in communities. MCOs sought to become 'one stop shops' to meet employer desires to minimize contracting and health benefits management costs. Those that had started as a product type (e.g. an HMO) now added other options (PPOs and, later, consumer-directed health plans (CDHP)). This allowed employers to make different benefit designs available to employees within a single contractual relationship with an MCO. MCOs, in losing their identification with a single product, became 'health plans' that offered an array of products to employers in different market segments.

At the same time, MCOs were losing the contracting leverage with providers that they had used to restrain rate increases in the past. Because they had to maintain relatively large provider networks to secure contracts with employers, plans could no longer credibly argue that providers would be rewarded with more patients and revenues if they accepted lower fees as preferred providers. Perhaps more important in the view of some analysts was the fact that providers (especially hospitals and specialty groups) merged in order to enhance their negotiating power, as health plans could not withstand significant 'holes' in their provider networks and yet be responsive to employer demands. Although the impact of managed care growth on provider consolidation is not clear, increased provider consolidation has important implications for employers; it makes it very difficult for their agents – the health plans – to hold down rate increases in contract negotiations or implement effective utilization control strategies. In fact, two studies had found that, post managed care backlash, higher HMO penetration in local markets was no longer associated with lower cost growth. And, research based on consumer surveys conducted in 1996–97 found no difference between HMOs and other insurance arrangements in the use of expensive services, but HMO enrollees reported less satisfaction with their care and less trust in physicians. Also, 2002 data pertaining to New York State suggested that a larger number of HMOs in a local health care market was associated with lower quality of care. Taken together, these findings suggest that the changes made by MCOs to meet employer demands had reduced their ability to contain provider prices or control utilization of services, leading some analysts to declare 'the end of managed care.'

Returning to the past by offering 'narrow network' benefit options in contracts with employers, similar in design to early HMOs, would be difficult for health plans, even assuming that employers were inclined to favor such options. In highly consolidated markets, it would be difficult for health plans to exclude any significant provider system and still offer a product that was valued by employees. And, because health plans now offer multiple products, if they exclude a health care system when forming a narrow network product, they risk the withdrawal of that system from other products that rely on having an extensive network for market success.

Faced with tight labor markets, and with the recent managed care backlash firmly in mind, some large employers began advocating for a new health benefits strategy known variously as consumer-centric benefits, or managed consumerism, or facilitated consumerism. At its core, this strategy focuses on creating cost-containing, quality-enhancing competition among providers for consumers, rather than competition among MCOs for enrollees in a situation where employers offer multiple MCO options. In this environment, MCOs compete for contracts with employers by offering new benefit designs that feature greater employee cost sharing, sometimes accompanied by an employer-funded health savings account, besides maintaining substantial freedom of choice among providers. MCOs are charged with providing employees with cost and quality information necessary to make informed choices of providers. Employers contract with MCOs or freestanding vendors to provide disease management programs, intensive care management programs, and 'healthy lifestyle' programs to their employees. To meet these new demands of their employer-customers, MCOs have attempted to 're-invent' themselves as organizations that encourage and facilitate the efforts of employees to more effectively managed their own health besides promoting cost-containing competition at the 'retail' as opposed to the 'wholesale' level.

By 2008, employers and MCOs had made credible inroads in modifying conventional managed care, by introducing elements of a managed consumerism strategy, although not without controversy. Skeptics argued that new 'CDHP' options offered by MCOs, featuring relatively high deductibles (in comparison to earlier benefit designs of HMOs and PPOs) coupled with health savings accounts, were simply mechanisms to facilitate greater cost sharing on the part of employees, and that MCOs were providing limited and not particularly useful information to employees to assist in the choice of providers. They also expressed concerns that CDHPs would be attractive to relatively healthy or higher income employees, but would increase costs disproportionately for sicker employees, and do little to modify employer costs or the growth in health care costs still more generally. In light of these concerns, and the reality that employers make health benefit decisions only once each year, it took several years for CDHPs to become established health benefit options for employers. However, buttressed by federal government actions that conferred tax benefits on the purchase of one type of CDHP (the 'Health Savings Account' plan) along with the experiences of early-adopting employers, 15% of employers were offering CDHP options to employees by 2010, including 34% of firms with more than 1000 employees, and overall 13% of employees were enrolled in these plans.

Research suggests that CDHP enrollees have higher incomes and are in somewhat better health than employees who do not choose to enroll in CDHPs. Employees who switched to CDHPs spent less on health care and used fewer services, but had lower levels of satisfaction with their plans, used less preventive care, and felt that they lacked sufficient information to make informed choices.

The onset of the worldwide recession in 2008 accelerated the implementation of at least one component of the managed consumerism strategy – increased employee cost sharing. Employers facing significant financial challenges focused their attention on the need to take immediate steps to reduce health care costs. Just as unemployment rates rose, employers too became less concerned regarding the possible impact of health care cost containment efforts on their ability to attract and retain employees. Large employers reduced employee compensation by increasing deductibles and coinsurance rates in PPO and CDHP plans as well as by reducing their percent contribution toward premium costs. For many employers, these actions led to year-to-year rates of increase in their health benefits costs of 5% or less.

Large employers also invested in disease management and healthy lifestyle programs to soften the impact of reductions in benefit coverage and, in some cases, because they believed that employee participation in these programs might reduce employer health benefit costs in the longer term. Targeted disease management programs, which include various utilization management components, are now a standard part of MCO offerings to employers, although the evidence that they reduce costs is decidedly mixed. More recently, MCOs have responded aggressively to employer demands to develop healthy lifestyle programs for employees. These programs reward employees for healthy behaviors and, in some cases, include benefit designs that penalize them for unhealthy lifestyles. In a growing number of case studies, programs have been identified that have favorable short-term returns on employer investments. However, other research suggests that there may be wide variation in the ability of such programs to contain costs.

The End of Managed Care?

Clearly, the concept and reality of managed care has changed substantially since the introduction of the HMO Act almost 40 years ago. Large, closed panel MCOs of the type that once exemplified managed care still exist, integrating health insurance with a health care delivery system. However, even these organizations have become 'health plans' in that they offer a variety of products to employers, including CDHPs and other high deductible benefit designs. And, despite the continued success of some limited network plans, the vast majority of employed Americans now are enrolled in health benefit options featuring broad networks, deductibles, coinsurance, and in relatively few intrusive efforts that manage the delivery of care by contracting providers. This would indeed suggest that the concept of 'managed care' has come full circle, reflecting in large part a response to the changing goals of employers for their health benefits offerings. Health plans now generally avoid the label of MCO, preferring to emphasize their evolving

role in supporting consumers both in choosing providers and engaging in healthy lifestyles.

Although some analysts suggest that 'the end of managed care' has occurred with the adoption of 'managed consumerism' by large employers, others refer to 'the changing face of managed care' instead. In fact, there are several reasons to believe that many aspects of the original concept of managed care remain relevant. First, the intensifying pressure on government to contain costs in public programs will continue to make public sector contracts with what now could be called 'traditional MCOs' appealing. At present, approximately two thirds of Medicaid beneficiaries are enrolled in managed care plans with limited provider networks, often aggressive care management, and an emphasis on primary care. (In addition, almost a quarter of Medicare beneficiaries are enrolled in a mixture of different private sector plan types, but these plans generally are less aggressive in managing care of enrollees.)

Second, some MCOs continue to practice utilization management in targeted areas, and some have reintroduced utilization management techniques that they had previously discarded. Perhaps the best example consists of efforts by MCOs to constrain the use of imaging procedures, especially as a first step in the diagnosis of lower back pain. Many MCOs conduct extensive retrospective review, and some require prior authorization and credentialing of imaging facilities and machines. It may be that there will be ongoing opportunities for MCOs to apply traditional managed care techniques to areas where growth in costs and service utilization seems excessive and indicative of poor quality, or where there are clear opportunities to substitute lower with higher cost service venues without jeopardizing quality.

There are also instances where essential aspects of early managed care, albeit controversial at a time, have now become accepted (if not always welcomed) as part of health practice. For instance, the use of data to 'profile' the practices of individual physicians and hospitals, with feedback of findings, was a standard tool employed by early MCOs to challenge provider 'outliers.' This practice continues today at a much more sophisticated and transparent level, with the results made available to some MCOs to their members, published in community reports and/or used to calculate financial rewards for providers. Early MCO support for the development and use of 'practice guidelines' is a second important example. Although providers saw these guidelines as tools used by plans primarily to contain costs, over time guidelines had achieved widespread acceptance as contributing to the practice of high quality evidence-based medicine. In this case, a utilization and quality management tool of the early MCOs has been widely adopted in the support of managed consumerism strategies, and its use will probably continue to expand, as care guidelines are increasingly being incorporated in electronic medical records.

One aspect of managed care that generally has not survived the transition to managed consumerism is the negotiation of capitated and other reimbursement arrangements that place providers at risk for costs exceeding budgeted amounts. However, this could change in the future, as some health plans are now negotiating 'shared gains' contracts with large integrated provider systems. Under these contracts, providers typically must meet quality and/or savings benchmarks in order to share a percentage of savings with health plans. In the United States, the Medicare program is encouraging providers to form Accountable Care Organizations (ACOs) that would contract with Medicare under shared savings arrangements. If a sufficient number of ACOs can be established, it shall accelerate the use of shared savings contracts between MCOs and providers in the private sector as well.

The prospects for MCOs to once again generate savings for purchasers through negotiation of deep discounts in fee-for-service contracts with providers seem to be less promising. In the early years of managed care, the considerable excess capacity in community health care systems was exposed as MCO enrollees used fewer services, especially inpatient care. Providers benefited from offering discounts so long as revenues from new MCO patients were sufficient to cover the fixed costs of unused capacity. Now, hospital occupancy rates are relatively high, and physician shortages are prevailing in many communities, reducing the value of new business to providers. Perhaps more importantly, provider consolidation limits the negotiating leverage of MCOs. It seems unlikely that this situation will change because MCOs will continue to find it difficult to restrain cost increases while negotiating favorable provider payment rates. Provider market power is also likely to inhibit the ability of MCOs to negotiate shared gain contracts that have strong incentives for cost control.

In summary, addressing the question of whether 'the end of managed care' has arrived is complicated. Some of the utilization techniques associated with traditional managed care have survived and may continue, in refined form, into the future. However, the increasing market power of providers, being supported by growing market consolidation, makes it rather unlikely for MCOs to be able to negotiate the sorts of risk sharing and discounted payment arrangements with providers that arguably were key elements to lower utilization of services and reduce costs during the early era of managed care. Interestingly, growing provider consolidation also threatens employers' managed consumerism strategy, which now depends on the willingness of a shrinking number of provider organizations to compete for patients.

See also: Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare. Health Insurance in the United States, History of. Health-Insurer Market Power: Theory and Evidence. Private Insurance System Concerns. Risk Adjustment as Mechanism Design. Value-Based Insurance Design

Further Reading

Buntin, M. B., Damberg, C., Haviland, A., et al. (2006). Consumer-directed health care: Early evidence about effects on cost and quality. *Health Affairs* **25**, w516–w530.

Christianson, J. B., Ginsburg, P. B. and Draper, D. A. (2008). The transition from managed care to consumerism: A community-level status report. *Health Affairs* **27**, 1362–1370.

Christianson, J. B. and Trude, S. (2003). Managing costs, managing benefits: Employer decisions in local health care markets. *Health Services Research* **38**, 357–373.

Claxton, G., DiJulio, B., Whitmore, H., et al. (2010). Health benefits in 2010: Premiums rise modestly, workers pay more toward coverage. *Health Affairs* **29**, 1942–1950.

Cutler, D. M., McClellan, M. and Newhouse, J. P. (2000). How does managed care do it? *The RAND Journal of Economics* **31**, 526–548.

Draper, D. A., Hurley, R. E., Lesser, C. S. and Strunk, B. C. (2002). The changing face of managed care. *Health Affairs* **21**, 11–23.

Glied, S. (2000). Managed care. In Cuyler, A. and Newhouse, J. (eds.) *Handbook of health economics*. Vol. 1A, pp. 707–753. Amsterdam, The Netherlands: North-Holland.

Hadley, J. P. and Langwell, K. (1991). Managed care in the United States: Promises, evidence to date and future directions. *Health Policy* **19**, 91–118.

Kongstvedt, P. R. (2007). *Essentials of managed health care*. 5th ed. Sudbury, MA: Jones and Bartlett Publishers.

Luft, H. S. (1978). How do health maintenance organizations achieve savings? Rhetoric and experience. *The New England Journal of Medicine* **298**, 1336–1343.

Mays, G. P., Claxton, G. and White, J. (2004). Managed care rebound? Recent changes in health plans' cost containment strategies. *Health Affairs – Web Exclusive*, W4-427–W4-436.

Regopoulos, L., Christianson, J. B., Claxton, G. and Trude, S. (2006). Consumer-directed health insurance products: Local-market perspectives. *Health Affairs* **25**, 766–773.

Vogt, W. B. and Town, R. (2006). How has hospital consolidation affected the price and quality of hospital care? *Research Synthesis Report, No. 9*. Princeton, NJ: Robert Wood Johnson Foundation.

Mandatory Systems, Issues of

M Kifmann, Universität Hamburg, Hamburg, Germany

Glossary

Adverse selection A situation in which the health insurance market is distorted because individuals are better informed about their probability of needing health care than insurers.

Community rating Regulation of the health insurance market requiring insurers to charge a uniform premium regardless of the state of health.

Libertarian paternalism A policy approach to protect individuals from making decisions against their own interest. Experts design specific arrangements which apply to individuals unless they deliberately opt out.

Open enrollment Regulation of the health insurance market requiring insurers to accept any applicant.

Introduction

A number of countries mandate that individuals purchase health insurance, a policy referred to as mandatory health insurance (MHI). It requires that all or a large part of the population purchase health insurance, which covers a substantial part of healthcare costs. This article reviews the reasons for this policy, considers issues in implementing MHI, and discusses the problems in enforcing the mandate to buy health insurance.

Rationales for Mandatory Health Insurance

Avoiding Free Riding

In most wealthy societies, there is a consensus that life-saving medical care should be made available to citizens in case of need. This creates a free rider problem on part of those with low income who can expect to receive this support when they become ill. By not buying insurance, they save the premium and enjoy a higher level of consumption as long as they remain in good health. However, as soon as sizable payments for medical care occur, these individuals qualify for free treatment. This opportunity to act as a free rider on the rest of the society can mean that *ex ante* individuals do not find it worthwhile to buy health insurance. A possible response would be to deny treatment to individuals who failed to buy health insurance. To a wealthy society, however, this is usually not acceptable or feasible. For instance, if the victim of an accident or a seriously ill person is rushed to hospital, it is unthinkable if not illegal in many countries to refuse to treat the patient because of doubts about the patient's financial means.

Making health insurance mandatory solves this problem. In this sense, it is similar to mandatory car insurance in protecting third parties from being damaged. MHI also has an efficiency advantage in this setting. To receive assistance by others in case of large payments for medical treatment, individuals may refrain from buying any coverage at all. This inefficiently exposes them to smaller risks when they do not qualify for assistance. MHI avoids this inefficiency.

Alternative policies are subsidies for buying insurance, taxes for not buying insurance, or a combination of both. However, high subsidies may be necessary to induce free riders to buy insurance, requiring substantial increases in public expenditure. Taxes have the same effect as MHI provided that they induce all free riders to take up insurance.

Paternalistic Motives

Because data on health risks are often complex, individuals have problems in making informed decisions. In particular, they may underestimate certain risks. Several studies show that individuals tend to have an 'optimistic bias' with respect to their vulnerability to health risks. For instance, individuals typically rate their personal risk with respect to health problems and other hazards between 'average' and 'less than average.'

If risks are underestimated, individuals will tend to buy too little health insurance. As in other branches of insurance, catastrophic risks, illnesses which are very costly, are likely to be insufficiently covered. As a consequence, individuals may suffer financial distress and may not be able to afford adequate treatment. This is unnecessary because insurance to cover these risks can be inexpensive, provided that the probability of getting the illness is low (Nyman, 1999).

A paternalistic response is to mandate individuals to buy health insurance contracts, which provide sufficient coverage, in particular for catastrophic risks. The requirements for these contracts would be based on the opinion of experts who assess health risks. An alternative would be to provide individuals with more information. Given that problems in processing information is the underlying cause for the interference in private decisions, however, this may only help those who have the time and capabilities of assessing in detail the risks they face. As a third way, Sunstein and Thaler (2003) propose 'libertarian paternalism,' a combination of paternalism and free choice. Experts would be consulted for designing a 'default' health insurance contract which would cover individuals unless they decide deliberately for buying alternative or no coverage.

Adverse Selection

Adverse selection in health insurance arises if individuals are better informed about their probability of needing health care

than insurers. A possible implication is market failure in the health insurance market. This argument is based on the famous analysis by Rothschild and Stiglitz (1976). In their model with two risk types, only a separating equilibrium can exist in which high-risk types obtain full insurance, whereas low risks buy partial coverage. Such an equilibrium may not be second-best efficient. Mandatory public health insurance with partial coverage can lead to a Pareto improvement.

The efficiency argument for mandatory public health insurance, however, hinges on the equilibrium specification of the model. In an alternative specification in which insurers anticipate the withdrawal of unprofitable contracts in response to their own actions and can cross-subsidize between contracts, the market equilibrium is second-best efficient, i.e., no Pareto improvement is possible given the self-selection and resource constraints (Crocker and Snow, 1985).

Starting from the premise that health insurance markets are not second-best efficient, Neudeck and Podczeck (1996), Encinosa (2001), and Finkelstein (2004) have examined the effects of MHI which is not tied to public insurance. A general finding is that this policy leads to redistribution from low-risk individuals to high-risk individuals. However, MHI is not able to implement a Pareto improving outcome compared with the unregulated market equilibrium. Whether a second-best outcome can be reached by MHI depends on the specific way of modeling the insurance market.

Enforcing Cross-Subsidies

For individuals with low income, health insurance may not be affordable. This also applies to those with a high risk of needing health care who have to pay high premiums in a market with risk rating. Mandating them to buy coverage usually does not solve this problem. However, in an indirect way, making health insurance mandatory for all can lower the price of health insurance for people with low income or high risks by enforcing cross-subsidies from others. This is the case in a social health insurance system with income-related contributions. In such a scheme, those with high income and low expected health expenditure cross-subsidize the poor and ill.

The problem that the poor cannot afford health insurance can also be solved by paying earmarked transfers for the purchase of health insurance contingent on income. Cross-subsidies between low and high risks, however, are more difficult to implement by transfers. These would need to reflect the risk type in the same way as insurers differentiate their premiums by risk type. This is a demanding task for a government transfer program and has not yet been implemented.

In the absence of a satisfactory transfer solution, MHI can establish transfers to high risks if it is combined with community rating, i.e., the regulation that insurers do not differentiate their premiums by risk type, and open enrollment by insurers. MHI is crucial in this context because otherwise low risks may prefer not to buy any health insurance to avoid cross-subsidizing high risks.

It should be noted that to some extent markets provide cross-subsidies from low- to high-risk individuals. For the individual health insurance market in the US, Pauly and Herring (2007) find that premiums are not proportional to risk, pointing to some risk pooling in the market. This can be partly explained by guaranteed renewable contracts which protect individuals from being reclassified if they turn into a high risk. However, such contracts cannot induce cross-subsidies to individuals who are already high risks at the onset of the contract.

Political Economy Considerations

Making health insurance mandatory increases the demand for health insurance. The private insurance sector may, therefore, have an interest in such a policy and may lobby to bring about such a mandate. Provided that the normative reasons above apply, this is not necessarily against the public interest. However, if competition is low in the health insurance sector, individuals may be forced to buy overpriced health insurance coverage which they do not need.

Implementing Mandatory Health Insurance

Designing Health Insurance Benefit Packages

MHI requires the definition of a minimum benefit package. Otherwise, individuals could bypass MHI by buying a health insurance contract with high deductibles at little cost to meet the mandate. The design of the minimum benefit package should follow the rationales for introducing MHI. To avoid free riding, expensive treatments for which treatment cannot be denied should be included. Paternalistic motives call for coverage of those risks which individuals tend to underestimate. If MHI is introduced to mitigate adverse selection or to enforce cross-subsidies, then the benefit package should conform to the preferences of the insured population.

To what extent these considerations play a role in the actual design of benefit packages in mandated systems has not yet been studied comprehensively. Usually, the health ministry or committees decide about which benefits are included. Sometimes, economic evaluations inform decision makers about the costs and benefits of treatments.

A rare example of an explicit process to determine a minimum benefit package is Chile's introduction of a guaranteed basic uniform benefit package. It applies both to public health insurance and to private health insurers among which individuals can choose (Vargas and Poblete, 2008). Implemented from 2005 to 2007, it is based on an algorithm of prioritization using multiple criteria (burden of disease, inequality, high costs, social preferences, cost-effectiveness, and the rule of rescue, i.e., the imperative to save the life of a person who is at risk of death even if the chances of success are low and costs are high).

Mandatory Health Insurance and Social Health Insurance

MHI on its own does not make health insurance affordable. Further measures need to be taken. MHI often goes along with social health insurance. These systems are characterized by two additional requirements. Open enrollment guarantees that high risks cannot simply be rejected by insurers. Community

rating prohibits insurers from charging risk-based premiums. Frequently, contributions are also income related, making insurance affordable to poor individuals. In Switzerland, by contrast, premiums are uniform. Health insurance is made affordable by premium subsidies that are financed out of tax revenue. Depending on the canton of residence, these subsidies are granted as soon as health insurance costs more than a certain percentage of taxable income of a household.

Mandatory Health Insurance without Social Health Insurance

MHI without social health insurance faces the challenge of making health insurance affordable to those with low income and high expected healthcare expenditure. In particular, this holds true for industrialized countries in which standard health insurance coverage usually includes access to advanced medical technology and is, therefore, expensive. Without social health insurance, health insurance has to be subsidized unless MHI refers only to a very modest benefit package. Several options are available to implement such subsidies. First, access to subsidized public systems like Medicaid in the US or FONASA in Chile can ensure that poor individuals obtain access to affordable health insurance. Second, insurers can receive subsidies if they accept low-income and high-risk individuals. Finally, a policy option is to make individuals eligible for public transfers if their expenditure on health insurance exceeds a certain percentage of their income. Such a system has been introduced as part of the 2006 Massachusetts health reform. It has also been proposed by Zweifel and Breuer (2006), who have made the case for risk-based premiums and wanted to target those with low income and high premiums through premium subsidies.

Problems of Mandatory Health Insurance

Enforcing Benefit Packages

As pointed out in Section 'Designing Benefit Packages,' MHI can only be effective if a minimum benefit package is defined. If MHI refers to a single insurer, there should be no problem in making sure that individuals actually obtain insurance with this coverage. With competing insurers, however, this task becomes more difficult. Given that some individuals do not want to buy insurance, insurers may satisfy this nondemand by selling policies which cover the minimum benefit package only on paper. Insurers, therefore, have to be monitored whether they really provide the benefit package. Furthermore, it is advisable to require insurers to build up sufficient loss reserves to secure that they can meet their obligations. Otherwise, there is the risk that the bill needs to be footed by the public, reintroducing the free rider problem at the insurance level.

Enforcing Mandatory Health Insurance

To what extent MHI can be enforced depends on the institutional context. In an economy where all individuals are employed in the formal sector and are required to spend a certain amount for health insurance, contributions for MHI can be collected via the employer and transferred directly to health insurers. This is typically the case in social health insurance schemes. By contrast, in countries with a large informal sector, MHI can be difficult to enforce. In these countries, subsidized schemes are essential in expanding coverage, because otherwise it will be hard to reach all parts of the population. A tax-financed national health system which covers the entire population may be preferable in such settings (Wagstaff, 2010).

For countries in which contributions are not automatically deducted from the wage bill, the question needs to be addressed how to treat those who do not obtain insurance or refuse to pay premiums. In Massachusetts, individuals face tax penalties if they have access to affordable health insurance and remain uninsured. In Switzerland, those who do not pay their premiums can be denied coverage for nonemergency services. This policy, however, effectively allows individuals to remain uninsured.

Limiting Consumer Choice

An evident drawback of MHI is some limitation of consumer choice. This arises from the need to specify a minimum benefit package which will not always correspond to the benefits individuals would choose according to their preferences. In social health insurance systems, this problem is particularly severe. If there is only one insurance fund, for instance, in Estonia, individuals have no choice at all unless they seek care in the private sector. Even if there are competing funds as in Germany, the Netherlands, or Switzerland, benefit packages are often tightly regulated.

The regulation of the benefits package in a social health insurance scheme with competing insurers can be a response to the incentives for risk selection which arise naturally in such a setting. The requirement to accept any individual at a uniform premium leads to expected losses with high-risk types and expected profits with low-risk types. Insurers, therefore, have an incentive to design their benefit package such that it is attractive for low but not for high risks. Regulation of the benefit package is a possible response (an alternative is risk adjustment which tries to set insurers' budgets according to the risk characteristics of their insured population, Zweifel *et al.*, 2009, Chapter 7). On one hand, minimum benefits can be specified, forcing insurers to offer benefits that are of importance for high risks, such as treatment of chronic diseases. On the other hand, imposing an upper limit on benefits can prevent insurers from including services that are of particular interest to low risks but not essential to health insurance, such as visits to sports centers. These benefits effectively reduce cross-subsidies to high risks (Kifmann, 2002).

Questionable Cross-Subsidies

As discussed in Section Enforcing Cross-Subsidies, MHI can be a means of enforcing cross-subsidies to other members of society. Equity considerations call for subsidies from high- to low-risk types. Also cross-subsidies from high-income to low-income individuals can be justified if these are not implemented through the general tax-transfer system. However, MHI can also lead to cross-subsidies which are difficult to

legitimate. For instance, individuals living in the countryside may have to subsidize those in urban areas with good access to medical care. If premiums are not differentiated according to age, then the young will cross-subsidize the elderly. Given the demographic trends in many countries, this can place a high burden on the young.

Conclusions

MHI can be implemented for several reasons. It can be a policy directed against free riding behavior by those who expect to be covered by others in case of emergency. To the extent that individuals insure too little because they underestimate their health risks, it can be part of a paternalistic intervention by the government. Combined with partial social health insurance, MHI may bring about efficiency improvements in a health insurance market characterized by adverse selection. It can help to enforce cross-subsidies from those with low health risks and high income to high-risk and low-income individuals.

An important point is that MHI by itself usually does not make health insurance affordable. It needs to be combined with social health insurance or with programs which make subsidized health insurance available for those with low income. MHI also requires the definition of a minimum benefit package. Otherwise, individuals could bypass the mandate by buying a health insurance contract with little coverage. When implementing MHI, regulators need to monitor that insurers actually offer the minimum benefit package. Furthermore, measures need to be taken to make sure that individuals actually buy health insurance.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Demand for Insurance That Nudges Demand. Risk Selection and Risk Adjustment. Social Health Insurance – Theory and Evidence. State Insurance Mandates in the USA

References

Crocker, K. and Snow, A. (1985). The efficiency of competitive equilibria in insurance markets with asymmetric information. *Journal of Public Economics* **26**, 207–219.

Encinosa, W. (2001). A comment on Neudeck and Podczeck's adverse selection and regulation in health insurance markets. *Journal of Health Economics* **20**, 667–673.

Finkelstein, A. (2004). Minimum standards, insurance regulation and adverse selection: Evidence from the medigap market. *Journal of Public Economics* **88**, 2515–2547.

Kifmann, M. (2002). Community rating in health insurance and different benefit packages. *Journal of Health Economics* **21**, 719–737.

Neudeck, W. and Podczeck, K. (1996). Adverse selection and regulation in health insurance markets. *Journal of Health Economics* **15**, 387–408.

Nyman, J. (1999). The value of health insurance: The access motive. *Journal of Health Economics* **18**, 141–152.

Pauly, M. and Herring, B. (2007). Risk pooling and regulation: Policy and reality in today's individual health insurance market. *Health Policy* **26**, 770–779.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay in the economics of incomplete information. *Quarterly Journal of Economics* **90**, 629–649.

Sunstein, C. and Thaler, R. (2003). Libertarian paternalism. *American Economic Review* **93**, 175–179.

Vargas, V. and Poblete, S. (2008). Health prioritization: The case of Chile. *Health Affairs* **27**, 782–792.

Wagstaff, A. (2010). Social health insurance reexamined. *Health Economics* **19**, 503–517.

Zweifel, P. and Breuer, M. (2006). The case for risk-based premiums in public health insurance. *Health Economics, Policy and Law* **1**, 171–188.

Zweifel, P., Breyer, F., Kifmann, M., et al. (2009). *Health economics*, 2nd ed. New York: Springer.

Market for Professional Nurses in the US

PI Buerhaus, Vanderbilt University Medical Center, Nashville, TN, USA
DI Auerbach, RAND, Boston, MA, USA

Introduction

The nursing workforce in the US is comprised of both professional nurses and nonprofessional workers. Professional nurses typically complete nursing education in a hospital-based diploma program, community college or university and are registered and licensed by the state to practice nursing. Professional nurses also include advanced practice nurses (APRNs) who are registered nurses (RNs) that have completed graduate education and practice as nurse practitioners (NPs), certified nurse midwives (CRNMs), clinical nurse specialists (CNS's), and certified nurse anesthetists. Nonprofessional nurses receive their nursing education in technical and vocational programs and are licensed by states as practical or vocational nurses. Supporting professional and nonprofessional nurses are assistive personnel, such as aides, orderlies, and personal care attendants, who have not completed formal education in nursing.

For several reasons, this article focuses on professional nurses. First, there is more complete data on RNs versus either practical or vocational nurses or the various personnel who assist nurses. Because RNs' educational preparation and legal scope of practice enable them to perform more complicated nursing services, RNs have a greater impact on the productivity of the nursing workforce, earn higher wages, and exert a greater effect on healthcare spending, quality of care, and patient safety. And, because APRNs can legally provide many of the services traditionally provided by physicians, these nurses have become a highly visible component of the professional nursing workforce.

The article begins with an overview of the key demographic, educational, and employment characteristics of RNs and then briefly summarizes the forces that affect their demand and supply. Following this, the authors examine the 'cyclical' nature of RN shortages, describe the impact of the recent recessions on hospital RN employment, and identify key issues facing the RN workforce. The article concludes with a discussion of the characteristics and challenges faced by APRNs.

Data for the tables and figures shown in this article are derived from two sources. Data to estimate RN employment growth and the age composition of the nursing workforce were derived from the US Bureau of the Census Current Population Survey (CPS) Outgoing Rotation Group Annual Merged Files. The CPS is a household-based, nationally representative survey of more than 100 000 individuals administered monthly by the Bureau of the Census. This data source is used extensively by the Department of Labor to estimate current trends in unemployment, employment, and earnings and has been used to estimate employment trends for RNs and project the age and supply of RNs and physicians (Auerbach *et al.*, 2007; Staiger *et al.*, 2009). The CPS survey contains information on roughly 3000 RNs employed in nursing each year.

The second source of data comes from the National Sample Survey of RNs (NSSRNs) conducted by the Health Resources and Services Administration (HRSA). The NSSRN is the most well-known and comprehensive source of data on individuals who have active licenses to practice in the US as RNs whether or not they are actually employed in nursing. The surveys have been conducted every 4 years from 1977 to 2008 and provide information on the number of RNs; their educational background and areas of clinical specialty; employment settings; positions; salaries; geographic distribution; and personal characteristics including gender, racial/ethnic background, age, and family status.

Key Characteristics of the Registered Nurse Workforce

Employment and Earnings

As shown in Table 1, RNs are employed in a variety of settings, including hospitals, extended care facilities, ambulatory care clinics, schools, public and community healthcare clinics, insurance companies, and others. Hospitals employ more than 60% of RNs, with the majority working on general medical and surgical care units, critical care and stepdown units, emergency departments, and hospital-based outpatient surgery and ambulatory care centers. Not surprisingly, RNs work in many different clinical and nonclinical positions both in hospitals and nonhospital settings (Table 2). Over the past few decades, data from the CPS (Figure 1) indicate that RN employment on an FTE basis has grown faster in nonhospital settings than in hospitals.

Table 1 Full-time equivalent employment of registered nurses in principal employment settings, 2008

Setting	*Full-time equivalent registered nurse employment*	*Total (%)*
Hospital	1 601 831	62
Nursing home/extended care facility	135 514	5
Academic education program	98 268	4
Home health setting	165 697	6
Community/public health setting	97 210	4
School health service	84 418	3
Occupational health	18 840	1
Ambulatory care setting (not hospital)	270 556	10
Insurance claims/benefits/utilization review	49 441	2
Other	51 947	2
Not known	22 875	1
Total	2 596 599	

Source: Reproduced from Health Resources and Services Administration (HRSA) (2010). *The registered nurse population: Findings from the 2008 National Sample Survey of Registered Nurses.* Rockville, MD: HRSA.

Table 2 Job title in principal nursing position, by hospital and nonhospital settings, 2008

Setting	*Total estimated number*	*Estimated number hospital setting*	*Estimated number nonhospital setting*
Staff nurse	1 711 271	1 232 586	478 685
Management/administration	322 790	145 574	177 216
Certified registered nurse anesthetist	29 538	23 856	5 682
Clinical nurse specialist	22 070	13 943	8 127
Nurse midwife	6 455	2 682	3 773
Nurse practitioner	98 487	36 533	61 954
Patient educator	18 405	9 053	9 352
Instruction	94 946	28 857	66 089
Patient coordinator	140 060	48 605	91 456
Informatics nurse	8 952	6 105	2 847
Consultant	23 115	3 788	19 327
Researcher	17 136	8 625	8 510
Surveyor/auditor/regulator	10 652	–	8 686
Other	92 720	39 658	53 062
Total	2 596 599	1 601 831	994 768

Source: Reproduced from Health Resources and Services Administration (HRSA) (2010). *The registered nurse population: Findings from the 2008 National Sample Survey of Registered Nurses*. Rockville, MD: HRSA.

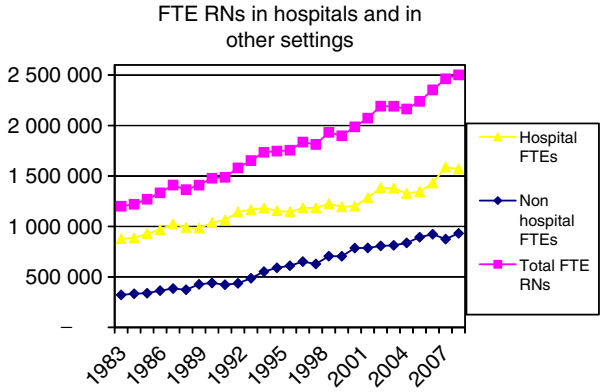


Figure 1 Total full-time equivalent (FTE) RNs by hospital and nonhospital settings, 1983–2010.

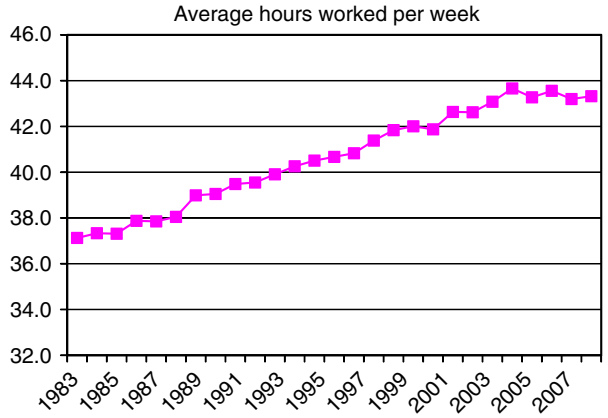


Figure 2 Average hours worked per week by RNs, 1983–2010.

Over the past several decades, the average number of weekly hours worked by RNs has been increasing. According to the CPS, the average number of hours worked by RNs during a given week increased by 2 h from 34.7 h in 1983 to 36.7 h in 2010 (Figure 2).

Using data on hourly earnings from the CPS, real (inflation adjusted) wages, for all RNs, increased 25% from 1983 to 2010. Increases in annual RN earnings were not gradual, however, as most of this increase occurred between 1983 and 1992 (Figure 3). From 1992 to 2000, real earnings stagnated or even dipped in some years, which suggests that excess capacity (too many RNs) may have existed in the nurse labor market during this period, perhaps as a result of the spread of managed care during the 1990s. During the last decade, real earnings among all RNs have increased less remarkably.

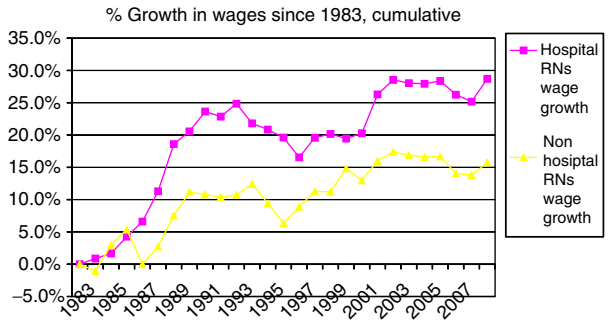


Figure 3 Cumulative percentage growth in RN wages since 1983. Reproduced from Current Population Survey.

Demographics

Although the racial and gender composition of the nursing profession has become gradually more diverse, in 2010 the vast majority of RNs were women (91%) and white (78%). RNs whose initial nursing education took place outside the US or in the US territories also contribute substantially to the RN workforce in the US. According to the HRSA (2010), internationally educated nurses (IENs) have grown as a percent of the US nursing workforce, increasing from 5.1% in 2004 to 8.1% in 2008. The dominant source country of the IEN

workforce is the Philippines (50%), followed by Canada at nearly 12%.

Age

The average age of the RN workforce has been increasing rapidly over the past several decades (Figure 4), from 37.1 in 1983 to 43.2 in 2010. Figure 5 shows the number of RNs participating in the workforce broken into three age groups: under 35 years, between 35 and 49 years, and more than 50 years. Among these groups, the number more than 50 quadrupled from roughly 200 000 in 1983 to nearly 900 000 in 2010. The number of middle-aged RNs (35–49) more than doubled over the same period from 400 000 to nearly 1 000 000, whereas RNs under 35 grew very little and in the present day are just above 600 000. These trends reflect the very large baby boom cohorts who entered nursing in unprecedented numbers in the 1970s and 1980s. In the decades that followed, other professional opportunities opened up for career-oriented women and entry into nursing declined (the groups following the baby boom were also smaller in size due to declining birth rates). Thus, as baby boom RNs have moved through the workforce, the average age has increased. Rapid renewed entry into the nursing profession over the past decade has stabilized the average age and lessened expected future shortages. Nevertheless, as large numbers of RNs of more than 50 years of age retire over the next decade, shortages of RNs may again develop.

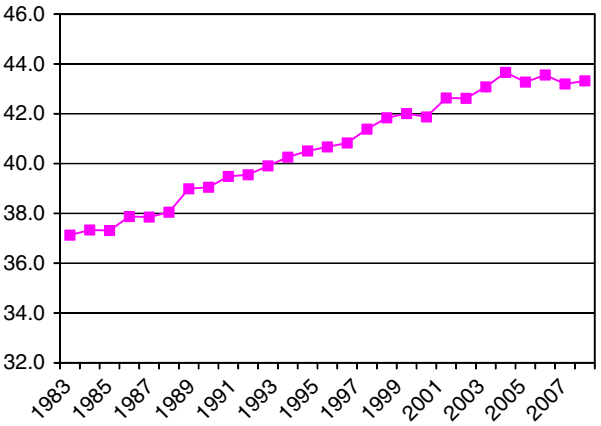


Figure 4 Average age of RNs, 1983–2010. Reproduced from Current Population Survey.

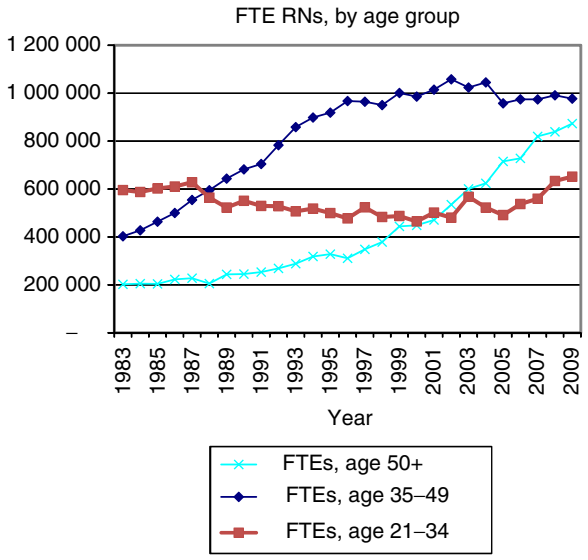


Figure 5 Full-time equivalent RNs by age group, 1983–2010. Reproduced from Current Population Survey.

Education

The educational preparation of RNs in the present day occurs in community colleges or in baccalaureate degree nursing education programs, whereas many of the large number of RNs born in the baby boom generation received their nursing education in hospital-based diploma programs. In 2008, according to the NSSRN, community colleges produced the majority of graduates in 2008 (Figure 6).

Overview of Factors Affecting the Demand and Supply of Registered Nurses

Like any labor market, the performance of the RN labor market as indicated by wages and output (number of RNs or hours worked) is determined largely by forces affecting the demand and supply of RNs. On the demand side of the market, forces arise from factors that determine society's overall demand for healthcare and from a different set of factors that healthcare organizations consider when deciding on the quantity of RNs to employ. The authors will not focus on forces that affect society's overall demand for healthcare, but rather, will focus on areas where that demand may particularly differ for RNs. With regard to the supply side of the market, which is more idiosyncractic to RNs, they distinguish

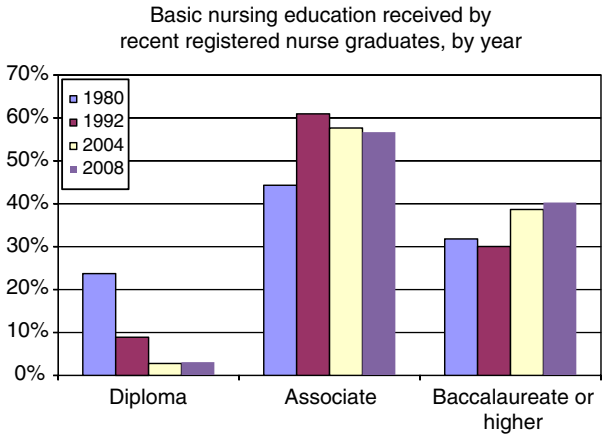


Figure 6 Basic nursing education received by recent RN graduates, 1980–2008. Reproduced from Health Resources and Services Administration (HRSA) (2010). *The registered nurse population: Findings from the 2008 National Sample Survey of Registered Nurses*. Rockville, MD: HRSA.

between forces that determine the long-run supply of RNs (the number of individuals choosing to become an RN) and forces that influence the short-run supply of RNs (participation in the labor market and number of hours worked by existing RNs).

Societal Factors Affecting the Demand for Registered Nurses

Factors that determine society's total demand for healthcare include changes in the health, size, age, and ethnic composition of the population; economic factors; and the organization of the healthcare system. As 60% of RNs work in hospitals, elements that particularly increase demand for hospital care would disproportionally increase the demand for RNs. Changes in the prevalence of diseases requiring hospital care such as congestive heart failure (or needs brought about by old age such as knee and hip replacement) could particularly result in increased need for RNs. The proportion of the US population that is more than 65 years of age will grow from 13% in 2010 to 16% in 2020 to 19% in 2030 – suggesting an increase in the demand for hospital care.

In the near future, the out-of-pocket price of healthcare services will decrease for many of the estimated 32 million Americans that will obtain health insurance in 2014 under the Patient Protection and Affordability Care Act, the health reform legislation passed in 2010, also potentially increasing the demand for healthcare. Although ambulatory care is generally more sensitive to out-of-pocket price than hospital care, it is possible that an increasing proportion of RNs will be employed in ambulatory care in the future if systems devote resources toward patient-centered medical homes and primary care-intensive preventive services (Sochalski and Weinder, 2011).

Organizations' Demand for Registered Nurses

Healthcare delivery organizations are in the business of producing goods and services to satisfy society's demand for healthcare. Because producing many of the goods and services requires RNs (other nursing personnel, other labor, and capital), the number and type of nursing personnel employed at any given time is a function of organizations' demand for nursing services. With respect to RNs, demand is determined by the productivity of RNs relative to nonprofessional nurses, assistive personnel, and capital, the wages and input prices of these other productive factors, and the ability to substitute one type of input for another in the heath production function.

Briefly, the higher (lower) the wages and fringe benefits, it must pay to hire RNs, organizations demand fewer (more) RNs, holding all else constant. The supply of RNs available at the time employers are seeking to hire additional RNs and the quantity of RNs demanded determine the wage RNs can command in the market, and hence the quantity of nurses that employers' can afford to employ. Although RNs command a higher wage than Licensed Practical Nurses, their productivity relative to their wage (marginal product) is greater because they can legally provide a greater number of nursing services. Thus, organizations' demand for RNs is influenced by whether the output organizations are producing requires nursing services that can only be provided by RNs or can be produced by using LPNs or others. For example, long-term care organizations typically provide patient care services that can be provided at less cost by LPNs, whereas the nursing services needed in most acute care hospitals require far more RNs relative to LPNs or nonlicensed personnel.

As the healthcare system has become increasingly focused on improving the quality and safety of care, hospitals have begun to pay more attention to the additional quality and safety that can be obtained by hiring RNs relative to other nursing personnel. Over the past few years, both public and private payers have begun to link hospital payment to patient outcomes that are sensitive to the care provided by RNs; should such incentives be expanded to outpatient and non-hospital settings, then demand for RNs could increase in these settings as well.

Organizations also consider the changing relationship between capital and labor when they determine their overall demand for RNs. Clearly, the combination of resources used to produce healthcare in a hospital a decade ago is not the same as those used to produce health services in the present day. With respect to nursing personnel, the roles and productivity of one type of nurse relative to another (e.g., an LPN versus APRN) have changed markedly over the years due to modifications in state practice acts, innovations in nursing education, changes in institutional policies, emergence of evidence-based practice, collective bargaining agreements that have expanded or restricted the performance of tasks by different types of personnel, and by efforts to mandate patient-to-nurse staffing ratios.

In sum, given the forces affecting society's demand for healthcare and assuming that enough RNs are available and willing to work at the wages and working conditions offered by employers, most healthcare organizations seek to employ the number and mix of RNs and other nursing personnel that can most efficiently produce the treatments and services consistent with the organization's objectives, budget, quality standards, and the ways that other healthcare personnel, capital, and technology can be productively combined.

Forecasts of Registered Nurse Demand

The Bureau of Labor Statistics (BLS) and HRSA have estimated the societal and organizational factors affecting the demand for RNs and both indicate increasing demand for RNs over the near-term future. Based on industry surveys, the BLS estimates overall job opportunities for RNs will increase by 22% from 2008 to 2018, a rate of growth that is much faster than the average of all occupations (averaging between 7% and 13% over the same time period). According to the BLS, growth will be driven by technological advances in patient care, an increase in preventative care and growth in the population of older citizens. Further, the BLS expects that employment growth in hospitals will be slower (17%) than in nursing care facilities (25%), home healthcare services (33%), and offices of physicians (48%).

In 2004, HRSA projected that the future requirement of RNs through 2020 would increase by more than 800 000 FTE

RNs more than 2000 levels. These projections, however, were made before the passage of health reform legislation in 2010, and thus demand is likely to exceed these projections as an estimated 32 million Americans gain greater access to health insurance coverage during the decade.

Factors Affecting the Supply of Registered Nurses

When thinking about the supply of RNs, it is useful to differentiate between the short- and long-run supply of RNs. The short-run supply of RNs refers to the decisions of existing RNs to participate in the labor market and number of hours to spend working. If, for example, we are interested in increasing the supply of RNs to help resolve a nursing shortage, then any increase in RN supply will come initially from stimulating the number of currently available RNs to participate in the workforce or, if they are already working, to increase the number of hours they are willing to work (or both).

Changing the short-run supply of RNs can be accomplished by manipulating factors that existing RNs consider when deciding whether to participate in the nurse labor market and the number of hours they are willing to work. In contrast, because it takes between 2 and 4 years for an individual to complete a basic nursing education program, the long-run supply of RNs refers to the number of RNs that will be available at some point in time in the future. Thus, an expansion in the long-run supply of RNs will not address a shortage of RNs that is being experienced in the present day but may help resolve a future shortage.

Short-Run Supply

RNs' participation and hours decisions are determined by economic and noneconomic factors. Economic factors include the RN's wage (and fringe benefits) and nonwage income (primarily the income of the RN's spouse). In economics, when wages change, both substitution and income effects are elicited. However, because the substitution and income effects exert opposite effects on labor supply decisions, whichever effect dominates will determine RNs employment decisions, holding the effects of other economic and noneconomic factors constant. Many studies of RNs' short-run labor supply show that, on average, increases in wages tend to exert a positive but relatively small impact on the number of hours worked by RNs and a greater impact on the decision of nonparticipants to rejoin the workforce (Sheilds, 2004). With regard to the impact of nonwage income on RNs' labor supply decisions, evidence from labor supply studies indicates that increases in nonwage income exert a negative and substantial impact on participation and hours worked, holding all else constant. Because a majority of RNs are married women, a spouse's income is a significant source of income for many RN households. The effect of spouse income is related to the observed counter-cyclical effect of RN labor supply and the economy as a whole. For example, Buerhaus *et al.* (2009) showed that RN employment tends to grow much faster during and immediately after recessions, with much of employment growth linked to older RNs rejoining the workforce or working more hours. Quantitatively, the authors calculate that a percentage point increase in the unemployment rate is associated with a 1% increase in RN labor supply.

Noneconomic factors also influence RNs' labor market decisions. These factors include: the presence of children (approximately 70% of RNs are married, and studies show that young children at home exert a substantial demand on the RN's time and hence a negative effect on RN participation and hours worked); older adults living in the RNs' household (few studies have examined the impact on the RNs' labor supply decisions, although studies of women in the overall workforce indicate that caring for older adults decreases participation and hours worked substantially); enrollment in education programs (many RNs are obtaining their bachelor's or master's degree and thus have less time available to work in the labor market), demographic characteristics such as age (older RNs work more hours), race (nonwhite RNs have higher participation rates and work more hours) and gender (most studies show that men have higher participation rates and work more hours than women). However, because it is difficult to change the noneconomic factors that affect RNs' decision to work, employers rely on changing wages and fringe benefits to influence the short-run labor supply decisions of existing RNs.

Long-Run Supply of Registered Nurses

In contrast to the short-run supply of RNs that involves the labor market decisions of existing RNs, the long-run supply of RNs concerns the total number of RNs, who will be available in the future. A key factor affecting the long-run supply of RNs is the number of women in the US population between the ages of 20 and 40 years that make up the largest pool of individuals from which nursing education programs draw applicants. As large numbers of women born during the baby boom generation (1946–64) entered their twenties, the size of the pool of women increased in the late 1960s and continued expanding for the next 20 years. Consequently, these pools 'produced' large numbers of RNs. Since 1985, however, the size of the population pool 20–40 has remained relatively stagnant and is projected to change very little over the next 10–15 years.

RN nursing students are drawn into nursing for a variety of personal interests and motivations. However, the growth in new career options for women in the 1980s and 1990s led to a declining propensity of women choosing a nursing career (at the same time that the size of applicant pools were no longer increasing). More recently, enrollments have expanded, suggesting that interest in becoming an RN has increased. Internationally educated RNs, who join the US nursing workforce, have also helped expand the long-run supply of RNs; since the mid-1990s, IENs have been increasing both in number and as a proportion of the nursing workforce in the US.

Economic factors such as tuition, time costs, and prospective earnings also influence the long-run supply of RNs. For some people, the tuition charged by nursing education programs relative to the tuition required by other careers the individual is considering is an important factor in making the decision to become an RN. The less time it takes for an

individual to recoup their investment in a nursing education, given his or her particular skills, the more likely they will become an RN. For individuals who are on the brink of deciding to choose nursing or a different career, RN wages in the nurse labor market can influence their decision, especially if wages are increasing and the individual is aware of the improving economic prospects in nursing.

The capacity of the nursing education system should respond to demand for RNs and interest in becoming an RN in the population – however, that response may be uneven due to institutional or other constraints. Shortages of faculty have been reported since the early 2000s – an oft-cited reason for thousands of qualified applicants being turned away from nursing education programs each year since 2002. That constraint seems to have eased recently, as the past few years have seen strong growth in nursing programs, new graduates and RNs entering the workforce.

Projections of the Long-Run Supply of Registered Nurses

Projections made by the authors in 2000 and HRSA in 2004 suggested that the number of RNs would grow slowly through the current decade, level off for several years, and then decline by 2020 as RNs retire from the workforce. Subsequent projections (Auerbach *et al.*, 2007; Buerhaus *et al.*, 2009) revealed that the future supply of RNs was beginning to grow in response to national initiatives to attract people into nursing. These initiatives appear to have had their desired effect on increasing enrollment into nursing education programs by both young people graduating from high school and by those in their 30s deciding to leave their nonnursing occupation and become an RN.

Registered Nurse Shortages

Perhaps no other topic related to the nursing workforce has dominated the attention of federal and state legislators, workforce planners, nursing organizations, and the media more than hospital shortages of nurses. Shortages have occurred frequently in the US and affect hospitals' (and other care delivery organizations) ability to operate safely and provide access to healthcare. From an economic perspective, a shortage of hospital RNs reflects market disequilibrium in which hospitals' demand for RNs exceeds the existing supply of RNs at the prevailing wage (including nonwage benefits). Thus, a shortage is a market disequilibrium in which labor demanded by hospitals exceeds labor supplied by RNs because the wage lies below the equilibrium wage – the wage level in which demand and supply are in balance. The shortage will not begin to disappear until wages increase to a level that brings about an increase in RNs' short-run labor supply (an increase in participation or hours worked, or both) that satisfies a hospital's demand. If, however, the hospital's demand for nurses continues to expand at the same time that wages are rising, the shortage of RNs will persist.

Figure 7 shows the supply and demand for hospital RNs, with the labor supply curve upward sloping, whereas the labor

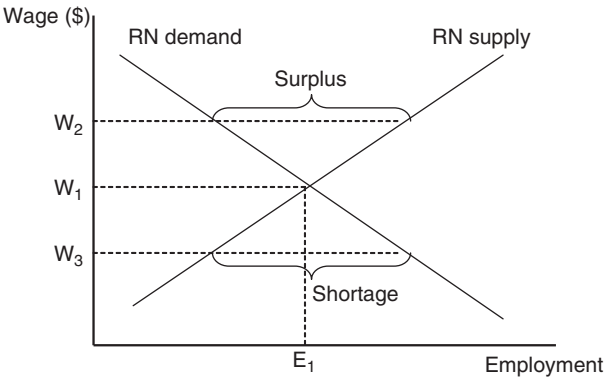


Figure 7 Equilibrium hours and employment in a competitive labor market.

demand curve is downward sloping. At the point where demand and supply intersect, the wage level (W_1) is such that the supply of RN labor is exactly equal to the demand for RN labor at employment level E_1. At any higher wage level, such as W_2, there will be a surplus of RNs seeking jobs because the higher wage increases supply while at the same time reduces hospitals' demand for RNs. In Figure 7, the surplus is reflected in the horizontal distance between the supply and demand curves at wage level W_2. Competition among the surplus RNs to obtain the limited number of hospital jobs will eventually place downward pressure on wages, decreasing wages toward W_1 until the supply and demand intersect W_1, the equilibrium wage.

Similarly, at any wage level below W_1, such as W_3, there will be a shortage of RNs as the lower wage decreases the supply of RNs and increases employers' demand for RNs. The shortage is shown in Figure 7 as the horizontal distance between the demand and supply curves at wage level W_3. Competition among employers to obtain the limited number of RNs will exert upward pressure on wages, pushing wages again back toward W_1. Thus, the point at which labor supply and demand intersect determines the unique equilibrium combination of wage and employment levels (W_1 and E_1) that will result in an equilibrium market. During a shortage, competition among employers to obtain the limited number of RNs will put upward pressure on wages, pushing wages back to their equilibrium level. Thus, shortages should not exist for extended periods, at least in a competitive labor market or in the absence of restrictions that prevent wages from increasing.

If hospitals' demand for RNs increases, the hospitals may find that there are not enough RNs willing to supply their services at the wages they are offering and a new shortage will develop. The development of the shortage is shown in Figure 8 and focuses initially on the long-run equilibrium wage, W_1, at which the short- and long-run supply of RNs are equal to the demand for RNs. In the short run, the outward shift in RN demand from D_1 to D_2 results in a shortage of RNs. The shortage develops because not enough RNs are willing to supply their time to hospitals at this prevailing wage rate, W_1. However, as soon as RN wages increase and reach a new equilibrium at the point where the new labor demand curve (D_2) crosses the short-run labor supply curve, the shortage will disappear. This movement along the short-run labor supply curve results in much higher wages and somewhat higher

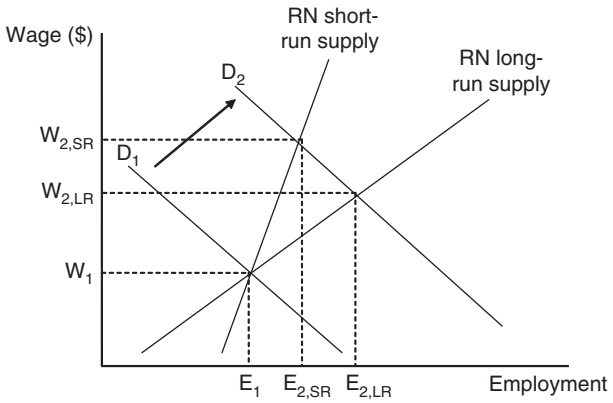


Figure 8 Impact of outward shift in labor demand on short- and long-run competitive equilibrium.

employment in the short run (increased participation and hours worked by existing RNs) from E_1 to $E_{2,SR}$. In other words, the increase in RN wages will first stimulate some existing RNs to respond in the short run by rejoining the workforce, moving from nonhospital settings into hospitals or by working additional hours (switching from a part- to full-time basis, working overtime or even working a second job). Eventually, in the long run new individuals will choose to become RNs drawn to nursing by the wage increase, thus multiplying the effect of the initial short-run response to the wage increase. Thus, over time, the new equilibrium will move to the point where the new labor demand curve (D_2) crosses the long-run labor supply curve at $W_{2,LR}$, in Figure 8. Thus, outward shifts in the labor demand curve tend to primarily increase wages in the short run, whereas having more of an impact on employment (and less on wages) in the longer run.

As noted above, shortages of RNs tend to be transitory and corrected by increases in the wage rate unless hospitals do not increase RN wages or for some reason are blocked from raising them. Unless demand is continuing to expand at the same time, the increased long-run supply of RNs will, in turn, exert downward pressure on wages. As the wage rate decreases, employers will be willing to hire additional RNs until they reach the point depicted by $E_{2,LR}$ and the labor market for RNs will once again adjust to a new long-run equilibrium wage and employment level of RNs.

If a hospital's demand for RNs should increase again, the hospital may find that there are not enough RNs willing to supply their services at the prevailing wage it is offering and consequently a new shortage will develop. The series of short- and long-run adjustments will begin a new, and eventually a new market wage will be reached where the long-run demand and supply of RNs are in balance and employment levels are higher. The repetition of this cycle of demand, supply, and wage adjustments is often referred to as the 'cyclical' shortage of nurses.

The speed with which shortages are resolved depends on several factors including: how high and quickly a hospital increases RN wages; how sensitive RNs are to the wage increase (the RN wage elasticity of supply); and how sensitive is hospitals' demand for RNs over the range of wage increases they are considering offering RNs (the employers' wage elasticity of demand for RNs).

History of Hospital Registered Nurse Shortages

Surprisingly, although hospitals have reported shortages of RNs in every decade since the 1960s, there is no agreement about how to define and measure shortages. One common indicator of shortages is the job vacancy rates reported by hospitals (the percentage of vacant FTE RN positions that hospitals are actively trying to fill). In general, reports of hospital shortages typically occur when FTE RN vacancy rates exceed 4%. Most, but not all, reported hospital RN shortages since the mid-1960s were driven by increases in the demand for RNs and were resolved after hospitals' increased real wages.

For several years before the creation of the Medicare and Medicaid programs in 1965, hospital RN vacancy rates were very high, exceeding 15%. The new financial resources provided by the Medicare program enabled hospitals to increase RN wages, which subsequently brought about an end to the shortage by the end of the decade. During the 1970s, demand for RNs continued to grow, but wage controls imposed by the federal government via the Nixon administration's Economic Stabilization Program combined with high inflation rates restricted the increase in RN wages from rising fast enough to bring hospitals' demand and supply of RNs into equilibrium. Consequently, hospitals reported double-digit RN vacancy rates and shortages of RNs during most of the 1970s.

Following large increases in real wages in the early 1980s, hospital RN vacancy rates began decreasing and the shortage ended quickly. However, with the beginning of the Medicare prospective payment system in 1983, hospitals faced new incentives to become more efficient and, among other adjustments, shifted less acutely ill patients to lower cost outpatient departments. Because RNs are more productive than LPNs at the prevailing wages, hospitals' demand for RNs increased. Throughout the decade, hospitals' demand for RNs increased approximately 3% annually and despite increases in real wages, the supply of RNs was unable to catch up to the significant growth in demand, resulting in hospitals once again reporting RN shortages during much of the latter part of the 1980s.

During the 1990s, the growth in the hospitals' demand for RNs slowed to approximately 2% per year as managed care developed rapidly. Both RN hospital vacancy rates and real wages soon decreased, bringing the shortage of RNs to an end. The increased number of RNs joining the RN workforce that were being supplied by nursing education programs resulted in an apparent surplus of RNs during the mid-1990s and real earnings and vacancy rates both decreased. However, by the end of the decade, yet another RN shortage was reported by hospitals but, in this case, the shortage was concentrated in intensive care units (ICUs) and operating rooms.

When data on RNs were analyzed by age categories, hospital unit, and educational background, it was discovered that the shortage broke out in these units because they were the first to experience the implications of underlying changes in the age and education composition of the RN workforce that resulted in a decrease supply of certain types of RNs. Shortages reported by ICUs and stepdown units resulted from a decrease in the number of younger RNs entering the workforce who have a greater propensity than older RNs to work in critical care units. Shortages in operating rooms resulted from the

decline in the number of older diploma-educated graduates who had a greater propensity to work in this setting, as they began to retire from the workforce. This analysis demonstrated that unlike the earlier shortages that were driven by increases in demand, the shortage that developed in 1998 was driven by supply-side factors reflected by the changing age composition of the RN workforce.

Impact of Recent Recessions on Hospital Registered Nurse Shortages

By the early 2000s, the hospital RN shortage spread throughout other nursing units as demand for hospital care began to increase. In 2001, hospital FTE RN vacancy rates exceeded double-digit levels. However, during 2001 a recession developed, and though it lasted only 8 months, unemployment rates remained high over the next 2 years (Table 3). Because the majority of RNs are married, increases in overall unemployment meant that many RN spouses either lost their job or feared that they could be laid off. To ensure the economic welfare of their households, many RNs who were not working at the time, particularly married RNs, rejoined the nursing workforce. During 2002 and 2003, hospital RN employment increased dramatically, shooting up by an estimated 185 000 FTEs. This burst in RN employment decreased the impact of the nursing shortage and reduced vacancy rates to approximately 8% by the end of 2006. As the shortage continued into 2007, it became the longest lasting shortage of hospital RNs in the past 50 years.

The second recession of the past decade began in December 2007 and lasted through June 2009, even though monthly unemployment rates continued to increase before peaking at 10.1% in October. Once again, hospital RN employment increased as nonparticipating RNs rejoined the hospital workforce and other RNs, who were already working increased their work hours. During 2007 and 2008, hospital RN employment surged, adding nearly one-quarter million FTE RNs. Moreover, more than 100 000 of this increased employment occurred among RNs older than 50 years of age, suggesting that some RNs who had retired rejoined the workforce. Another 50 000 RNs left their positions in nonhospital settings for higher paying hospital jobs, which also offered richer benefits (particularly health insurance coverage) and flexible work hours. Although national estimates of hospital RN vacancy rates are not available to assess the effect of this employment increase, anecdotal reports suggest that the national shortage of RNs that had begun a decade earlier in 1998 had finally come to an end for many hospitals (Auerbach *et al.*, 2011).

The Future of the Registered Nurse Workforce

Over the decade, the RN workforce will be dominated by older RNs in their 50s. Because the number of RNs in their 50s is so large (approximately 900 000), it will be very difficult to replace these RNs with new entrants, and thus new shortages are expected. These shortages could be larger than previous shortages of RNs experienced since the 1960s and could take an extended period of time for the labor market to adjust and establish a new equilibrium in which the shortage disappears.

Currently, the looming threat of large retirements from the workforce is masked by the lingering effects of the recent recession on the labor supply decisions of the current workforce. Average monthly unemployment rates remain above 8% and appear to be continuing to stimulate record high participation in the labor market by the existing stock of RNs. Reports of new graduates of nursing education programs having substantial difficulty finding jobs suggest that the hospital labor market may be in equilibrium. However, once the economy strengthens and there is a strong jobs recovery, many currently employed RNs, particularly older RNs, may retire from the workforce. If large numbers of RNs exit and if they withdraw from the workforce rapidly, then a new shortage is likely to develop. However, the stock of new graduates waiting for new jobs to develop may be large enough to enter the labor market and replace those exiting and thereby decrease the risk of new shortages developing.

Beyond these uncertainties, the nursing profession faces other challenges, particularly in an era of health reform. Many of these challenges are described in a recent report by the Institute of Medicine (IOM), The Future of Nursing: Leading Change, Advancing Health. The IOM report offered four key messages and eight recommendations aimed at strengthening the nursing workforce (Table 4). Several of these were aimed at strengthening patient centered, high quality, coordinated, primary care that is expected to be in great demand as the number of the insured grow while physicians increasingly move toward specialization. Although it is beyond the scope of this article to discuss these messages and recommendations

Table 3 Changes in national unemployment rates and full-time equivalent (FTE) registered nurse (RN) employment in the US, 2001–10

Year	*National unemployment rate (%)*	*FTE RN employment and change from prior year hospitals*	*FTE RN employment and change from prior year nonhospital*
2001	4.7	1 201 003	786 387
2002	5.8	1 285 718 (84 715)	787 564 (1177)
2003	6.0	1 384 482 (98 764)	807 498 (19 934)
2004	5.5	1 378 116 (– 6366)	813 317 (5819)
2005	5.1	1 326 914 (– 51 202)	837 269 (23 952)
2006	4.6	1 345 711 (18 797)	894 162 (56 893)
2007	4.6	1 429 989 (84 278)	923 165 (29 003)
2008	5.8	1 588 226 (158 237)	875 260 (– 47 905)
2009	9.3	1 569 496 (– 18 730)	932 406 (57 146)
2010	9.6	1 608 453 (38 957)	926 907 (5499)

in detail, one recommendation calling for removing barriers restricting NPs' scope of practice has received considerable attention from the media, health policy makers, and many in the medical profession. Therefore, the authors conclude the discussion of the professional nursing workforce by providing a brief overview of APRNs.

The Advanced Practice Nurse Workforce

The term APRN refers to four types of nurses who have received advanced education and training beyond that required to become an RN and include NPs, certified RN anesthetists (CRNAs), CNS's, and CRNMs. Most states require APRNs to complete a master's degree in nursing, and the vast majority of state legislature have delegated to state boards of nursing the authority to establish requirements for certification examinations in each type of APRN and in the various advanced practice subspecialties. In a few cases, the state board of medicine holds this authority (Cuningham, 2010). Data from the NSSRN indicate that approximately 8% of RNs or 220 000 of 2.6 million RNs employed in nursing in 2008 were APRNs. Below, the authors briefly describe the roles and key characteristics of each type of APRN (see Table 5).

Table 4 Key messages and recommendations

Key messages

1. Nurses should practice to the full extent of their education and training
2. Nurses should achieve higher levels of education and training through an improved education system that promotes seamless academic progression
3. Nurses should be full partners, with physicians and other health professionals, in redesigning healthcare in the US
4. Effective workforce planning and policy making require better data collection and an improved information infrastructure

Recommendations

1. Remove scope-of-practice barriers
2. Expand opportunities for nurses to lead and diffuse collaborative improvement efforts
3. Implement nurse residency programs
4. Increase the proportion of nurses with a baccalaureate degree to 80% by 2020
5. Double the number of nurses with a doctorate by 2020
6. Ensure that nurses engage in lifelong learning
7. Prepare and enable nurses to lead change to advance health
8. Build an infrastructure for the collection and analysis of interprofessional healthcare workforce data

Source: Reproduced from Institute of Medicine (2011). *The future of nursing: leading change, advancing health*. Washington, DC: The National Academies Press.

Nurse Practitioners

NPs are the largest and most rapidly growing APRN. According to the HRSA (2010), an estimated 130 000 NPs were working in nursing in 2008, double the number estimated a decade earlier. The role of NPs was established in the mid-1960s and focused on serving women and children in rural and underserved inner city areas where physicians were scarce. In the present day, NPs work across many populations and geographic regions and their focus have expanded to include family care, pediatrics, geriatrics, adult health, women's health, psychiatry, neonatology, and acute hospital care of adults and children. Currently, 39% of NPs work in hospital settings, particularly in specialized inpatient care units and hospital-affiliated primary care clinics, approximately 36% provide primary care in traditional ambulatory care settings (including retail clinics), and 12% work in public and community healthcare agencies and in schools. In 2008, full-time NPs earned $83 000 on average in 2008, compared to $67 000 for all RNs HRSA (2010).

Clinical Nurse Specialists

Although the number of CNS's has declined recently, CNS's are the second-largest type of APRN and are estimated to number approximately 45 000 in 2008. The role of a CNS is to improve clinical care, primarily in hospitals and extended care facilities, by providing advanced clinical nursing expertise to help coordinate care for individuals, educate nursing personnel who provided direct care, and help identify and improve aspects of the health system organization that affect patients and nursing staff. CNS's have expertise in one or more clinical areas such as oncology, pediatrics, geriatrics, psychiatric/mental health, adult health, obstetrics, acute/critical care, and community health. Although about half of CNS's are employed in hospital settings, often in administrative or supervisory roles, CNS's also work in ambulatory care, public health, and academic settings. Nearly 64% are over the age of 50 years, making them the oldest group of APRNs. Average earnings for a full-time CNS were $86 000 in 2008.

Table 5 Number and characteristics of advanced practice registered nurses, 2008

	Nurse practitioners	*Certified nurse midwives*	*Certified nurse anesthetists*	*Clinical nurse specialists*
Number employed in nursing, 2008	132 000	15 000	30 000	45 000
Number employed in nursing, 2004	118 000	11 000	28 000	57 000
% More than age 50 years	50%	54%	50%	64%
Average salary (full time)	$83 000	$75 000	$135 000	$84 000
Physician type with most overlap	Family or general practitioner	Obstetrician/gynecologist	Anesthesiologist	N/A
Main setting of work	Primary care settings	Hospitals	Hospitals	Hospitals

Source: Reproduced from Health Resources and Services Administration (HRSA) (2010). *The registered nurse population: Findings from the 2008 National Sample Survey of Registered Nurses*. Rockville, MD: HRSA.

Certified Registered Nurse Anesthetists

Nurses have been providing anesthesia since the Civil War and in the present day provide approximately 32 million anesthetics annually in the US and represent two-thirds of anesthetists in rural hospitals (American Association of Nurse Anesthetists, 2011). Most (82%) CRNAs not only work in a hospital operating room, but they also deliver anesthetics in birthing centers/obstetrics departments, dental offices, emergency rooms, plastic surgery centers, and outpatient surgery facilities. CRNAs play a particularly important role providing anesthesia in the military and the Veterans Administration and in hospitals located in rural areas. There were roughly 30 000 CRNAs working in 2008, a 16% increase from 2000. On average, CRNAs are younger and more likely to be male (more than 40%) than other APRNs. In 2008, CRNAs reported average earnings of $136 000, which is much higher than all other groups of RNs or other APRNs HRSA (2010). Unlike anesthesiologists, CRNAs are more likely to work in rural rather than urban areas.

Certified Nurse Midwives

CNMs care for women before, during, and after childbirth. Their role began in the nineteenth century to fill a particular need in impoverished urban and rural areas with limited access to physicians. Nurse midwifery arose both in New York City and Kentucky in the late nineteenth and early twentieth centuries. The earliest US nurse midwifery programs were designed to meet the needs of special populations in urban, rural, and impoverished populations with limited access to physicians HRSA (2010). Most CNMs work in hospitals, with 42% specializing in labor and delivery, 34% in obstetrics, and 14% in gynecology or women's health. One-fourth worked in ambulatory care settings in 2008. There were approximately 15 000 CNMs employed in nursing in 2008, making them the smallest group of APRNs, although their numbers have grown since 2004. CNMs earned $75 000 on average in 2008, and 55% were more than age 50 years.

Overlap with Physicians

CNMs, CRNAs, and particularly NPs perform roles throughout the healthcare delivery system and provide many services that overlap considerably with those of physicians (respectively, obstetricians/gynecologists, anesthesiologists, and physicians proving primary care such as internists, pediatricians, and family practitioners). That overlap has several implications. First, areas of the country that have difficulty attracting physicians (particularly rural or inner city areas) have relied on APRNs to fill workforce gaps. Consequently, APRNs are more likely to work in rural and inner city areas and serve patients that are less likely to have private insurance. Second, the degree to which APRNs can substitute for physicians has resulted in a growing literature and policy debate about whether APRNs provide care of comparable quality to their physician counterparts. Most studies find that NPs can successfully handle up to 80% of primary care visits, and that the care received by patients seeing either an NP or a primary care physician is comparable in terms of quality or resource use (Newhouse *et al.*, 2011). Some studies have employed randomized clinical trials, for example, assigning patients to either NPs or primary care physicians (Laurant *et al.*, 2004). In light of projections of shortages of primary care physicians by 2020 (Association of American Medical Colleges, 2011) and because provisions in the Patient Protection and Affordable Care Act will expand the insured population and demand for primary care (Ku *et al.*, 2011), the demand for NPs is likely to grow as will the controversy surrounding policies that call for expanding the NP workforce to make up the gap in primary care (Naylor and Kurtzman, 2010; Pohl *et al.*, 2010). Similar debates over quality and practice restrictions also involve CRNAs (Dulisse and Cromwell, 2010).

All states regulate the boundaries of practice governing the services each type of APRN is permitted to perform. Particularly in the case of NPs, 'scope of practice' laws regulate aspects of practice such as the required level of physician supervision and collaboration and the ability to prescribe medications. Critics of these laws assert that they reduce access and increase the cost of care by forcing patients to seek care from physicians who typically charge higher prices than NPs. Defenders argue that they are necessary to protect patients from low-quality care and unsafe practice. Currently, such laws vary widely from state to state and this variation is viewed by some as hampering reimbursement by private insurers to NPs in certain states (Tine-Hanson-Turton *et al.*, 2006).

An analysis of state scope of practice laws that governed NPs and CNMs from 1992 to 2000 suggested that the state laws were becoming less restrictive during this period (HRSA, 2000). Considerable variation and restrictions remain, however, as illustrated in Missouri where new patients who had an initial visit with an NP had to be seen by a physician within 2 weeks or in South Carolina that requires a supervising physician be available at all times for consultation. Some states only permit NPs to prescribe certain medications or to refer patients for laboratory tests on an approved list (Lugo *et al.*, 2007; HRSA, 2000). As of 2010, 16 states allowed NPs to practice independently from a physician (Fairman *et al.*, 2011).

Scope of practice issues for CNMs are similar to those of NPs, with states varying on the degree of prescriptive authority, supervision by physicians, and the extent to which CNMs can be reimbursed directly (HRSA, 2000). Yet, despite extensive state variation in scope of practice for NPs and CNMs, there is little research or evidence as to the effects of the laws on processes and outcomes of care.

Summary

Much of the production and distribution of personal healthcare services in the US depends on professional nurses, RNs and APRNs, who are employed in wide variety of clinical and nonclinical positions in countless organizations. Professional nurses deliver basic and advanced nursing care services, practice nursing independently, and function as both complements and substitutes to physicians. The RN workforce is dominated by the large and aging baby boom cohorts who are expected to retire in large numbers over the decade, threatening to create a new nurse shortage in hospitals and other

settings, particularly as the demand for healthcare expands due to the implementation of health reform, the aging of the baby boom generation, and other factors. How long it will take for a new equilibrium to be reached in the nurse labor market will depend on how many RNs retire, whether there will be enough new RNs to replace them, how much demand grows, and how effectively organizations adjust to these changes. Increasing demand for healthcare also affects physicians, particularly primary care physicians whose supply is projected to fall below the estimated demand by the end of the decade. Because NPs are viewed as being good substitutes for the majority of primary care services provided by physicians, healthcare policy makers are focusing on efforts to allow NPs and other APRNs to practice to the full extent of their education and training by reforming restrictive state laws and nurse practice acts.

See also: Aging: Health at Advanced Ages. Competition on the Hospital Sector. Health and Health Care, Need for. Health Care Demand, Empirical Determinants of. Home Health Services, Economics of. Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity. Long-Term Care. Managed Care. Monopsony in Health Labor Markets. Nurses' Unions. Occupational Licensing in Health Care. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Primary Care, Gatekeeping, and Incentives. Public Health Profession

References

American Association of Nurse Anesthetists (2011). About AANA http://www.aana.com/aboutaana.aspxid=46 (accessed 09.05.11).

Association of American Medical Colleges (2011). Physician shortages to worsen without increases in residency training. Available at: https://www.aamc.org/download/150584/data/physician_shortages_factsheet.pdf (accessed 08.05.11).

Auerbach, D., Buerhaus, P. and Staiger, D. (2007). Better late than never: Workforce supply implication of later entry into nursing. *Health Affairs* **26**(1), 178–185.

Auerbach, D., Buerhaus, P. I. and Staiger, D. O. (2011). Registered nurse supply grows faster than projected amid surge in new entrants ages 23–26. *Health Affairs* **30**(12).

Auerbach, D., Buerhaus, P. I. and Staiger, D. O. (2011). Registered nurse supply grows faster than projected amid surge in new entrants ages 23–26. *Health Affairs* **30**(12), doi: 10.1377/hlthaff.2011.0588.

Buerhaus, P., Auerbach, D. and Staiger, D. (2009). The recent surge in nurse employment: Causes and implications. *Health Affairs* **28**(4), w657–w668, Web Exclusive, June 12.

Cuningham, R. (2010). Tapping the potential of the health care workforce: Scope-of-practice and payment policies for advanced practice nurses and physician assistants. National Health Policy Forum, background paper no. 76, The George Washington University, Washington, DC.

Dulisse, B. and Cromwell, J. (2010). No harm found when nurse anesthetists work without supervision by physicians. *Health Affairs* **29**, 1469–1475.

Fairman, J., Rowe, J., Hassmiller, S. and Shalala, D. (2011). Broadening the scope of nursing practice. *New England Journal of Medicine* **364**(3), 193–196.

Health Resources and Services Administration (HRSA) (2010). *The registered nurse population: Findings from the 2008 National Sample Survey of Registered Nurses*. Rockville, MD: HRSA.

Ku, L., Jones, K., Shin, P., Bruen, B. and Hayes, K. (2011). The states next challenge – Securing primary care for expanded Medicaid populations. *New England Journal of Medicine* **364**(6), 493–495.

Laurant, M., Reeves, D., Hermens, R., et al. (2004). Substitution of doctors by nurses in primary care. Cochrane Database of Systematic Reviews, Issue 4 Art. No.: CD001271. DOI: 10.1002/14651858.CD001271.pub2.

Lugo, N., O'Grady, I., Hodnicki, D. and Hanson, C. (2007). Ranking state NP regulation: Practice environment and consumer health care choice. *American Journal for Nurse Practitioners* **11**(4), 8–24.

Naylor, M. and Kurtzman, E. (2010). The role of nurse practitioners in reinventing primary care. *Health Affair* **29**(5), 893–899.

Newhouse, R. P., Stanik-Hutt, J., White, K. M., et al. (2011). Advanced practice nurse outcomes 1990–2008: A systematic review. *Nursing Economic$* **29**(5), 1.

Pohl, J., Hanson, C., Newland, J. and Cronenwitt, L. (2010). Unleashing nurse practitioners' potential to deliver primary care and lead teams. *Health Affairs* **29**(5), 900–905.

Sheilds, M. (2004). Addressing nurse shortages: What can policy makers learn from the econometric evidence on nurse labour supply? *Economic Journal* **114**, F464–F498.

Sochalski, J. and Weinder J. (2011). Health care system reform and the nursing workforce: Matching nursing practice and skills to future needs, not past demands. Appendix F: *The Future of Nursing: Leading Change, Advancing Health*. Washington, DC: The Institute of Medicine.

Staiger, D., Auerbach, D. and Buerhaus, P. (2009). Comparison of physician workforce estimates and supply projections. *Journal of the American Medical Association* **302**(15), 1674–1680.

Tine-Hanson-Turton, T., Ritter, A., Rothman, N. and Valdez, B. (2006). Insurance barriers create barriers to health care access and consumer choice. *Nursing Economic$* **24**(4), 204–211.

US Department of Health and Human Services, Health Resources and Services Administration (2000). A comparison of changes in the professional practice of nurse practitioners, physician assistants, and certified nurse midwives: 1992 and 2000. This study was funded by the National Center for Health Workforce Analysis Bureau of Health Professions Health Resources and Services Administration under Contract No. HRSA 230-00-0099.

Further Reading

Buerhaus, P., Staiger, D. and Auerbach, D. (2008). *The Nursing Workforce in the United States: Data, Trends, & Implications*. Boston, MA: Jones-Bartlett, Inc.

United States Department of Labor, Bureau of Labor Statistics (2011). Registered nurses. Occupational Outlook Handbook. 2010–11 ed. Available at: http://www.bls.gov/oco/ocos083.htm (accessed 06.05).

US Department of Health and Human Services, Health Resources and Services Administration (2004). What is behind HRSA's supply, demand and projected shortage of registered nurses?

US Department of Labor (2011). Labor force statistics from the Current Population Survey. Available at: http://www.bls.gov/cps/home.htm (accessed 06.05.11).

Markets in Health Care

P Pita Barros, Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal
P Olivella, Universitat Autonoma de Barcelona and Barcelona GSE, Cerdanyola del Valles (Barcelona), Spain

Glossary

Capitation A payment from a third-party payer such as a health authority or an insurer to a supplier such as a health plan or a health-service provider that is made per enrollee (in the case of health plans) or per individual in the population residing in the catchment area (in the case of paying for specific healthcare services).

Complementary (or supplementary) private health insurance (PHI) Private health insurance cover for copayments in the public sector.

Copayment If both a patient and a third-party payer share the payment of some service, the part that the patient bears.

Duplicate PHI In the presence of a national health service that covers part or the whole population and a (large) portfolio of services, private insurers offer insurance covering a similar portfolio of services.

Health plan Usually an insurer that receives its premium from a third-party payer rather than from (or on top of) the individual's out of pocket premium.

National Health System or National Health Service (NHS) In an NHS, health insurance and healthcare services are integrated into a single health authority, which either owns its own network of final providers or subcontracts with (usually) nonprofit private hospitals. The whole system provides a large portfolio of services and is financed trough general taxation and limited copayments.

Premium The price of an insurance contract. Also used to describe a third party's payment to a health plan for each of its enrollees, in this case sometimes also referred to as capitation or capitation rate.

Supplementary PHI The Organization for Economic Co-operation and Development is proposing to relegate the term supplementary to PHI that covers services that are not covered by the national health system, some dentistry services being a usual example (see Complementary PHI).

Yardstick competition In health care, the use of comparative performance indicators, usually by a health authority, to design payment mechanisms for providers.

Introduction

The first question one should ask when addressing healthcare markets is whether health care is any different from other goods. If it is not, economic theory states that, as with apples and pears, the unfettered competitive market will lead to efficient outcomes and that equity can be reached by appropriately redistributing purchasing power *ex ante*. This brings the question of why in some countries healthcare markets do not even exist or are severely restricted. In national health services (NHSs) like those in the UK or Spain, a single door provides access to most health goods and services (pharmaceutical products usually being an exception), and market forces may disappear or be relegated to the stage where the health authority subcontracts with providers (doctors and hospitals). Less extremely, why is regulation of healthcare markets desirable?

Before answering these questions, let the reader be first warned that a broad interpretation of what a market is has been taken here. For instance, in the market for prepaid health plans, the individual may not pay a price either when enrolling a plan of her choice nor when she uses the services this plan provides. Even farther away from the usual idea of a market, in a NHS, the health authorities may base the remuneration of the hospitals they own or subcontract with on relative performance evaluations (RPEs). In this case, a hospital's revenue depends on how its performance compares to the average. For those who want to read more, their web search should include the terms 'yardstick competition' and 'contests.' Even an individual's choice between seeking treatment in his or her NHS or resort to a private hospital can be seen as a market. There, the 'price' in the public provider may be the time the individual has to wait. All the examples given have one thing in common: as a provider, your revenues fall if your performance (the combination of price and quality) falls as compared with your rivals or to any other existing outside option (e.g., an alternative treatment). The broad term 'market forces' is used to refer to this effect.

Now, the questions posited above can be readily answered. Health care is either publicly provided or its market severely regulated because society does not believe that free markets do such a good job. The differences between health and other goods, as well as the differences between health care and other services, can be traced back to the work of Arrow (1963). Since then, the role of markets in the allocation of resources in the health sector has been scrutinized from many angles. The role of ethics and societal judgments was, and is, widely discussed. The societal value of health care is not necessarily restricted to the standard notion of economic welfare (measured by the difference between an agent's willingness to pay for a service and the costs the agent bears, the so called 'agents' surplus'). Other considerations like happiness, freedom to choose, and absence of pain, for example, are often included as relevant for assessment of resources allocation, in addition to the mere utility from consumption of health and health care. Health in itself, by its nature, does not have a market for transaction. The role of markets in determining welfare (and the proper meaning of welfare) is distinct in the health sector. For an introduction to the discussion, see the essays contained in Cookson and Claxton (2012).

Still, markets are one mechanism that allocates resources in the health care in many ways. There are several reasons why

 doi:10.1016/B978-0-12-375678-7.01301-8

markets may behave differently in this sector, irrespective of the social judgment one makes about the adequateness of the resulting allocation. The focus is on the particular features of the market mechanism (without discussing here the welfare reasoning associated with each of those features) and the resulting reasons for government intervention.

The first reason – and the one that is best understood – is that some healthcare goods have external effects. That is, their consumption affects (positively or negatively) other individuals. The first example that comes into mind is vaccines. It is well known that markets for vaccines will not work well because individuals often ignore their beneficial external effect on everybody else's well-being. A free market will lead to under vaccination. Note that this does not directly apply to a hip replacement. One could say that taking care at home of a family member who can hardly walk puts lots of strain on the other family members. However, this is a type of externality that is likely to be internalized by the individuals themselves, because they probably care as much about the well-being of their family as for themselves.

The second reason – also well understood for a long time – is that providers could hold some degree of the so-called 'market power.' The idea is that if very few providers are available then they will be able to restrict supply in a way that price will be too high as compared to its optimum value: the cost of providing the unit that is least valued by society. This brings two effects. One is on equity, as consumers are worse-off whereas suppliers are better-off. The other is on efficiency; too few units are enjoyed by society. Why are healthcare goods prone to such a situation? The obvious reason is the existence of patents in some areas, which restrict the number of providers to one. Now patents are there for other reason that would take it too far. The other is that there exist some treatments that imply extremely large fixed costs. It does not make any sense to have two CT-scan machines in a small village. Such issues are taken up in Chapter 9.14 Health-Insurer Market Power: Theory and Evidence (00914). The existence of insurance coverage by making the patient (buyer) less sensitive to price induces the providers (sellers) to increase the price at the expense of the third-party payer.

The third reason involves the presence of privileged information (or asymmetric information) in one or both sides of the market. Health care is plagued with examples of this. On the consumer side, he or she may be more knowledgeable about his or her own health risks and needs (say intensity of pain). On the supplier side, healthcare goods are usually expert goods, meaning that providers are more informed of the true benefits (or long run side effects) of certain treatments. Again, markets perform quite poorly under these circumstances.

The fourth reason is the presence of risk. However, risk *per se* is not really problem for markets. It only implies that the notion of a good becomes a little more sophisticated. Given a single service, say a hip replacement, there are two or more goods, one for each possible circumstance that the individual may encounter. A hip replacement when the individual is perfectly healthy is not the same good as a hip replacement when the individual is under excruciating arthritis pain. The solution is to create an insurance market. In such a market firms (now called insurers) offer to exchange goods (in this case money and hip replacement) between these circumstances. The provider offers a hip replacement for free in case of severe arthritis pain (and only in that case) for free in exchange for x monetary units *ex ante*. One refers to x as the insurance premium. The market for insurance should perform well if none of the aforementioned problems is present. However, what does *ex ante* mean? What if an individual has privileged information of the likelihood of needing a hip replacement? Individuals who expect to need a hip replacement are more willing to pay the insurance premium. This leads to all sorts of problems. One of the solutions to these problems is quite drastic: get rid of insurance markets altogether and have a monolithic (vertically integrated) public-health system, perhaps financed with taxes or the like.

The fifth reason, intimately linked with the previous one, is the presence of moral hazard in either or both the consumer and the provider's side. On the demand side, the idea is that the individual, once insured, bears the consequences of a bad event less than fully and keeps a behavior that is detrimental to his/her health. Let this be illustrated with two very simple and very practical examples. If an insured individual becomes severely arthritic, he or she will obtain a hip replacement at a low price, perhaps even for free. Hence he or she will keep practicing vigorous sports. The same goes for stomach reduction and eating in a disorderly manner. These are of course rather extreme examples, but there seems to be some evidence that such behavioral decisions are present. Note that a real issue only exists if these behaviors cannot be contracted on because they are not publicly observable. Hence some researchers include moral hazard under the umbrella of asymmetric information, but they are not the same thing and require different cures. (See **Box 1** for a taxonomy of informational problems and market/government responses to them.) In the supply side, moral hazard refers to situations where the actual actions of the provider (be it a doctor's diagnose effort or care in treating a patient, be it the manager of a hospital who fails to contain costs) are unobservable. Again the idea is that the decision maker (doctor, manager, and nurse) may not bear the full impact of his or her actions. This is the case for example if the doctor receives a fixed monthly remuneration (independent of health outcomes) and medical errors go unpunished. It is worth mentioning here that the term moral hazard is also used to describe the phenomenon by which an individual who faces a low or zero price due to insurance may overuse the services. To distinguish both phenomena the literature (not unanimously) uses the terms *ex ante* moral hazard, which refers to the change of behavior that may lead to greater needs, and *ex post* moral hazard, which refers to the increase in service usage for a given need. It is easy to see that empirically identifying which is the source of higher usage is a complex question. On this, see Chapter 9.16 Moral Hazard (00916).

The last reason is the presence of the so-called bounded rationality. In this concept, misperceptions of one's own health status, lack of the mathematical capability to appraise extreme probabilities, intertemporally inconsistent behavior, or purely irrational behavior are included (perhaps too broadly). Since individuals are unable to make proper choices, the government (or a delegate of the government) makes them for them.

Box 1 A taxonomy of informational problems and market/government responses, with health-care examples

Informational problems

1. *Asymmetric information:* Some party in a relationship has privileged information on the environment.
   1.1. *Adverse selection:* The uninformed party is the first to make a decision or take some action, to which the informed agent responds. Synonym: *Hidden type.*
   1.2. *Signaling:* The informed party is the first to make a decision or take some action, to which the uninformed party responds.
2. *Imperfect information:* Once the relationship between two parties has already been established, i.e., through a contract, one of the parties takes an action that is unobservable to the other or which cannot be contracted on either for legal reasons or because a judge cannot enforce the contract. Synonyms: *moral hazard; unobservable action.*

Responses

0. General
   0.1. Uninformed party gathers more information
      0.1.1. Monitoring and auditing
      0.1.2. Prompt informed party to report his or her private information
      0.1.3. Prompt the informed party to show unfalsifable information
      Note: the two last procedures may lead to a signaling problem, since both reporting and providing information are subject to strategic/opportunistic behavior.
1. To adverse selection
   1.1. Offer a menu of contracts (screening)
   1.2. Public provision
2. To signaling
   2.1. Forbid information disclosure (difficult to implement)
   2.2. Public provision
3. To moral hazard
   3.1. Risk sharing: make undesired actions more costly
   3.2. Risk equalization and risk adjustment: so that all actions lead to the same benefits

Examples of adverse selection

Informed party	*Uninformed party*	*Information*	*Decision by uninformed party*	*Decision by informed party*	*References*
Potential insuree	Insurer	Current health status	Price and coverage to offer	Purchase insurance and from whom	Rothschild and Stiglitz (1976)
Doctor	Patient	True health benefits/long run side effects of some treatment	Prescribe A or B /refer to specialist or not	Accept/reject; change doctor	Garcia-Mariñoso and Jelovac (2003)
Manager of the hospital	Hospital board/owners/government	Risk mix in the catchment area	Allot a budget/Set P4P incentive scheme	Accept/reject	

Examples of signaling

Informed party	*Uninformed party*	*Information*	*Decision by informed party*	*Decision by uninformed party*	*References*
Potential insuree	Insurer	Current health status	Choose an expensive full coverage contract or a cheap partial coverage contract	Reject/accept the agent	Grossman (1979)
Potential insuree	Insurer	Results of a blood test	Refuse to show results of genetic test	Premium and coverage	Hoy and Polborn (2000)
Doctor	Patient	Doctor's altruism	Time spent in diagnosis	Change doctor	Jack (2005)

Examples of imperfect information

Informed party	*Uninformed party*	*Decision by informed party*	*References*
Potential insuree	Insurer	Engage in some risky behavior	Zeckhauser (1970)
Doctor	Patient	Diagnostic effort to be exerted	Dranove (1988)
Manager of the hospital	Hospital board/owners/government	Managerial effort (e.g., cost containment); Personal interest purchases	Ma (1994)
Owner of a n HMO offering a health plan	Government paying a premium for each enrollee	Distort the quality of some services upwards or downwards in order to avoid high risks/attract low risks	Frank *et al.* (2000)

These are not the only reasons for government intervention. For example, social valuation of market outcomes may lead to government action as well. Generally speaking, dominance of one resource allocation mechanism over the other cannot be presumed (market vs. state). How the markets operate is of concern here.

Having said this, the next question is whether regulation or public provision do indeed palliate any of these problems present in healthcare markets. In doing so, a guide throughout the most related entries in the encyclopedia will be offered. Section Introducing Market Forces in a NHS briefly addresses the introduction of competitive forces in a national health system. Section Market Regulation addresses the regulation of private health insurance and provision markets, Section Duplicate Systems discusses the interaction between public and private insurance. Section A Closer Look at the Provision of Goods and Services in Health takes a closer look at the delivery of health goods and services. Section A Teachers' Guide offers a teacher's guide to these issues. Section Concluding Remarks offers some concluding remarks.

Introducing Market Forces in a NHS

Market forces in NHSs have been implemented basically trough two different policies. On one hand, in some national health systems like that in England, Denmark, Sweden, and Norway, patients are allowed to choose between a set of hospitals in case of needing specialized care. (*see* Chapter 13.10 The Impact of Competition on the Hospital Sector (01310) for an empirical appraisal of the effects of such policy.) On the other hand, some national health systems are remunerating hospitals according to the results of RPE. (*see* Chapter 13.13 Comparative Performance Evaluation: Information on Quality (01313)).

In the first case, the idea is that increasing patients' choice where patients face no copayment would foster competition in quality, as long as quality is observable by patients (or by doctors, but then patients and doctors incentives should be aligned). However, this depends in turn on how the chosen hospital is remunerated. If the remuneration is per episode and is fixed (like in a diagnosis-related group system) and quality is observable, theory predicts an increase in quality, since hospitals offering lower qualities would lose market share. However, it is not so clear that this assumption is satisfied to a sufficient degree. Moreover, even if a specific episode is reimbursed at a fixed fee, a hospital could be in financial trouble if it faces a catchment area where individuals bring higher costs for the same ailment. Things are even worse if hospitals are allowed to set their fees, since in that case they could raise a given service fee without compromising demand (recall that the patient obtains the service for free).

As for the effects of RPE-based remuneration systems under public provision, the Chapter 13.13 Comparative Performance Evaluation: Information on Quality (01313) and Chapter 13.14 Heterogeneity Across Hospitals (01314) review, respectively, issues related to obtaining information on quality of care and on hospital performance at large. The main problem is the asymmetry of information between patients and payers, on one hand, and providers, on the other hand, on provider's effort to deliver the adequate amount of care at the right cost. Asking the providers information faces evident problems related to truth telling and monitoring. Comparing across providers is, in this setting, a natural way to obtain information as long as performance of different providers is correlated. Fichera *et al*. discuss the instruments for quality comparison. One important decision is to which type of quality measurement is more informative, and whether attention should be put in quality of outcomes, in quality of processes, or in quality of inputs. It is not surprising that instead of a single quality indicator, a set is used. Using a set of indicators, however, poses the question of how to aggregate these indicators into a single variable. Without such aggregation it may be impossible to obtain a clear ordering of hospitals according to quality, as some hospitals may fare above the average in some services and worse in others.

On a different but related line, Chapter 13.14 Heterogeneity Across Hospitals (01314) addresses the question of how to structure payments to providers (hospitals) in a way that accounts for their performance and heterogeneity. A prospective payment equal for all providers may induce incentives for selection and for lower unobservable quality. Although in Chapter 13.13 Comparative Performance Evaluation: Information on Quality (01313) the focus in on instruments that may help to measure quality, Chapter 13.14 Heterogeneity Across Hospitals (01314) uses the distinction between long-term and short-term sources of cost heterogeneity, and uses the payment system to not pay for short-term inefficiencies and to keep efficiency incentives. The control of performance in unobservable characteristics is made implicitly through the payment rule.

Market Regulation

Every market is characterized by demand, supply, and a 'mechanism' that connects both to each other. In the standard textbook treatment of markets, that mechanism is the price of the good. Markets included in the health sector often deviate from this simple framework. Market analysis applied to health care or health insurance has to adjust for particular features involved.

First of all, the health sector involves several types of markets. Three of them are highlighted: the market for labor inputs, the market for goods and services, and the market for health insurance. Each of them has their own specific set of issues. Indeed, several entries in the encyclopedia are devoted to pharmaceutical and medical equipment industries.

Take a healthcare good or service. As mentioned above, the uncertainty about the moment and intensity of need for health care leads to the existence of insurance mechanisms (public, private, or both). Such insurance implies markets for healthcare goods and services have a third agent, the insurer, who decouples the price received by the provider from the price paid by the consumer. This third agent may take a passive role, as in traditional reimbursement models, or may take an active role. The active role can range from establishing conditions under which demand can exert (eventually limited) choice of providers, to contracting and paying directly providers or even integrating vertically insurance and

provision. Of course, if a single enterprise implements such integration it is got back to NHS (Figure 1).

The Private Health Insurance Market

Even in a basic private health insurance system, like the one present in Switzerland or in the US for those not covered by Medicare or Medicaid, the market incorporates a vertical dimension that is depicted in Figure 2. In the final stage of the vertical relationship, doctors and hospitals are contracted by insurers in order to provide healthcare goods and services. In the intermediate stage, consumers seek insurance contracts from insurers. Each of these stages is in itself a market, and often researchers have concentrated attention in one of them either by taking the outcomes in the other as given or by assuming that the other performs efficiently. In other words, in studying the insurance market, all the problems listed above are often assumed away in the relation between insurer and provider. Conversely, the insurers' revenue is taken as given when one studies the contracting phase between providers and insurers.

This is not to say that the lessons learned in studying one market are not useful in studying the other. Indeed, in both markets, a major issue is how to set the payments given that each firm (be it an insurer or a final provider) faces a heterogeneous set of consumers. Indeed, an individual may have a high or a low probability of falling ill, which matters for the insurer, and the same individual, once ill and hence requiring a specific treatment, may bring high or low costs, which matters for the service provider. In the first instance and if the insurer is able to choose and collect premia from the individual, the issue is whether and how does the insurer charge (or is allowed to charge) different premia to individuals presenting different characteristics (age or gender for instance). This practice is usually termed risk classification or risk categorization (see Section Risk classification). If the insurer instead receives the premium from a third-party payer (like in Medicare or in the Netherlands), the same issue is termed risk equalization or risk adjustment. In the relationship between the insurer and the provider, one speaks of designing patient classification systems and their use to set risk-adjusted payments.

The overall idea is that premium should fit the expected cost for every individual and that health treatment price should fit its cost. Otherwise, the supplier (again, insurer, or provider) will have an interest in attracting the individuals or serving the treatments where payment exceeds cost (cream skimming or cherry picking) and avoid the individuals or treatments where the inequality is reversed (dumping or skimping). Both opportunistic behaviors fall under the term risk selection.

The importance of these mechanisms in the working of healthcare markets is today clear. Chapters 13.3 Risk Adjustment, the European Perspective (01303), 13.7 Risk Adjustment as Mechanism Design (01307), and 9.18 Risk Selection

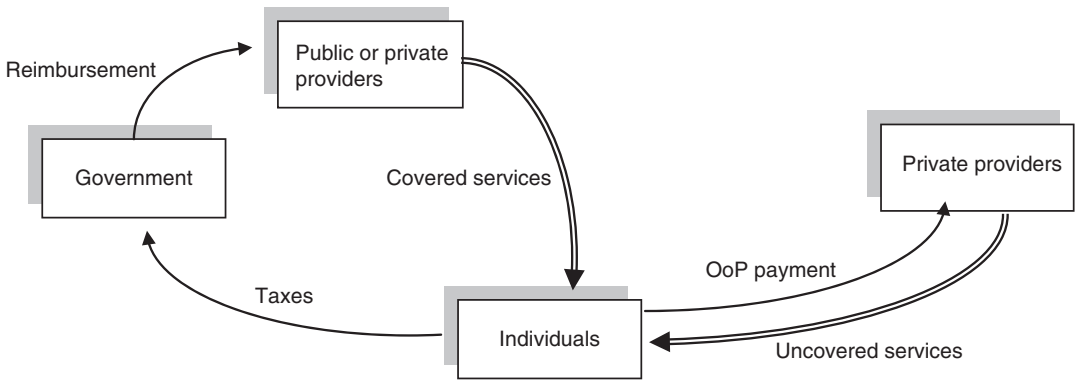


Figure 1 A pure NHS system. **OoP** stands for out-of-pocket payment.

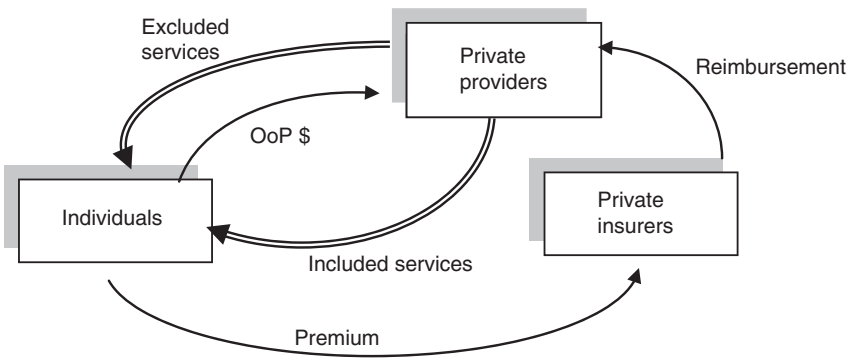


Figure 2 A pure private health system.

and Risk Adjustment (00918) address the role and characteristics of risk adjustment/risk equalization, while Chapter 13.16 Risk Classification and Health Insurance (01316) discusses risk classification. The latter issue is addressed first.

Risk classification

Important issues arise from information asymmetries between agents. A natural intervention is the demand for further information to be incorporated in decisions. In this vein, risk classification aims at reducing informational asymmetries between health insurers and individuals.

Under perfect risk classification, an insurer conditions its contracts on so many individual characteristics that the individual and the insurer have the same information (and therefore the same expectations) about future costs. In this case an individualized premium fitting these expected costs can be set. Such elimination of asymmetric information will lead to an efficient market allocation. However, high risks will face higher premia than low risks. The government can then set appropriate taxes and transfers to improve the welfare of the former at the expense of the latter. Voluntary participation by the low risks may put a limit to such cross subsidization, unless the government makes purchasing insurance mandatory. Incidentally, equalizing premia by the government is not to be confused with some naive 'community rating' where the insurer is not allowed to set different premia to different individuals. Such a policy might either lead to risk selection or to the self-exclusion of the low risks (the so called 'spiral of death').

The problem is that risk classification can only be performed on the basis of observable characteristics of the population. Moreover, collecting data on such characteristics may be quite costly. In any case, only a small set of variables is actually used to design contracts, and this implies that these variables fall very short of being perfect predictors of health risk. As a consequence, the market usually reacts by implementing self-sorting menus of contracts (be at the industry or at the firm level). By this it is meant that individuals, who now have privileged information on their true health risks, reveal this information by choosing one contract instead of another. Such self-sorting menus can only be constructed, however, by reducing the coverage and premia of the contracts aimed at attracting the low risks. Better risk classification could reduce these distortions according to some authors.

Note that the importance of improving – or limiting – risk classification depends on the extent of asymmetric information existing at the outset. This needs to be tested empirically. Current empirical work has progressed but is still far from a definite answer to the question of how significant and pervasive are the asymmetric information problems. Some studies found effects of relevant magnitude whereas others found less impressive implications. A well-recognized problem in testing for asymmetric information is that individuals may have privileged information on dimensions other than risk, and that differences in one dimension could be countervailed by differences in another. For instance, more risk-averse individuals (in principle more willing to pay for coverage) may at the same time have safer habits or lifestyles, which reduces their willingness to pay for coverage.

Switching costs

Whenever more than one possibility of health insurance coverage exists, patients will typically face trade-offs in choosing one health insurer over the other, and issues of switching across health insurers cannot be neglected. Chapter 13.12 Switching Costs in Competitive Health Insurance Markets (01312) details the knowledge in this particular point, individuals' switching across health insurance plans. Taking the Swiss long-lasting experience with health plans' competition, a review of it is of interest to the countries promoting choice of insurance contract. As for consumers' choice and consequently for switching behavior, several issues are particularly relevant: choice overload, the resistance to change (status quo bias); and the existence of risk selection.

Preferred providers

It is stated above that the vertical structure Individuals–insurers–providers is seldom studied as a whole. Some exceptions exist, however. An important issue is with which provider each insurer decides to work. The selection of providers by insurers will also determine how demand for health services is directed toward providers. Chapter 13.15 The Preferred Provider Market (01315) takes up the implications of this market relationship. The main issue is how insurers define the size of the network of providers they use, and how that size depends on the specific rules used to define which providers belong to the network. Selecting a subset of market providers as preferential providers (insiders henceforth) changes the strategic incentives of providers to compete in the market. More specifically, by being included in the network of a health insurance plan (public or private), insiders gain a competitive advantage vis-à-vis the outsiders. To see this, notice that patients will pay less (or even not pay at all) when choosing to be treated by an insider than when selecting an outsider. Hence insiders will face a demand for health care with lower price elasticity, bringing higher prices in equilibrium. This harms the insurer since it must bear higher prices herself. Providers will compete to become members of the preferred network of providers in order to obtain this competitive advantage. Equilibrium may have most or even all providers as preferred ones.

In this context, the payment rules set to providers gain importance as a way to induce competitive pressure. Indeed, the third-party payer can reintroduce competitive pressure by having the patient pay less for the outsider treatment the higher the price of the insider treatment is. In other words, if the insider sets a higher price, the reimbursement received by the patient when choosing an outsider is also larger. Some demand is then diverted from the insider to the outsider. Provider networks are also a form of competition between insurers as well. Patients may opt for one or the other networks based on the list of providers in each network. Competitive forces will be present both across providers and across insurers.

Demand-side issues

The demand side of health care is also characterized by information problems. Patients are not fully knowledgeable about their own health condition, and they are not completely informed about treatment options. Patients rely on physicians to guide them through the healthcare system in order to

restore their health condition. This agency relationship is taken up in Chapter 13.11 Primary Care, Gatekeeping and Incentives (01311), where the interaction of the referencing of patients to other providers (hospitals and specialists) and incentives is discussed. Gatekeeping receives particular attention. In itself, gatekeeping is a constraint on freedom of choice by patients, in the context of a trade-off between free choice and more informed choice. The advantage of using a gatekeeping organization does depend on the incentives of physicians acting as gatekeepers, which raises the issue of incentives faced by gatekeepers to perform their role. Primary care concentrates several roles (health promotion and prevention, diagnosis and treatment, referral, and long-term care). By restricting freedom of choice, gatekeeping is expected to be associated with lower patient satisfaction but also with lower (unnecessary) use of health services and lower expenditures. The empirical evidence on this trade-off has not yet produced results that account for confounding factors introduced by financial incentives faced by physicians (systems with freedom of choice use, generally, fee for service payments, whereas systems with gatekeeping use capitation).

The Role of Risk Adjustment in Market Competition among Health Plans

Countries with competition in health insurance want to ensure, at the same time, affordability of health care to all citizens and nondiscrimination of contribution based on individual risk. Health insurers, however, for the same value of contribution, prefer to contract with the better risks. This led to a role for risk adjustment in market competition among health plans. The solution adopted was to set a two-steps system (Figure 3). First, contributions not based on individual risk are used to build a pool. Second, risk-adjusted payments from the pool to health insurers (or health plans) aim at the double objective of providing enough funds and avoiding incentives for selection of good risks. In most cases these payments are made on a capitative basis, that is, per enrollee in the health plan, hence the term capitation rates. Risk adjustment is an essential element of market competition but its accurate definition is a difficult task. This system is in place, with some variations, in the Medicare sector in the US, the Dutch, Belgian, and German systems, as well as the system in place for public servants in Spain. The European approach to it has been mainly data driven (hence term statistical risk adjustment), attempting to find the more adequate system of risk adjustment based on observables like age, gender, and even prior use of healthcare services. The intricacies of this way to adjust payments are discussed in Chapters 13.3 Risk Adjustment, the European Perspective (01303) and 9.18 Risk Selection and Risk Adjustment, the latter from a US perspective.

A different approach is discussed in Chapter 13.7 Risk Adjustment as Mechanism Design (01307). Although statistical risk adjustment takes it for granted that an insurer will not engage in risk selection if expected costs (calculated *ex ante*) are close enough to the capitation rate, the latter entry admits the possibility that even in this case individuals may make use of their privileged information on their true health risks. Equivalently, there may be observables that are correlated with expected costs but cannot be used in the risk adjustment formula (either due to non discriminatory laws or inherent uncontractability). This implies that both individuals' choices and insurers' behavior must be taken into account when designing the capitation system. This leads these authors to seek an adequate distortion in weights of the risk adjustment model to provide the correct incentives to providers. This approach requires empirical work in understanding how providers (or health insurers) react to risk adjustment rules in order to design these rules taking into consideration such reactions. The empirical challenge is not to find the best statistical fit, but to measure reactions in behavior of health plans (or providers).

Duplicate Systems

In countries with a NHS, the main insurance protection is Government provided. Private (voluntary) health insurance has then a duplicate role of coverage (see Figure 4). Chapter 13.17 The Interaction Public and Private Providers of Health Services (01317) looks into the rationale and implications of such duplicate health insurance coverage in countries with NHS. Duplicate health insurance coverage means private health insurance to cover for the same risks as the NHS. The reasons suggested include promotion of population health, containing health expenditures, increased population choice, and health system 'responsiveness,' whenever the NHS fails to deliver health care to the extent desired by the population. The empirical support for the reasons behind duplicate private health insurance is not fully conclusive.

A different set of questions is related to the impact of double coverage on use of services, and whether it adds, or substitutes for, NHS expenditures. On this aspect, no conclusive evidence is available. Double coverage seems to be associated with higher use of healthcare services, though some of it can be diverted from the NHS. Overall, there is no evidence or

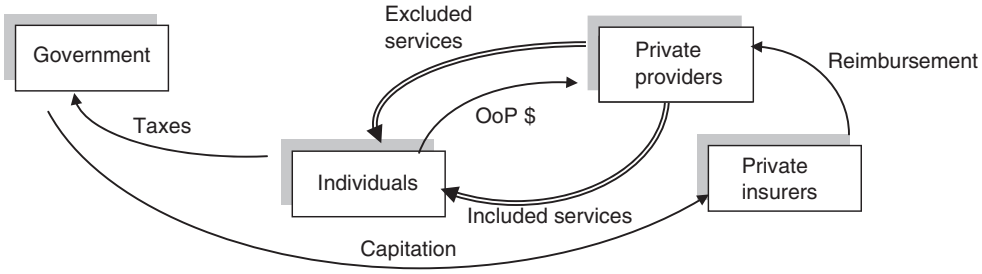


Figure 3 Competition among prepaid health plans.

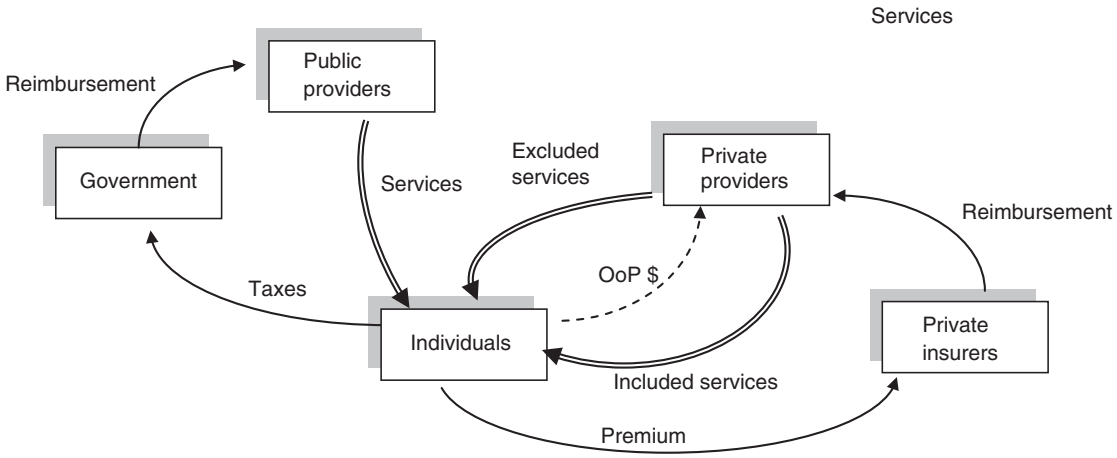


Figure 4 A duplicate system.

theory justifying a strong presumption of more efficient (less costly) healthcare provision through the duplicate coverage than under the NHS. That the private and public systems implicitly compete for patients brings two other issues that are discussed in Sections Waiting Lists and Specialists: the possibility that a physician works for both sectors and the role of waiting as a rationing device.

A Closer Look at the Provision of Goods and Services in Health

On the supply side of healthcare markets, many different providers exist, according to the particular good or service. The main ones are hospitals, primary-care services, imaging services, and pharmacies. Firms operating in other markets have several instruments to compete and attract demand: price, quality, and advertising are the main ones. Their use in health care is often restricted by regulation. Three nonprice competition variables are addressed below, namely, advertising, waiting time, and quality; and two specific providers, Pharmacies and Specialists. Other providers are dealt with in the encyclopedia: *see* Chapters 11.2 The Market for Professional Nurses in the US (01102) and 11.4 Nursing Unions (01104) for nurses, and Chapter 11.11 Dentistry (01111) for dentists.

Advertising

The role of advertising in healthcare markets is discussed in Chapter 13.9 Advertising (01309). Advertising from healthcare providers is often subject to strong restrictions if not banned at all. Advertising has long-been recognized has having two different roles, information and persuasion. Both types of advertising direct demand to the product or service being advertised. Although the informative advertising is usually taken to produce positive effects (as it reduces information asymmetries), persuasive advertising is considered socially wasteful, as it distorts preferences. An important aspect is the ambiguous impact of advertising on unobserved quality of health care, both in theory and in empirical evidence. Advertising seems to matter for competition between healthcare providers, as when bans on advertising are lifted, the later increases. But again the nature of advertising matters, as persuasive advertising may soften competition (and lead to high prices) whereas informative advertising increases price competition.

Waiting Lists

In some healthcare markets, price is not the only instrument to match supply and demand. Owing to random demand and random treatment times, setting demand to equal supply in each moment in time leads to excess capacity and idle resources. In private markets, these random elements are diversified across the several existing providers. In the presence of NHSs (or integration of insurance and provision), the diversification role of random arrivals for treatment and random treatment times cannot be done by choosing available providers. Instead, waiting lists and waiting times are used as an alternative mechanism to balance the system. Determining access to health care based on the price paid is often considered unfair and undesirable, and using time is preferred. Discriminating time to access based on clinical need (prioritization) is acceptable whereas doing it on the basis of ability to pay the price usually is not. Chapter 13.4 Waiting Times (01304) takes up this issue, discussing the role of competition in reducing waiting times in a context where waiting times work as a rationing device to allocate patients across providers. Waiting times are common in NHSs. Waiting times perform several roles. Waiting time works as a variable that balances demand and supply as a substitute to price, as health insurance protection and equity considerations entail prices having a much smaller role. Waiting times lead patients to make a trade-off between faster treatment and price paid when a private sector having no waiting times is available. The use of waiting times and waiting lists can also be seen as an alternative device to redistribute resources, as patients resorting to the private sector to skip waiting lists pay twice.

Quality

In many health systems, prices of health of health care are regulated, either by decision of a NHS or by agreement set

with health insurers. Providers, nonetheless, would like to guide demand toward their own services and products. When price is not available, other competition instruments have to be found. One of these instruments is quality, as long as quality is observable by the key choice maker (which in some cases is the patient and in other cases is the medical doctor, acting as agent of the patient). A main concern is whether competition will lead to lower quality, as providers attempt to save costs, or in higher quality, as providers look at ways to increase demand. According to Chapter 13.10 The Impact of Competition on the Hospital Sector (01310), the second force seems to be stronger.

Pharmacies

Pharmaceuticals are probably the most diffused type of good. Pharmaceutical products can be used by patients during treatment episodes, during admission for treatment (e.g., at hospitals), but are also widely used in ambulatory care. Many chronic conditions are treated on a daily basis with pharmaceutical products. Physicians prescribe the treatment and patients will buy the product from specialized retail outlets, pharmacies. The retail distribution of pharmaceutical products is, therefore, one more aspect of competition in the health sector. There are often constraints on price (regulation of pharmaceutical prices) and margins (distribution margins may be regulated), as well as constraints on entry. In some countries only pharmacists can own pharmacies. In some countries, a new pharmacy can only open on authorization from a regulatory body. In some countries, opening of a new pharmacy needs to obey rules related to population size and distribution as well as distance to competing pharmacies. Chapter 13.5 Pharmacies (01305) presents a thorough and extensive review of how different countries regulate pharmaceutical retail distribution, and they treat the implications of the different regulatory regimes. Price regulation and entry regulation interact strategically and getting it right requires careful analysis.

Specialists

Healthcare services are intensive in labor. Several dedicated professions exist, like physicians, nurses, and pharmacists, for example. The working of labor markets is therefore of importance. This is particularly true for physicians. They have the ability to 'guide' demand (patients) across services and providers. Chapter 13.2 Dual Practice in Duplicate Private-Public System (01302) addresses the positive and normative aspects of how the working of physicians' labor market affects market equilibria. Chapter 13.6 Specialists (01306) looks into the role of physicians as experts acting as agents of patients and of other healthcare providers. Physicians direct demand, which, under public or private health insurance arrangements, is often insensitive to prices. Then, the incentives faced by physicians will be a crucial element in determining how they allocate demand to providers. Gonzalez takes aim at a particular setup for doctors' decisions. In this setup, physicians may work in both a public and a private healthcare provider. Their decisions will define how demand splits across the two sectors. Physicians' decisions are again sensitive to the incentives they face. Possible policies include bans to working in a second job. Theoretical treatments of dual practice provide ambiguous effects on efficiency and quality of care. This ambiguity is not solved by empirical research, leaving room for further work. Although at first inspection a ban on dual practice would be welfare enhancing as it reduces the incentives for physicians to shift patients from public to private healthcare providers, there are other effects at play. Allowing dual practice has the benefit to the public sector of a lower cost to retain highly qualified professionals. Thus, a careful analysis of each institutional setting is called for.

A Teachers' Guide

For a focus on funding and payment in health care that pays special attention to insurance mechanisms, Chapters 13.16 Risk Classification and Health Insurance (01316), 13.17 The Interaction Public and Private Providers of Health Services (01317), 13.12 Switching Costs in Competitive Health Insurance Markets (01312), 13.3 Risk Adjustment, the European Perspective (01303), and 13.7 Risk Adjustment as Mechanism Design (01307) provide an overall view of issues related to market competition and instruments used by health insurance institutions. These readings complement chapters from other parts on health insurance, especially Chapters 9.17 Supplementary Private Insurance in National Systems in the USA (00917) and 9.24 Supplementary Private Health Insurance in National Health Insurance Systems (00924) on supplementary private health insurance and Chapter 9.18 Risk Selection and Risk Adjustment (00918) on the US experience with risk adjustment.

Demand and supply side issues related to health care are dealt within stand-alone entries and therefore some basic knowledge of the microeconomics of consumer behavior would be recommended to fully benefit from these entries.

When the interest lies in vertical market interactions, concentration should be put on Chapters 13.5 Pharmacies (01305) and 13.15 The Preferred Provider Market (01315). As for horizontal competition Chapters 13.6 Specialists (01306) and 13.4 Waiting Times (01304) are of interest.

Concluding Remarks

Since the seminal work of Arrow (1963) the application of economic theory to health care has evolved tremendously. The application of the usual apparatus of demand and supply to healthcare markets has been questioned in many aspects including how society assesses allocations of resources. Both demand and supply side features received attention. Markets of health insurance, healthcare goods and services, and health professions were and are today explored in detail.

Since early the asymmetric information aspects of demand and supply of health insurance were explored. Inefficiencies of market allocations and the issue of existence of market equilibrium were identified. It is now understood to some extent how health insurance markets work, what motives there are

for regulation, but challenges remain. Two of the main ones, deserving both theory and applied research, are risk categorization and risk adjustment and consumers' switching behavior. Market equilibrium is often determined by freedom of choice of consumers, and competitive pressure for efficient supply comes from consumers' exerting choice. Therefore, reducing information asymmetries and understanding how choice of consumers occurs is likely to originate further research. Regulation in health insurance markets is often intertwined with reducing impact of information asymmetries (here the risk pooling funds in certain countries, the mandates for health insurance in other countries and the existence of NHSs as mandatory insurance can be named) and with promotion of consumers' choice of health plans (e.g., such as rules and periods of switching health plans).

The markets for provision of health care also deviate from standard textbook analysis, as patients, the final consumers, often use the services of experts (doctors) to guide their demand of health care. They often have health insurance (either public or private), which makes them less (or even totally) insensitive to price at the moment of use. As a result, consumption decisions are distorted and market prices, if left totally unregulated, too high. Consequently, third-party payers (health insurers, sickness funds, and NHSs) over time moved into a more active role, in both the demand and the supply side. Market equilibrium then becomes the result of the interaction of demand, supply, and third-party payers. The continued growth of healthcare costs leads to interest in how the market allocates resources and how such allocation can be influenced. The role of nonprice market equilibrium mechanisms also ranks high in the agenda. Waiting lists can be named, where time is the rationing device, but also supply-side management such as medical guidelines and protocols, procedure authorizations, and so on. Advertising restrictions in healthcare markets are common, and competition in quality often substitutes for price competition. Bringing together these different aspects into the analysis of market equilibrium and its properties has resulted in a large stream of research in specific topics.

The development of new products and services, innovation, is another area of interest, not the least because technology is considered one of the main drivers of rising healthcare costs. The so-called 'bending the cost curve' in health care will certainly be related to the rate of growth and costs of delivering innovation.

Society's values have an impact on the way market equilibria in healthcare markets are looked at and consequently on the regulation imposed. Access to health care is a major concern. Ensuring access implies the development of networks of providers. The network can be centrally defined and built, as it is the case in NHSs, or can result from decisions of health insurers. Defining networks of providers of health care affects market equilibrium and competition both in health insurance markets and in healthcare provision markets. As third-party payers are increasingly active in managing demand and supply, this area is likely to receive further research attention.

Input markets in health care have own specific characteristics as well. Training healthcare professionals, in particular medical doctors, takes a long time and it is highly costly. Allocation of doctors to training vacancies and specialties is an important issue in many countries. Doctors are special input factors as they determine demand (when acting as agents for patients) and provide services (as suppliers of healthcare services). But also nurses and other professions have markets for their services, and the scope of health professions is changing in response to market forces. As an example of these changes there is the ability, in some places, of qualified nurses to prescribe pharmaceuticals for common health problems of the population. Or the issues associated with dual practice by doctors, especially when both public and private healthcare provision coexist (and compete). The way training and tasks of health professions evolve will potentially change market equilibrium in input markets, influencing the supply side of healthcare provision. Market equilibria will change as well in provision of healthcare services and ultimately in health insurance markets.

Many forces shape market equilibria and regulation in health care. Understanding the economics of markets in health care is an unfinished task, and future research will certainly develop issues addressed in the Encyclopedia (and likely open new areas of research as well).

See also: Advertising Health Care: Causes and Consequences. Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Dentistry, Economics of. Health-Insurer Market Power: Theory and Evidence. Heterogeneity of Hospitals. Interactions Between Public and Private Providers. Market for Professional Nurses in the US. Nurses' Unions. Pharmacies. Physicians' Simultaneous Practice in the Public and Private Sectors. Preferred Provider Market. Primary Care, Gatekeeping, and Incentives. Risk Adjustment as Mechanism Design. Risk Classification and Health Insurance. Risk Equalization and Risk Adjustment, the European Perspective. Risk Selection and Risk Adjustment. Specialists. Supplementary Private Health Insurance in National Health Insurance Systems. Switching Costs in Competitive Health Insurance Markets. Waiting Times

References

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**, 941–973.

Cookson, R. and Claxton, K. (2012). *The humble economist – Tony Culyer on health, health care and social decision making*. York: Office of Health Economics and The University of York.

Dranove, D. (1988). Demand inducement and the physician/patient relationship. *Economic Inquiry* **26**, 251–298.

Frank, R. G., Glazer, J. and McGuire, T. G. (2000). Measuring adverse selection in managed health care. *Journal of Health Economics* **19**, 829–854.

Garcia-Mariñoso, B. and Jelovac, I. (2003). GP's payment contracts and their referral practice. *Journal of Health Economics* **22**, 617–635.

Grossman, H. (1979). Adverse selection, dissembling and competitive equilibrium. *Bell Journal of Economics* **10**, 336–343.

Hoy, M. and Polborn, M. (2000). The value of genetic information in the life insurance market. *Journal of Public Economics* **78**, 235–252.

Jack, W. (2005). Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* **24**, 73–94.

Ma, C. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy* **3**, 93–112.

Rothschild, M. and Stiglitz, J. E. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics* **90**, 629–649.

Zeckhauser, R. (1970). Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**, 10–26.

Further Reading

Barros, P. P. and Siciliani, L. (2012). Public–private interface in health and health care. In Pauly, M., McGuire, T. and Barros, P. P. (eds.) *Handbook of health economics*. Amsterdam: Elsevier Science.

Blomqvist, Å. (2011). Public sector health care financing. In Glied, S. and Smith, P. C. (eds.) *The oxford handbook of health economics*. Oxford: Oxford University Press.

Dowd, B. and Feldman, R. (2012). Competition and health plan choice. In Jones, A. M. (ed.) *The Elgar companion to health economics*, 2nd ed. Cheltenham: Edward Elgar Publishing Limited.

Glazer, J. and McGuire, T. G. (2012). Optimal risk adjustment. In Jones, A. M. (ed.) *The Elgar companion to health economics*, 2nd ed. FALTA. Northampton, MA: Edward Elgar Publishing, Inc.

Iversen, T. and Siciliani, L. (2011). Non-price rationing and waiting times. In Glied, S. and Smith, P. C. (eds.) *The oxford handbook of health economics*. Oxford: Oxford University Press.

Zweifel, P. (2011). Voluntary private health insurance. In Glied, S. and Smith, P. C. (eds.) *The oxford handbook of health economics*. Oxford: Oxford University Press.

Markets with Physician Dispensing

T Iizuka, University of Tokyo, Tokyo, Japan

Introduction

In many countries, physicians play the dual role of prescribing and dispensing medicines. Although this practice is mostly prevalent in Asia, it also exists in some regions of Europe and in some African countries. In these regions, the dispensing physicians profit from selling medicines to their patients. Policy makers have long been concerned about this practice because financial incentives can distort physicians' prescription decisions. This article examines markets in which the physicians dispense medicines by reviewing the experiences of three Asian healthcare systems, Japan, South Korea, and Taiwan, all of which implemented policies to separate prescribing and dispensing. Interestingly, each of these governments responded to this challenge differently, providing valuable insights on the economic consequences of physician dispensing.

Potential Conflict of Interest

In Asia, physicians have long played dual roles as both prescribing physicians and dispensing pharmacists. It is a tradition in Oriental medicine to not differentiate the roles of physicians and pharmacists and for patients to receive drugs directly from their physicians. Healthcare systems in Asia, including China, Hong Kong, Japan, South Korea, Taiwan, Thailand, and Malaysia have followed this tradition. In these healthcare systems, although retail prices (or reimbursement prices) are commonly regulated by the government, wholesale prices (or purchase prices) are not. This situation allows physicians to legally profit from the margin between the retail and wholesale prices. In fact, pharmaceutical companies routinely set a wholesale price that is below the regulated retail price in an attempt to induce demand for their medicines. This practice creates the natural concern that these financial incentives could distort a physician's prescription decisions. Physicians may not choose the best medicine for their patients in terms of efficacy, safety, and/or cost; instead, they may choose a medicine that provides them with the highest margin.

The margin received by physicians can affect their prescribing decisions in three ways. First, the physician's margin can induce therapeutic substitution between brand-name drugs with different active ingredients, all of which could be used to treat the same disease. In such a case, the physician may choose a drug with a higher profit margin even when the drug is suboptimal for the patient. The second possibility is generic substitution, which involves the substitution between brand-name and generic drugs with the same active ingredients. The difference in the physician's margin between the two versions may affect the physician's decision to prescribe and dispense generics. The third possibility is overprescribing. Physicians can increase their profits by simply prescribing and dispensing more medicines to their patients. In some cases, physicians may prescribe and dispense medicines even when none are necessary. All these concerns are realistic; it may be difficult for patients and insurers to verify the appropriateness of the physician's choice even after the patient has taken the medicine.

Governments have been aware that physician dispensing can create a serious conflict of interest between physicians and their patients or payers. However, separating prescribing and dispensing is often difficult because physicians are highly dependent on the profit from dispensing medicines to their patients. In all the three healthcare systems that is discussed below, physician fees have been set relatively low, and profits from dispensing drugs have been a major source of income. Accordingly, physicians have been against the separation policy. An extreme case occurred in South Korea, where a series of nationwide boycotts by physicians occurred when the government mandated the separation of prescribing and dispensing starting on 1 July 2000.

Physician Dispensing in Japan

As in other Asian healthcare systems, physician dispensing is deeply rooted in Japanese society. Following the tradition of Kampo medicines, physicians have customarily prescribed and dispensed medicines to their patients. Although the Meiji government in 1874 considered the separation of prescribing and dispensing as one of the goals of the modern healthcare system in Japan, the actual separation of prescribing and dispensing was virtually nonexistent before the 1970s (Kosaka, 1990; Jeong, 2009).

Prescription drug prices are regulated in Japan. Specifically, although the retail price (or reimbursement price) is regulated by the government, the wholesale price (or physician's purchase price) is not. Thus, physicians can earn margins by both prescribing and dispensing drugs. Moreover, doctors are paid on a fee-for-service basis. Therefore, physicians can increase their profits by overprescribing and dispensing high-margin drugs. As in other healthcare systems in Asia, physicians were highly dependent on the profits from dispensing medicines. Government surveys showed that on average physicians' margins accounted for approximately 25% of the reimbursement price in the early 1980s (Tomita, 2009).

The potential incentive problem created by physician dispensing was not unnoticed. However, the physicians' association (the Japan Medical Association) and the pharmaceutical companies were against any drastic reforms, which made it difficult for the government to mandate the separation of prescribing and dispensing. Instead, the government has instituted two types of incentives to induce physicians not to dispense medicines. First, the government adjusted its price-control rule so that physicians' margins were reduced. The government updates the regulated retail price in April every alternate year, based on both the previous period's retail price

and the average wholesale price. Specifically, the retail price for drug k at year $t+1$ follows the pricing formula below, assuming that the retail price is revised in year $t+1$:

$$P^R_{kt+1} = P^W_{kt} + R_t \times P^R_{kt} \quad [1]$$

where P^R_{kt} and P^W_{kt} denote the retail price and the average wholesale price for drug k at time t, respectively. The government does not allow the retail price to increase over time: if the computed retail price at $t+1$ exceeds the retail price at t, the retail price at $t+1$ is set equal to the retail price at t.

To reduce the physicians' margins, the government has reduced the value of R_t in eqn [1] over time. To see how R_t may affect the physician's margin, M_{kt}, note that M_{kt} is simply the difference between P^R_{kt} and P^W_{kt}. Then, eqn [1] can be rewritten as follows:

$$P^R_{kt+1} = P^R_{kt} - M_{kt} + R_t \times P^R_{kt} \quad [2]$$

Equation [2] implies that if the physician's margin, M, is greater than $R \times P^R$ at t, the retail price for drug k has to decline in the next period. The government reduced R_t from as high as 0.15 in the early 1990s to 0.02 in recent years. The reduction of R_t makes it difficult for pharmaceutical companies to offer a deep discount without substantially lowering the retail prices in the next period. Indeed, as the government hoped, average margins declined over time as R_t decreased, from as high as 25% of the retail price in the early 1990s to approximately 7% in recent years.

Second, to reduce physician dispensing, the government substantially increased prescription-issuing fees, which physicians receive when they write a prescription to be filled at an outside pharmacy. Prescription-issuing fees were 100 yen (approximately $1.3) per prescription in the early 1970s but increased to 500 yen in 1974. The fees were approximately 700 yen (approximately $9.1) in 2005.

Although no existing analysis formally quantifies the effects of these policies on separating prescribing and dispensing, a large number of physicians have stopped dispensing medicines from their offices during the past few decades. To examine whether the reduction in physician dispensing has resulted in lower drug spending, Graph 1 shows how pharmaceutical spending changed between 2001 and 2010, focusing on outpatient office visits. According to the Japan Pharmaceutical Association, the percent of drugs dispensed by pharmacists increased from 45% to 63% during this period. The graph shows that as the separation of prescribing and dispensing increased, pharmaceutical spending at pharmacies doubled, whereas the amount of drugs dispensed by physicians decreased by approximately 20% during the same period. Thus, as expected, the separation policy substantially increased the role of pharmacists. Total outpatient drug spending, which does not include dispensing fees, has increased by 2.4% annually, which is higher than the annual rate of increase in national health expenditures (1.8% between 2001 and 2009).

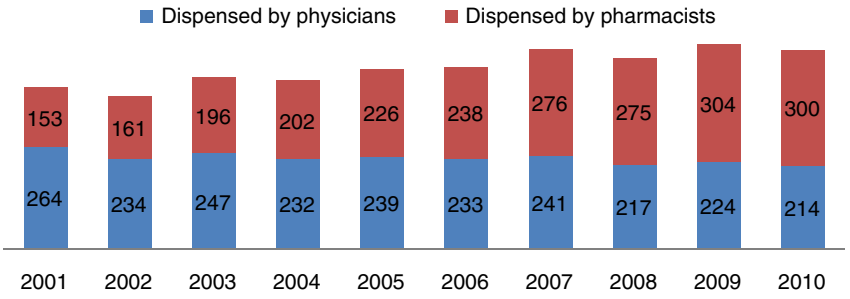
Although these simple comparisons suggest that the separation policy has not necessarily reduced pharmaceutical spending, it is difficult to isolate the impact of the separation policies from the impact of other healthcare policies, such as healthcare financing and provider payment reforms. In the section Overprescribing and Therapeutic Substitution, studies that examine more directly the impact of physician dispensing on pharmaceutical spending and medical expenditures are reviewed.

Overprescribing and Therapeutic Substitution

As noted previously, policy makers have been concerned that physician dispensing may result in overprescribing and substitution toward higher margin drugs. The latter situation results when the physicians' margins differ across drugs. Iizuka (2007) considered these possibilities by empirically examining the Japanese hypertension drug market, which consists of more than 40 brand-name drugs. The data were aggregated at the product level and covered the period between 1991 and 1997, when a majority of physicians dispensed medicines from their offices. Iizuka (2007) assumed that the physician acts as an agent for his/her patient and chooses a hypertension drug from more than 40 brand substitutes.

To examine whether financial incentives lead to overprescribing and substitution toward high-margin drugs, it is necessary to have data on the physicians' margins. Although no official data exist for the physician's margins, Iizuka (2007) calculated them by taking advantage of the pricing rule (i.e., eqn [1]). Specifically, in eqn [1], retail prices at t and $t+1$ (P^R_{kt} and P^R_{kt+1}, respectively) are publicly known. It is then easy to determine the average wholesale price at t (P^W_{kt}) using only the publicly available data. An average physician's margin at time t for each medicine can be obtained simply by taking the

Graph 1: Drug spending in Japan (2001–10)
Outpatient visits only. June of each year (billion yen)



Source: Survey of medical care activities in public health insurance, ministry of health, labour and welfare, 2001–2010.

difference between the retail price and the wholesale price at time t. It should be noted, however, that physicians' margins can be obtained only on average, and if the bargaining power of medical institutions differs substantially, this approach may not be valid.

Utilizing the obtained physicians' margins, Iizuka (2007) estimated a utility-based random coefficient discrete choice model, in which physicians choose one hypertension drug from more than 40 alternatives. Physicians were assumed to choose a drug by taking into account the patient and utility of each drug. In addition to physician's margin, other factors, such as the patient's out-of-pocket cost and the attributes of the drug, were also considered in the estimation. The results indicated that physicians respond to the size of the margin associated with each drug, suggesting that financial incentives created by physician dispensing distort the physicians' prescribing patterns.

To understand the magnitude of the distortion that the physicians' margins create, using the estimated parameter values, Iizuka (2007) conducted a counterfactual analysis in which prescribing and dispensing are hypothetically separated. This analysis was conducted by simply removing the physician's margin from their objective function. Under the assumption that the retail price and other factors do not change, it was shown that the elimination of the physicians' margins reduces total prescribing and pharmaceutical spending by 10.6% and 15%, respectively. This finding implies that the current spending on hypertension drugs is inflated 4.4% from substitution with high-price, high-margin drugs and 10.6% by overuse of drugs. These results support the ongoing concern that physician dispensing results in overprescribing and substitution toward high-margin drugs.

Although the simulation is valuable for quantifying the extent of the distortion potentially created by physicians, at least three issues exist that are outside the model but are important to actual policy making. First, it is likely that when the separation of prescribing and dispensing is mandated, physicians will be compensated for their lost income by, for example, higher physician fees. In turn, this may increase total medical expenditures. Second, the counterfactual simulation assumed that pharmaceutical prices would stay the same after the separation of prescribing and dispensing. However, this may not be true because the government may need to fund any additional payments to physicians by lowering reimbursement prices for pharmaceuticals. Third, as in Japan, the government may attempt to induce the separation by increasing prescription-issuing fees, resulting in overprescribing and higher medical expenditures. Policy makers need to carefully evaluate these possibilities when implementing policies.

Generic Substitution

Although Iizuka (2007) provided valuable insights regarding the effects of physician dispensing on both overprescribing and therapeutic substitution, the analysis is limited because it examines physician decisions only at the aggregate level. Moreover, it does not take into account the dynamic nature of the prescription process, and it does not examine whether physician dispensing affects generic substitution. Iizuka (2012) attempted to overcome these shortcomings by using rich, micro-level panel data covering more than 360 000 observations of over 40 drugs that faced generic competition after 1998. In this study, physician heterogeneity, such as a physician's general preference for generic drugs or whether the physician dispenses a drug from his/her office, is observable. Using this detailed microdata and estimating dynamic probit models, Iizuka (2012) examined the factors that affect the choice between brand-name versus generic drugs with the same active ingredient. As in Iizuka (2007), a physician was assumed to be an agent for his/her patient and to take into account both his/her own and the patient's utility when making a decision regarding which version of the drug to prescribe. During the data period (i.e., August 2003–December 2005), generic substitution was not allowed in Japanese pharmacies. Thus, the study focuses on physicians' generic adoption decisions. The patient's out-of-pocket costs, the physician's margin from each version, state dependence, and patient–physician heterogeneity were also considered as factors that could affect patient and/or physician utility from generic drugs. The physicians' margins were computed in the same way as in Iizuka (2007).

Based on a simple tabulation, Iizuka (2012) showed that generic drugs are more frequently used in small clinics (as opposed to large hospitals). Among the small clinics, generics are more often used by dispensing physicians (as opposed to nondispensing physicians). While dispensing physicians chose generic drugs 50.1% of the time, nondispensing physicians chose generics only 18.5% of the time. In terms of the margins that brand-name and generic drugs offer, the study found that generic drugs typically provide the largest margins immediately after they enter the market, and these margins are substantially larger than those for brand names during the period. However, the generics' advantage in margins quickly disappears after the first period, so the margins offered by brand-name and generic drugs no longer differ substantially. Iizuka (2012) argued that this phenomenon is a direct consequence of the government's price-control rule, as given by eqn [1]. That is, the rule makes it difficult for generic firms to continuously provide large margins because offering a large margin in one period reduces the room for a price discount in the next period.

Estimation results indicated that the dispensing physician's choices are affected by the difference in the margins between brand-name and generic drugs. Thus, as in the case of therapeutic substitution, financial incentives matter in generic substitution. In contrast, the study showed that when prescribing and dispensing are separated, physician prescription choices are not influenced by the difference in margins. This result is expected because nondispensing physicians do not earn the margins and therefore should not be affected by them. The results also indicated that, while dispensing physicians are responsive to patient costs, nondispensing physicians fail to internalize patient costs. This partly explains why substantially cheaper generic drugs are infrequently adopted in Japan. Iizuka (2012) speculated that dispensing physicians are more price sensitive because they directly purchase drugs from wholesalers and thus know more about the price difference between brand-name and generic drugs than do

nondispensing physicians. One implication of this result is that the separation of prescribing and dispensing reduces physician price sensitivity, which, in turn, may increase pharmaceutical spending when pharmacists do not have incentives to substitute generics for brand names. Physicians were also found to differ substantially in their preference for generic drugs, and this heterogeneity plays an important role in the choice between brand-name and generic drugs.

Physician Dispensing in South Korea

As in Japan, physicians in South Korea have long prescribed and dispensed medicines to their patients. However, an interesting difference has existed. In South Korea, not only physicians but also pharmacists have prescribed and dispensed medicines to their patients. On 1 July 2000, the South Korean government implemented a law for mandatory separation between the roles of physicians and pharmacists. After the implementation, physicians no longer dispensed drugs, and pharmacists no longer wrote prescriptions. The policy was intended to address the ongoing concern that physician dispensing (and pharmacist prescribing) induces overprescribing and inappropriate use of medicines, as seen in other Asian healthcare systems. The South Korean government expected that the separation policy would reduce the cost and misuse of medicines and improve drug efficiency (Kim and Ruger, 2008).

South Korea's experience is unique; unlike Japan and Taiwan, it was able to switch from a full integration of prescribing and dispensing to a complete separation of the two functions. Jeong (2009) argued that the president's political leadership and progressive civic groups have played key roles in the drastic reform. The radical change may make it possible to infer the impact of the reform on the outcomes related to our interests. However, existing studies do not employ rigorous identification strategies and simply compare the outcomes before and after the separation policy. Thus, care should be taken when interpreting the results.

With this caveat in mind, the following sections review the literature that examined the effects of the separation policy in South Korea on (1) therapeutic and generic substitution, (2) overprescribing, (3) pharmaceutical spending, and (4) health outcomes.

Therapeutic and Generic Substitution

Evidence indicates that after the separation policy, physicians shifted away from cheaper drugs and toward more expensive drugs (Kim and Ruger, 2008; Kwon, 2009). According to Kim and Ruger (2008), high-priced prescriptions for outpatients increased their market share from 16.0% (in March 2000) to 34.4% (in March 2001) at clinics and from 59.4% to 73.2% at general professional hospitals. The authors also reported that, as a result, sales by multinational companies rose consistently after the reform. Regarding generic substitution, several authors noted that physicians shifted away from cheaper generic drugs toward more expensive brand-name drugs after the separation policy.

These substitution patterns appear to indicate that before the separation policy, lower-priced brand-name and generic drugs provided higher margins for physicians than their substitutes. This conjecture is supported by Iizuka (2007), who demonstrated that when the physicians' margins are eliminated, pharmaceutical demand will shift away from former high-margin drugs toward low-margin drugs. Although no systematic evidence exists on the extent of the margins for drugs in South Korea, Kim *et al.* (2004) noted, "Physicians no longer had any incentive to prescribe cheaper drugs to outpatients after the policy was implemented" (p. 272), which also supports the conjecture. The observed shift away from low-priced drugs after the implementation of the policy makes sense if these drugs provided higher margins for physicians before 2000.

Alternatively, the shift toward more expensive drugs can be explained if physicians became less price sensitive after the separation policy. To the author's knowledge, no empirical study on the South Korean market has shown this relationship. However, as noted previously, Iizuka (2012) showed that, in the Japanese market, dispensing physicians are responsive to price differences, whereas nondispensing physicians are not. This result supports the hypothesis that physicians become price insensitive after the separation policy was implemented.

Overprescribing

Kim and Ruger (2008) reported that the number of prescribed medicines per visit declined approximately 4.8% between 1999 and 2001. Similarly, Kim *et al.* (2004) noted that the prescription rate of antibiotics declined by approximately 4.7% after the separation policy. These results indicate that the quantity of medicine dispensed declined after the separation policy. This is not surprising because the separation of prescribing and dispensing removes any financial incentives to overprescribe, holding all other factors constant. In fact, these numbers may underestimate the impact of the separation policy. After the separation policy, the South Korean government introduced a separate prescription-issuing fee, which created a new incentive to write more prescriptions (Kwon, 2009).

Pharmaceutical and Medical Spending

The above evidence indicates that, on one hand, the separation policy in South Korea reduced overprescribing, but, on the other hand, it caused a shift away from cheaper brand-name and generic drugs. Because these effects potentially cancel each other out, in theory, the total impact of the separation policy on pharmaceutical spending is ambiguous. However, authors agree that pharmaceutical spending increased dramatically after the separation policy was implemented. Kim *et al.* (2004) noted that compared to the first half of 2000, drug spending for outpatient visits increased by 41.6% in the first half of 2001, whereas drug spending for inpatients increased by 22.5%. Kwon (2009) showed that the rapid increase in drug spending continued until 2006. The fact that total drug spending increased after the separation

policy suggests that the shift toward more expensive drugs outweighed the cost savings because of a reduction in overprescribing.

The reader may note that this increase in pharmaceutical spending is not consistent with the counterfactual simulation presented by Iizuka (2007). In the Japanese case, Iizuka (2007) showed that the separation of prescribing and dispensing would reduce pharmaceutical spending both by reducing overprescribing and by increasing the use of less expensive drugs. The latter occurred because of the price-control rule in Japan; high-priced drugs generally provide higher margins to physicians than low-priced drugs. It is difficult for low-priced drugs to continuously provide high margins to physicians because doing so will substantially reduce the retail price of the drug in the following period. This comparison also suggests that the effect of a separation policy on pharmaceutical spending is likely to depend on which drugs provided higher margins to physicians before the implementation of the separation policy.

From the perspective of medical expenditures, it is also important to examine whether the separation policy affected nondrug expenditures, including physician consultation fees. If the latter are raised in exchange for a reduction in drug spending, total medical expenditures may increase. This is an important issue because physicians strongly resisted the separation policy because of their dependence on the profits from drug sales. Kwon (2009) noted that the revenue from drugs typically accounted for over 40% of total revenue. To compensate for the lost income, the South Korean government increased physician consultation fees five times between November 1999 and January 2001 (Jeong, 2009), for a total fee increase of 49%. Kim and Ruger (2008) found that medical expenditures, as a percentage of the gross domestic product, drastically increased after 2000, from approximately 4.5% before 2000 to approximately 6.0% in 2005. Although it is not clear how much of this sharp increase is because of the implementation of the separation policy, concerns were raised that health care expenditures in South Korea were out of control (Kim and Ruger, 2008).

Health Outcomes

As noted previously, one of the objectives of the separation policy was to reduce the inappropriate use of medicines. It was widely known that South Korean physicians and pharmacists were prescribing excessive amounts of antibiotics to their patients (Park *et al.*, 2005; Kim and Ruger, 2008). Before the separation policy, the rate of antibiotic resistance in South Korea was one of the highest in the world, and the overuse of antibiotics was considered to be the main cause (Kim and Ruger, 2008). By reducing the incentive to overprescribe drugs, governments hoped that the inappropriate use of drugs would be reduced.

Park *et al.* (2005) examined whether the separation policy reduced the inappropriate use of antibiotics. They looked at physician prescription choices in January of 2000 and 2001 and examined whether antibiotic prescribing in cases of viral illness, for which antibiotics are inappropriate, declined after the reform in comparison to cases of bacterial illness, for which the use of antibiotics may be justified. The author found that antibiotic use declined in both groups, but the reduction was larger for patients with viral illness (from 80.8% to 72.8%) than for patients with bacterial illness (from 91.6% to 89.7%). Kwon (2009) also reported that, before the separation policy in January 2000, 57.7% of prescriptions included antibiotics, but that number decreased to 45.6% after the separation policy in January 2002. These numbers appear to indicate that the separation policy had reduced antibiotics usage. However, the use of antibiotics remains very high among patients with viral illness, even after the separation policy. Moreover, to the author's knowledge, no direct evidence has shown that the separation policy improved health outcomes. Clearly, the impact of physician dispensing on health outcomes is understudied, suggesting the need for additional research on this important issue.

Physician Dispensing in Taiwan

As in Japan and South Korea, physicians in Taiwan have traditionally prescribed and dispensed medicines to their patients and have thus earned margins. The physicians' margins in Taiwan appear to be large. Chou *et al.* (2003) noted that unofficial estimates indicate that physicians' margins represent half of drug reimbursement prices. Patients pay copayments, but they are relatively low (Liu *et al.*, 2009).

The Taiwanese government has been concerned that financial incentives might lead to an excessive use of medicines. In 1997, the government implemented a separation policy that prohibits physicians from directly dispensing drugs to their patients. However, as in other countries, physicians were against the separation policy because they were dependent on the revenue generated by dispensing medicines. To gain support from physicians and pharmacists, the government increased physician consulting fees and pharmacist dispensing service fees. Furthermore, the government made a major concession as part of the separation policy: physicians were allowed to dispense drugs from their offices if they hired an on-site pharmacist. This is in contrast to the South Korea's separating policy, which prohibited all medical institutions from employing pharmacists or having on-site pharmacies (Kim *et al.*, 2004). As a result, although almost no clinics in Taiwan had on-site pharmacists before the separation policy, nearly 60% of them subsequently hired on-site pharmacists (Chou *et al.*, 2003). Thus, a large number of clinics continued to dispense drugs even after 1997.

An important aspect of the separation policy in Taiwan was that the policy was phased in between 1997 and 2000, which allowed researchers to rigorously examine the impact of the separation policy by implementing the difference-in-differences approach. This approach identifies the effect of a policy through a before-and-after comparison with a control group. Chou *et al.* (2003) conducted such a study and reached the following three findings. First, the separation policy reduced the drug prescription rate and drug spending per visit by 17–34% and 12–36%, respectively, for visits to non-dispensing clinics relative to the control group. This shift is consistent with the studies previously discussed and indicates

that the separation of prescribing and dispensing reduced prescribing. The reduction in drug spending largely results from the reduction in number of drugs prescribed, suggesting that no clear shift occurred toward either more or less expensive drugs in Taiwan.

Second, in contrast to the effect of the separation policy on drug spending, Chou *et al.* (2003) did not find that the policy had an impact on medical expenditure, which includes drug prices, lab tests and diagnostic expenses, dispensing fees, and consultation fees. This lack of impact implies that the reduction in drug spending was offset by physician fees and dispensing fees, both of which were intentionally raised to gain support for the policy from physicians and pharmacists (Chou *et al.*, 2003).

Third, the study found that the separation policy had no effect on drug spending for the clinics that hired on-site pharmacists. By permitting physicians to hire on-site pharmacists, the separation policy failed to alter physician prescribing behavior. This example demonstrates that, as Hsieh (2009) argued, it is critical to break the link between profit margins and physician prescribing behavior to prevent the inappropriate use of medicines.

Research Agenda

As reviewed in this article, a growing number of papers have examined the impact of physician dispensing on physician prescribing patterns, pharmaceutical spending, and medical expenditures. This frequency is not surprising given the prevalence of physician dispensing in Asia and its potential impact on health outcomes and medical expenditures. However, most existing studies simply compare the outcomes before and after the separation policy without controlling confounding factors that would also influence the outcomes. Because other policy reforms, such as healthcare financing, provider payments, or pharmaceutical pricing reforms may occur simultaneously, these studies face difficulties in isolating the effects of the separation policy. Only a limited number of studies have rigorously quantified the impact of physician dispensing on physician prescribing behavior and medical expenditures. Clearly, more research is needed to improve our understanding of this important issue.

Research that examines the impact of physician dispensing on health outcomes is even more scarce. A major concern regarding physician dispensing is that physician dispensing could adversely affect health outcomes as a result of overprescribing or inappropriate medicine choices. The literature that most directly investigates these issues consists of studies that examined whether physician dispensing increased the rate of antibiotics prescriptions (e.g., Park *et al.*, 2005). As previously noted, important progress has been made on this front. To the author's knowledge, however, it is still unknown whether physician dispensing practices ultimately affect health outcomes. Given the importance of this issue, more research is needed to clarify the effect of physician dispensing on health outcomes. Indeed, without such analysis, one has to be very careful about discussing the welfare implication of physician dispensing.

Conclusions and Lessons Learned

This article examined markets with physician dispensing, focusing on the impacts of physician dispensing on their prescribing patterns, drug and medical expenditures, and health outcomes. The experiences of three Asian healthcare systems, Japan, South Korea, and Taiwan, were reviewed. Although these systems faced the same concerns that physician dispensing could lead to overprescribing and inappropriate use of medicines, the governments intervened in the markets differently, providing valuable insights on the impact of physician dispensing.

Japan did not ban physician dispensing but instead created financial incentives to encourage physicians to refrain from dispensing drugs. That is, the physicians' margins were gradually reduced, whereas prescription-issuing fees were raised. As a result, according to Japan Pharmaceutical Association, the percent of drugs dispensed by pharmacists increased from 12.0% in 1990 to 20.3%, 39.5%, 54.1%, and 63.1%, in 1995, 2000, 2005, and 2010, respectively.

Although it is apparent that physician dispensing has decreased over the past 20 years, it is not clear whether the reduction in physician dispensing has reduced overprescribing, drug spending, or medical expenditures. To induce the separation of prescribing and dispensing, the government has substantially increased the prescription-issuing fees, which may have encouraged overprescribing and resulted in higher medical expenditures. Dispensing fees for pharmacists were also substantially raised, further increasing medical expenditures. Thus, the total impact of the separation policy on drug spending and medical expenditures is not clearly known.

In contrast to the gradual approach taken in Japan, South Korea enforced the separation of prescribing and dispensing, making physician dispensing illegal after 1 July 2000. This drastic approach faced strong protests by physicians and resulted in substantial fee increases for them. Evidence also indicates that after the separation policy, physicians shifted away from low-priced drugs toward high-priced drugs, which substantially increased pharmaceutical spending. Both of these changes appear to have contributed to the sharp increase in pharmaceutical and medical expenditures after the separation policy.

Beginning in 1997 Taiwan also made physician dispensing illegal. However, when faced with the strong opposition of physicians, Taiwan created a major loophole: clinics were allowed to continue dispensing as long as they hired an on-site pharmacist. The majority of clinics were therefore allowed to continue both prescribing and dispensing medicines, even after 1997. As a result, the separation policy had little impact on physician prescribing behavior. For the small number of physicians who stopped dispensing drugs, the separation policy appears to have reduced total prescribing. However, the reduction in drug spending was offset by higher physician fees, resulting in little change in total medical expenditures.

The lessons for policy makers can be summarized as follows. First, consistent with an ongoing concern, evidence indicates that physician dispensing distorts physician prescribing decisions by creating financial incentives to both overprescribe and substitute toward higher margin drugs. Thus, holding everything else constant, eliminating the physicians' margins

will mitigate these distortions. As a result, separating prescribing and dispensing can potentially improve health outcomes.

Second, although the separation policy may remove the incentive to overprescribe and to substitute toward high-margin drugs, it does not necessarily reduce pharmaceutical spending. For example, by eliminating the physicians' margins, demand for cheaper brand-name and generic drugs could decline if these drugs provided higher margins before the separation policy. Alternatively, physicians could become less price-sensitive after the separation policy because they would no longer purchase drugs directly from the wholesalers. Thus, if the goal of the separation policy is to reduce drug spending, additional policies – such as global budgets, which will increase the price sensitivity of physicians – may also have to be implemented.

Third, the experiences of the three healthcare systems suggest that the aforementioned assumption that 'everything else is constant' does not usually hold true. That is, if the physicians' margins are eliminated, the physicians' lost income must be compensated by, for example, higher physician fees or prescription-issuing fees, both of which increase medical expenditures. Moreover, if the prescription-issuing fees are set higher than the marginal cost of writing a prescription, overprescribing will be encouraged even when the physicians' margins are eliminated, further increasing drug spending. Because pharmacists are assuming new tasks, separate fees may also have to be paid to the dispensing pharmacist. Policy makers should be aware that this additional spending is difficult to avoid. Unless this spending is funded by a reduction in pharmaceutical spending, separation policies may result in a substantial increase in total medical expenditures. Reduction in drug spending can be achieved, for example, by a decrease in overprescribing or by reducing pharmaceutical prices. The latter may be justified because the pharmaceutical companies will no longer pay margins to physicians.

This discussion indicates that the success of a separation policy critically depends on how policy makers construct the details and take into account the interdependence of healthcare policies, such as physician dispensing, pharmaceutical pricing, and provider payments. The author hopes that this short article helps policy makers anticipate the key issues to be considered before designing policies related to physician dispensing.

See also: Physician-Induced Demand

References

Chou, Y. J., Yip, W. C., Lee, C. H., et al. (2003). Impact of separating drug prescribing and dispensing on provider behaviour: Taiwan's experience. *Health Policy and Planning* **18**(3), 316–329.

Hsieh, C. R. (2009). Pharmaceutical policy in Taiwan. In Eggleston, K. (ed.) *Prescribing cultures and pharmaceutical policy in the Asia-Pacific*, pp. 109–125. Baltimore, MD, USA: Brookings Institution Press.

Iizuka, T. (2007). Experts' agency problems: evidence from the prescription drug market in Japan. *RAND Journal of Economics* **38**(3), 844–862.

Iizuka, T. (2012). Physician agency and adoption of generic pharmaceuticals. *American Economic Review* **102**(6), 2826–2858.

Jeong, H. S. (2009). Pharmaceutical reforms: Implications through comparisons of Korea and Japan. *Health Policy* **93**, 165–171.

Kim, H. J., Chung, W. and Lee, S. G. (2004). Lessons from Korea's pharmaceutical policy reform: The separation of medical institutions and pharmacies for outpatient care. *Health Policy* **68**, 267–275.

Kim, H. J. and Ruger, J. P. (2008). Pharmaceutical reform in South Korea and the lessons it provides. *Health Affairs* **4**, 260–269.

Kosaka, F. (1990). *Iyaku Bungyo-no Jidai (The era of separation of prescribing and dispensing)*. Tokyo, Japan: Keiso Shobou. (in Japanese).

Kwon, S. (2009). Pharmaceutical policy in Korea,. In Eggleston, K. (ed.) *Prescribing culture and pharmaceutical policy in the Asia-Pacific*, pp. 31–44. Brookings Institution Press.

Liu, Y. M., Kao Yang, Y. H. and Hsieh, C. R. (2009). Financial incentives and physicians' prescription decisions on the choice between brand-name and generic drugs: Evidence from Taiwan. *Journal of Health Economics* **28**(2), 341–349.

Park, S., Soumerai, S. B., Adams, A. S., et al. (2005). Antibiotic use following a Korean national policy to prohibit medication dispensing by physicians. *Health Policy and Planning* **20**(5), 302–309.

Tomita, N. (2009). The political economy of incrementally separating prescription from dispensation in Japan. In Eggleston, K. (ed.) *Pharmaceutical policy in the Asia-Pacific*, pp. 61–76. Baltimore, MD, USA: Brookings Institution Press.

Further Reading

Eggleston, K. (ed.) (2009). *Prescribing cultures and pharmaceutical policy in the Asia-Pacific*. Baltimore, MD, USA: Brookings Institution Press.

Iizuka, T. (2009). The economics of pharmaceutical pricing and physician prescribing in Japan. In Eggleston, K. (ed.) *Prescribing cultures and pharmaceutical policy in the Asia-Pacific*, pp. 47–59. Baltimore, MD, USA: Brookings Institution Press.

Measurement Properties of Valuation Techniques

PFM Krabbe, University of Groningen, Groningen, The Netherlands

Introduction

In medical decision analysis and economic evaluation of health care, states of illness or disability (hereafter called 'health states') are commonly valued on a scale from zero to unity. A value of 0 is assigned to the state of being dead (or a state equivalent to being dead), whereas a value of 1 is assigned to 'full health.' The values are called preference scores or utilities and may be used to weigh life years in evaluations of health outcomes. Several techniques can be used to elicit values for health states from individuals, including the standard gamble (SG), time trade-off (TTO), rating scale, magnitude estimation (ME), person trade-off (PTO), Thurstone scaling, and extensions of this latent scaling model, the class of discrete choice (DC) models. They are based on different theoretical assumptions and stemming from different disciplines (e.g., health economics, psychology, and public health). Empirical studies on the relationship between the outcomes of these valuation techniques have shown that there are differences in the values elicited by the different valuation techniques and in their measurement properties. So far, there is little agreement about which technique is the most appropriate.

For health state values to be useful to decision makers, the numbers should accurately represent the genuine value or attitude of the subjects from whom they were elicited toward the health states in question. The extent to which this is the case depends on the psychometric or measurement properties of the elicitation techniques used to establish the values. In the context of health economics, the most salient psychometric properties are validity and reliability. From the area of clinimetrics the concept 'responsiveness' has been introduced as an important property of health outcome measures. More general is the idea of 'level of measurement,' which is related to the field of measurement theory, and that is more directed on the information level of the responses captured by various measurement approaches. This article explains what is meant by each of these and reviews the valuation techniques mentioned above with respect to these properties.

This overview is only dealing with valuation of health states derived from a group of respondents. In the area of clinical decision making, often individual patients are involved in eliciting values for health states that concerns possible outcomes related to their own disease and optional treatment modalities. The measurement properties of these patient values will not be discussed and presented in this article. The main reason for refraining to incorporate these types of values is because many measurement properties cannot be (directly) estimated on an individual basis or are rarely performed.

Validity

Validity refers to the degree to which an instrument really measures what it intends to measure. Another definition used in educational and psychological testing is that it is an overall assessment of the degree to which evidence and theory support the interpretation of the scores entailed by proposed uses of the instrument. Validity is thus concerned with the nature of 'reality' and the nature of the entity being measured. Especially for (partly) subjective phenomena, such as the valuation of health states, the determination of validity seems to be a process that involves the incremental accumulation of evidence rather than one definitive comparison. As opposed to outcomes such as temperature, blood pressure, or survival, health status is not directly observable and its appraisal is to some extent normative.

Validity encompasses three main aspects each with a rather broad scope: content validity, criterion-related validity, and construct validity. Content validity refers to the question: 'Is the instrument really measuring what we intend to measure?' For the purpose of this study, this implies a discussion about the 'real' meaning and interpretation of values elicited by valuation methods. Are they really representing individual expressions of health state preferences? Criterion-related validity is only applicable if one method can be identified as superior, i.e., a 'gold standard.' As these issues are part of an ongoing debate, content and criterion-related validity are not addressed. Here convergent validity which may be regarded as a type of construct validity is primarily dealt with. In convergent validity research, the degree of association (i.e., correlation coefficients) between measures of constructs that theoretically should be related to each other is estimated, that is, patterns of intercorrelations among measures are looked at. Correlations between theoretically similar measures should be 'high.' A detailed discussion about the validity in valuation techniques, which revolves around the content of the HRQoL concept, would go beyond the scope of this article.

Convergent Validity

A variety of relationships between values from different valuation techniques has been reported in the literature. From these studies it is clear that the different valuation techniques produce different value functions. In general, SG and TTO values are all higher than visual analog scale (VAS) values. Under highly controlled experimental circumstances, a study by Krabbe *et al.* (1997) showed that in students the SG and the TTO are producing equivalent valuations to a large extent, despite their apparent conceptual difference. Results from this study can be compared with the few existing studies that have examined this issue, taking into consideration (**Figure 1**) that in the latter studies the numbers of health states and/or participants have usually been small and the statistical techniques rather global. A paper by George Torrance published in 1976 showed a coefficient of determination (R^2) of 0.95 between SG and TTO. These coefficients are based on the mean values of only six health states assessed by local alumni of McMaster University. In Torrance's study, the very bad and the very good

 doi:10.1016/B978-0-12-375678-7.00504-6

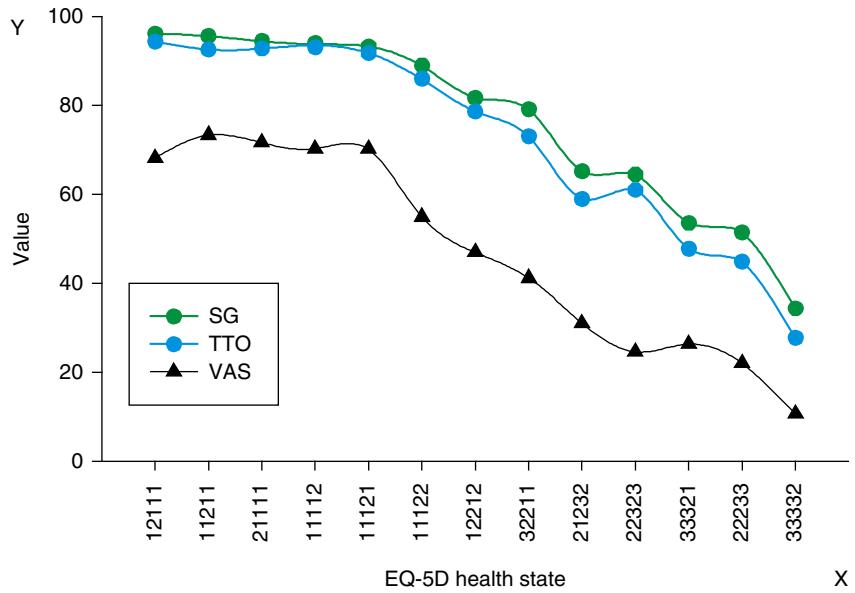


Figure 1 Valuations (means) for 13 EuroQol health state descriptions elicited by three valuation techniques (ordered by SG values). Reproduced from Krabbe, P. F. M., Essink-Bot, M. L. and Bonsel, G. J. (1997). The comparability and reliability of five health-state valuation methods. *Social Science & Medicine* **45**, 1641–1652.

health states were excluded, which may have improved the coefficients. Comparison of mean values obtained (from a set, N = 52, of physicians, therapists, family members, and patients) with the SG and the TTO for 35 disability levels in a study conducted by Alan Wolfson and colleagues in 1982 resulted in an R^2 of 0.84. In 1984, Leighton Read and colleagues presented a Pearson correlation coefficient of 0.65 between the SG and the TTO based on assessments made by 67 physicians. Their study was based on the valuation of only two health states. John Hornberger and colleagues in 1992 reported a Spearman rank correlation of 0.31 between the SG and the TTO. Their results were based on 58 individual patients' valuations of their own health. Comparisons of these methods are inevitably problematic as the techniques used vary across studies (study design, framing) as well as mechanisms for transforming raw values, such as done not only for the TTO (states worse than dead) but also for the VAS (based on position of 'dead').

In the earlier mentioned study of Krabbe *et al.* (1997), valuations based on a VAS (a type of rating scale) were distinct from, but strongly related to, values derived from the two trade-off methods. A simple one-parameter power function sufficed to transform VAS values to SG or TTO. A smaller study by Eric Bass and colleagues from 1994, focused on deriving values for health states in gallstone disease, demonstrated a consistent and substantial difference between values derived by a rating scale technique and those obtained by an SG technique.

In the rise of health state valuation techniques (late 1970s and early 1980s), some studies have been investigating ME. The most well known one is probably the study by Rachel Rosser and Paul Kind from 1978. However, in this study, ME was not compared with another valuation technique. One year later in 1979, Robert Kaplan and colleagues showed that ME responses are compressed at the lower end of scale near death, which seems inconsistent with their VAS results.

In one of the rare studies focused on the comparison of PTO with other valuation methods, Joshua Salomon and Christopher Murray performed in 2004 a head-to-head study in which they calculated the following Spearman's rank correlations based on responses from 69 public health professionals: PTO versus VAS 0.85, PTO versus TTO 0.84, and PTO versus SG 0.86. For the other combinations they found: VAS versus TTO 0.94, VAS versus SG 0.94, and TTO versus SG 0.92.

Benjamin Craig and colleagues applied in 2009 a secondary analysis on data for 8 countries collected by the EuroQol Group, which enabled them to compare VAS, TTO, and DC values (derived from rank data). They observed between VAS and TTO coefficients ranging from 0.61 to 0.80 (Kendall's tau) and ranging from 0.60 to 0.92 for the strength of the relationship between VAS and DC. In a recent study published in 2010, Stolk *et al.* (2010) observed a convergent validity of 0.93 (ICC; intraclass correlation coefficient) between TTO and values derived under a DC model. Responses in this study were collected from 209 students (**Figure 2**). Another study from the Netherlands by Denise Bijlenga and colleagues in 2009, based on participation of 97 community persons, found convergent validity of 0.72 (ICC) between VAS and TTO. For the comparisons DC between TTO and VAS they transformed the values of these two methods into binary scores (to create data comparable to raw DC input data, e.g., preference data). Coefficients were 0.79 between VAS and DC (Cohen's kappa) and between TTO and DC 0.77 (Cohen's kappa). In an explorative study, Joshua Salomon performed in 2003 a secondary analysis on the original the Measurement and Valuation of Health, University of York, 1995 data that were used to construct the EuroQol-5D valuation function. Transforming this data into rank data followed by DC modeling, a very high similarity between the original TTO data and the DC-derived values was reached: 0.97 (ICC).

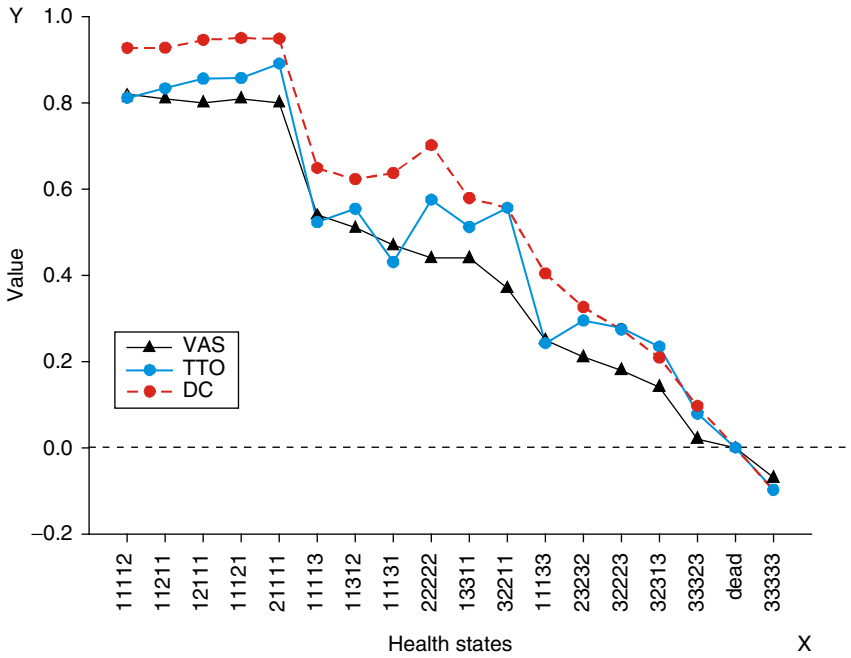


Figure 2 Comparison of values elicited from the student sample: VAS and TTO values for the 17 empirically measured EQ-5D health states and the derived values of the same 17 states based on the DC task. Reproduced from Stolk, E. A., Oppe, M., Scalone, L. and Krabbe, P. F. (2010). Discrete choice modeling for the quantification of health states: the case of the EQ-5D. *Value in Health* **13**(8), 1005–1013.

Reliability

Reliability deals with the stability of measurements, all other things being equal, and with the congruence between raters in the case of the assessment of stimuli (e.g., health states). To a large extent, achieving reliability is a technical matter (e.g., larger samples sizes, repeated measurements, and increasing number of health states). Two distinct types of reliability coefficients can be distinguished in the field of health state valuation: test–retest and interrater. Test–retest reliability is dealing with the reproducibility of a method. If a method is reliable, it should evoke the same outcomes on a second occasion if there is no alteration due to change expected. The most appropriate way of testing this is by computing the ICC between the test and retest. The interrater (or interobserver) reliability seems less suitable for valuation methods, as here assessment is focused on the scaling of health states. Hence, health states that are related to each other with an underlying natural ordering on a unidimensional scale yields another type of data. However, from a more fundamental measurement perspective the interrater reliability can be used as an indicator for the fulfillment of basic measurement requirement in general. In the following, various data on the reliability of individual responses are reported. The reliability of mean responses in groups of people can be much greater, depending on the size of the group, because random variations in individual responses go both ways and tend to cancel each other out.

Test–Retest

In 1976, George Torrance reported in one of the earliest studies involving both SG and TTO a test–retest reliability coefficient (Pearson correlation based on replications) of 0.77 for both SG and TTO. Test–retest was also studied by Donald Patrick and colleagues in 1973 for the rating scale (numbers from 1 to 11) and produced in a group of health leaders a coefficient of 0.79 (Pearson correlation).

There is little evidence for the test–retest reliability of ME and PTO techniques. Of these two techniques, ME would appear to be most promising in terms of reliability. Rosser and Kind reported test–retest reliability for ME at 97% (percentage of agreement). But it is not clear from their publication whether this is done for the real ME tasks or only for the preceding ranking task. In one of the earliest valuation studies, Patrick and his colleagues applied ME for which they presented a test–retest reliability of 0.74.

In 1995, Erik Nord reported relatively poor test–retest findings for the PTO at the individual level, 40% measured by the percentage of agreement, but stressed that group-level reliability could, nevertheless, be satisfactory. In 1997, Christopher Murray and Arnab Acharya reported a lowest correlation coefficient of 0.87 among 9 different groups that performed the same PTO task. Of course, this statistic is only an approximation for the test–retest as not the same individuals were applied.

The classical method of Thurstone scaling (or paired comparison) has been studied in the area of quantifying health states by Paul Kind and David Hadorn in the early days of health-state valuation. Kind applied the classical Thurstone model and an extension of it, the Bradley–Terry–Luce model, in 1982. However, neither reliability statistics nor comparisons (validity) with other methods were performed. Hadorn and colleagues performed in 1992, a Thurstone scaling analysis, but based on an incomplete and selective design

of 54 (59%) of the total number of pairs. In addition, this response mode and analytical steps seem a bit different than the standard approach. Nevertheless, they reported test–retest correlation of 0.79 for Thurstone scaling as well as for the rating scale.

Denise Bijlenga and colleagues also explored a DC model in their 2009 study; the researchers found test–retest results of 0.77 for the VAS (ICC), 0.70 for TTO (ICC), and 0.78 (Cohens's kappa) for DC values.

Interrater

Item Response Theory models, and in particular the Rasch model, are built to deal with 'objective' measurement of subjective phenomena. The most important claim of the Rasch model is that due to the mode of collecting response data in combination with the conditional estimation procedure of the model, the derived measures may fulfill the invariance principle. This is a critical criterion for fundamental measurement. Invariance means that the comparison between two (or more) health states should be independent of the group of respondents that performed the comparisons, and judgments among health states should also be independent of the set of health states being compared.

This invariance principle is closely related to an (implicit) assumption made in the field of health state valuation, namely, that in general people evaluate health states similarly, which permits the aggregation of individual valuations to arrive at group or societal values. The invariance principle seems also related to the IIA assumption (independent of irrelevant alternatives) made in DC models. Therefore, it is important to determine how similar people's judgments actually are for particular valuation techniques, as heterogonous responses (or even distinct response structures) of individuals may indicate that a valuation technique is less appropriate as it may not yield unidimensional responses. Such an analysis can be performed with intraclass correlation statistics (interrater reliability) or specific mathematical routines closely related to factor analysis.

For these reasons, the author wants to assess, additional to test–retest reliability, the consistency across subjects in their task of rating health states (i.e., group level). This type of reliability is indicated as interrater reliability. To compute this reliability coefficient for all health states together, based on a variant of analysis of variance, a global interrater coefficient can be estimated. Formally, this coefficient is a simple adaptation of the conventional Cronbach's alpha (internal consistency measure); instead of multiple items, multiple raters are now being investigated. Although the interrater reliability is formally a statistic that expresses the homogeneity of the responses among raters, this statistic may also be seen as evidence for the content validity. Because a high interrater coefficient may only be expected if most of the raters have a rather similar understanding of the valuation task and in addition come up with comparable preference scores for the valued health states.

By the use of Generalizability Theory, a specific application of analysis of variance, Krabbe *et al.* (1997) were able to reveal various sources of measurement error in the elicited values for health outcomes. Although all the methods to some extent seem to be biased, the valuation methods yield health state valuations that were satisfactorily reliable at the group level: SG 0.58, TTO 0.65, and VAS 0.77. These findings support the validity of constructing societal values for health states based on aggregated data. In an earlier postal survey, which was also conducted using EuroQol health-state scenarios, VAS interrater reliability coefficients in the range 0.77–0.84 were observed by Marie-Louise Essink-Bot and colleagues in 1993. Both results confirm the relatively good properties of VASs with regard to the interrater reliability of the responses. It should be noted that Leighton Read and colleagues in their 1994 study also applied a type of analysis of variance analysis that approximates the G-theory approach. They also found that the variability of responses among respondents is considerably greater for SG than for VAS. Denise Bijlenga and colleagues estimated interrater ICCs for the VAS (0.73), TTO (0.33), and the DC (0.64).

The independency of the set of health states (invariance principle) to be positioned on the VAS has been rejected in two Dutch studies by Han Bleichrodt and Magnus Johannesson in 1997 and Paul Krabbe and colleagues in 2006. Both studies clearly showed that different values will be collected with a multiitem VAS for a fixed set of health states if these are part of varying other states. It is reasonable to assume that these biases may even be larger in the case of measuring health states on a VAS state by state.

Responsiveness

The concept of responsiveness (or sensitivity) has arisen over the past 20 years and refers to the ability of an outcome measure to reflect change. To be of value, an instrument should be stable when no change occurs, although reveal differences in case of improvement or deterioration of a person's health status. The concept of responsiveness has drawn considerable attention among the users of descriptive HRQoL instruments (questionnaires). Most of these users are working in the field of medicine, where responsiveness is part of the clinical framework of health measurement, called 'clinimetrics.' This term was coined to describe an approach to scale development in the area of health that is ostensibly different from the more traditional approach known as 'psychometrics.' These two approaches differ from both in a conceptual and a methodological viewpoint.

Many within the field of descriptive HRQoL or patient-reported outcomes research agree that responsiveness is important, yet there is no consensus on how to quantify it. The confusion even extends to the conceptualization, study design, and measurement of responsiveness. Conspicuously absent is a theory on its relationship to the two classic psychometric concepts, reliability and validity. Responsiveness seems to have a bearing on validity because an instrument first has to measure what it was designed to measure in order to measure accurately. Responsiveness also seems to have a bearing on reliability; if an instrument is unreliable it will not be responsive to changes. Formal research fields in the social sciences (e.g., psychometrics, mathematical psychology, and measurement theory) offer no empirical, theoretical, or mathematical support for the notion of responsiveness. Nevertheless, responsiveness is used here as a theoretical construct that can only be examined by means of

comparison with other measurement instruments and practical experiences.

So far, it seems that the responsiveness has not been investigated for preference-based instruments. This is to some extent explainable as most often valuation techniques are used to quantify certain health states or conditions. They are far less applied to measuring changes between two measurement occasions (in the case of estimating the test–retest property, everything is done to reduce possible changes in the health status of individual). Accordingly, for the applications of the valuation methods there are arguments why this has not been done. Of course, for the use of the so-called preference-based multiattribute systems, such as the EuroQol-5D (EQ-5D), the Health Utility Index Mark III, and the Short-Form 6D it is more informative and more important to have results about the responsiveness of these systems.

Level of Measurement

Apart from theoretical and methodological differences between the valuation techniques, the general underlying assumption is that individuals possess implicit preferences for health states that range from good to bad and that, in principle, it should be possible to reveal these preferences and express them as quantitative or semiquantitative values. The implication of this is that the values should be characterized as interval level data or cardinal data. So, differences between health states should reflect the increment of difference in severity of these states. For that reason, informative (i.e., metric) outcome measures should be at least at the interval level. This means that measures should lie on a continuous scale, whereby the differences between values reflect true differences (i.e., if a patient's score increases from 40 to 60, this increase is the same as from 70 to 90).

Although there have been interest from the onset of quantifying health states in the classical psychometric reliability statistics (validity remains a difficult factor in this area of subjective measurement), far less attention have been directed on the basics of measurement theory in general. Unfortunately, it seems that certain crucial conditions of measurement are hardly recognized by scientists working in the field of quantifying health states. In particular, to arrive at health state values that are characterized as having cardinal or interval level measurement properties, certain basic conditions are required. This involves the invariance principle in collecting response data, but another requirement is unidimensionality of the measurement scale.

Economists tend to claim that responses to the TTO and the SG have interval scale properties, whereas responses to rating scales, including the VAS, tend not to have interval scale properties, given that in the latter, no trade-offs are expressed. Around 1990, Erik Nord and Jennifer Morris/Allison Durand published two papers showing that when subjects locate a set of health states on a straight value line ranging from 0 to 100, most subjects do not intend to express more than ordinal preferences. In a later attempt to find empirical evidence to support that mean health state values collected with a (multiitem) VAS can be characterized roughly as interval data, based on a rank-based scaling model (unfolding), Heleen van Agt and colleagues observed in 1994 a very strong relationship that support the interval property of the raw VAS data. Confirming results were found in a study by Paul Krabbe and colleagues in 2007 that applied nonmetric multidimensional scaling on data (metric and ranks) that were derived from VAS values. Parkin and Devlin (2006) argued that there is no more evidence for interval scale properties in TTO responses and SG responses than in VAS responses.

In a paper by George Torrance and colleagues in 2001, results are presented of a more detailed analysis between the relationship of SG and the VAS. Based on their own study outcomes and incorporating studies from other published studies (all aggregated group means), they show that there is a clear concave curve that passes through 0 and 1 (**Figure 3**). Similar results were also found for the relationship between TTO and VAS. In fact, if the relationship between two different valuation techniques is nonlinear, this implies that at least one of these two methods cannot be regarded as a true metric scale (cardinal or interval differences between the mean values of the health states).

Based on the results of one of the earliest studies to derive health state values based on DC modeling, Salomon concludes that predicted health state valuations derived from a model of ordinal ranking data can provide a close match to observed differences between cardinal values for different states. The model may be used to generate robust predictions on an interval scale, with predictive validity rivaling that of a model estimated directly from TTO values.

To find empirical evidence to support that health state values are overly representing a unidimensional structure, Paul Krabbe in 2006 used a basic mathematical routine to dissect valuation data into underlying dimensions. This study revealed deviating response behavior among the respondents in their health state valuation elicited with the TTO, whereas a

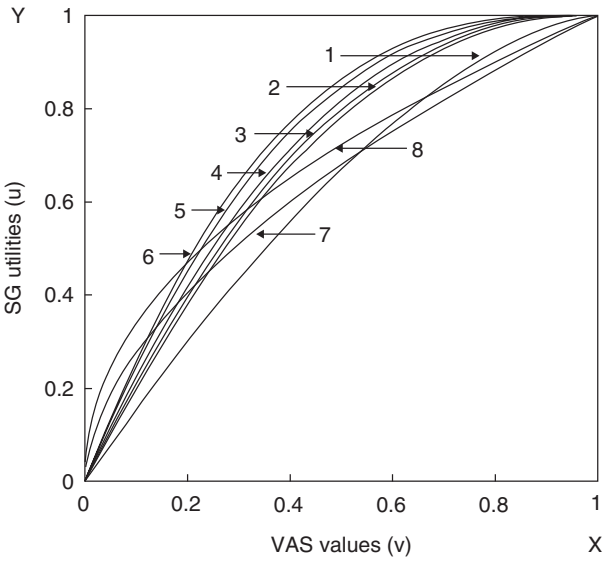


Figure 3 Relationship between mean SG scores and mean VAS scores for health states. Reproduced from Torrance, G. W., Feeny, D. and Furlong, W. (2001). Visual analog scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making* **21**(4), 329–334.

similar analysis on VAS data showed a single dimension. A logical explanation for the absence of unidimensionality of the TTO is that this method is measuring two distinct phenomena (health states and longevity) simultaneously.

Conclusion

It is not surprising that the results found by the author are heterogeneous. Most studies about comparing different valuation techniques were conducted years ago. Certainly in the beginning, most studies were relatively small, often clinically oriented, and there was less harmony about the way valuation methods should be performed. Moreover, in each valuation technique the subjects are faced with a cognitive task that differs from that used with other techniques. In addition, several of the techniques exist in different versions that frame the decisions in different ways. In general, studies focused on comparing different valuation techniques can be differentiated in terms of the type of descriptions of the health states, selection of study population, number of health states, and types of health states. Health states can be divided into hypothetical states and actual or hypothetical health states pertaining to treatment outcomes or particular stages of disease.

Conventionally, the values for different health states used in economic evaluations are derived from a representative community sample. Subjects who value the hypothetical health states need not be familiar with specific illnesses. However, it is reasonable to assume that in many situations healthy people may be inadequately informed or lack good imagination to make an appropriate judgment about the impact of (severe) health states. Many authors assert that individuals are the best judges of their own health status instead of unaffected members of the general population. Numerous studies have found discrepancies in valuations for health states between the general population (healthy people) and people who actually experience illness (patients). Several of these discrepancies can be explained by referring to adaptation mechanism made by patients, but for the frequently applied TTO, it is above all the central element time that likely induce different values for different respondents.

See also: Multiattribute Utility Instruments and Their Use. Quality-Adjusted Life-Years. Valuing Health States, Techniques for

References

Krabbe, P. F. M., Essink-Bot, M. L. and Bonsel, G. J. (1997). The comparability and reliability of five health-state valuation methods. *Social Science & Medicine* **45**, 1641–1652.

Parkin, D. and Devlin, N. (2006). Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Economics* **15**, 563–564.

Stolk, E. A., Oppe, M., Scalone, L. and Krabbe, P. F. (2010). Discrete choice modeling for the quantification of health states: The case of the EQ-5D. *Value in Health* **13**, 1005–1013.

Further Reading

Brazier, J. E., Ratcliffe, J., Salomon, J. and Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University.

Froberg, D. G. and Kane, R. L. (1989). Methodology for measuring health-state preferences – II: Scaling methods. *Journal of Clinical Epidemiology* **42**, 459–471.

Kind, P. (1982). A comparison of two models for scaling health indicators. *International Journal of Epidemiology* **11**, 271–275.

Nord, E. (1992). Methods for quality adjustment of health states. *Social Science & Medicine* **34**, 559–569.

Richardson, J. (1994). Cost-utility analysis: What should be measured? *Social Science & Medicine* **39**, 7–21.

Streiner, D. L. and Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford University Press.

Tengs, T. O. and Wallace, A. (2000). One thousand health-related quality-of-life estimates. *Medical Care* **38**, 583–637.

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal. *Journal of Health Economics* **5**, 1–30.

Measuring Equality and Equity in Health and Health Care

T Van Ourti, Erasmus University Rotterdam, Rotterdam, The Netherlands, and Tinbergen Institute Rotterdam, Rotterdam, The Netherlands
G Erreygers, University of Antwerp, Antwerpen, Belgium
P Clarke, The University of Melbourne, VIC, Australia

Introduction

Health economics is a relatively young subdiscipline, and the measurement of inequalities in the health domain has only relatively recently received attention from health economists. Nevertheless, and perhaps unsurprisingly, the topic has a very long history outside health economics, in particular in public health, demography, sociology, and epidemiology. The notion of a 'gradient in health' across measures of socioeconomic status has been the subject of empirical analysis and speculation regarding its causes for more than a century. For example, in the mid-nineteenth century, William Farr proposed a law relating mortality with population density. At around this time, the famous political economist William Stanley Jevons examined variation in the rate of mortality in different English cities, attributing differences to the proportion of poor Irish immigrants (Jevons, 1870). In the early part of the twentieth century, there were also several empirical examinations of income-related gradients in mortality, including analyses by Hibbs (1915) and Woodbury (1924) of the gradient in infant mortality in the US using information collected from household surveys. A key issue then (as now) was whether the relationship between health and income was purely a correlation, or implied some form of causation. However, most of these early studies reported the health–income gradients only in a tabular or graphical form and did not apply any of the measures the authors examine here.

In this article, the authors give a nonexhaustive overview of the techniques that economists have developed to measure inequality and inequity in health and health care. These measures have their origins in univariate measures such as the Gini coefficient and the Lorenz curve that were developed in the early twentieth century to measure income inequality. Economists also developed bivariate inequality measures, particularly for quantifying the distribution of categories of expenditure across income (Wiśniewski, 1935). Some of these early studies used measures such as concentration curves and indexes to examine health care spending as a component of household expenditure at different levels of income (Iyengar, 1960; Ghezelbash, 1963).

It has only been in the past few decades that these measures have been used specifically for health economics applications. Probably the first proposal for the use of the Gini coefficient in a health economics context can be attributed to Chen (1976), who formulated the K index as a proxy measure of health care quality. The rationale for using this measure was as a way of penalizing situations where avoidable morbidity was concentrated in a small number of individuals rather than being spread more evenly across a community. Le Grand (1987) also applied the Gini coefficient to quantify inequalities in age at death in his international comparisons across a range of high and middle income countries. However, more recent applications of the Gini are less common than studies focusing on bivariate inequality, i.e., the correlation between health and measures of socioeconomic status such as income. Here, the measure traditionally adopted is the concentration index, stemming from proposals of Wagstaff *et al.* (1991), which has been widely employed in international inequality comparisons (e.g., *see* Van Doorslaer *et al.*, 1997). In the past few years, there has been a considerable interest in developing new uni- and bivariate health inequality measures, in part to address some of the aspects of health such as the bounded nature of many health measures (e.g., rates of mortality must fall in the 0–1 range).

The authors' overview focuses on the most important contributions since 2000 and is intended primarily as a catalog of what is available at present. They therefore confine themselves to a short presentation of the various measurement techniques developed by economists; for more in-depth discussions or the literature on the causal mechanisms linking health and income the authors refer to the literature list at the end of this article. The remainder of this article contains four sections. The authors discuss the measurement of inequality in the next section. Next, the authors deal with decomposition methods and introduces methods to measure health inequities. The final section concludes. For brevity, the authors refer to health variables in what follows, but all methods described in this article can be applied to any variable measuring health, health care use, and health care expenditures.

Measurement of Inequality

Measurement of Total (i.e., univariate) Health Inequality

The initial focus of this entry is measurement of the degree of inequality within a given health distribution. The literature on the measurement of this type of health inequality borrows heavily from the literature on the measurement of income inequality.

Throughout this entry, the authors consider a population of n individuals that are ranked by their health levels, i.e., each individual $i=1,\ldots,n$ is characterized by a health level h_i and $h_1 \leq h_2 \leq \ldots \leq h_n$. They always assume that the health variable h_i has a well-defined, finite lower bound $h^{\min}$. With regard to the upper bound, they distinguish between the infinite and finite case, i.e., the authors have either $h_i \in [h^{\min}, +\infty]$ or $h_i \in [h^{\min}, h^{\max}]$. When the health variable is of the ratio-scale type, they assume that $h^{\min}=0$.

 doi:10.1016/B978-0-12-375678-7.00205-4

For ratio-scale variables with an infinite upper bound, the most popular inequality indicator is the Gini index, which can be written as a weighted sum of health shares (Lambert, 2001):

$$G(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_i}{\overline{h}}\left(2R_i^h - 1\right)\right] \quad [1]$$

where $\overline{h}$ denotes average health and R_i^h is the fractional rank of individual i (in the absence of ties we have $R_i^h = (i - 0.5)/n$; in the presence of ties the definition is slightly different). The Gini index is a relative inequality index: it focuses on the relative health differences between individuals. If one wants to stress the absolute differences between individuals, one could use the generalized Gini index, which is an absolute inequality index obtained by multiplying the Gini index by average health:

$$GG(h) = \frac{1}{n}\sum_{i=1}^{n}\left[h_i\left(2R_i^h - 1\right)\right] \quad [2]$$

In principle, any relative or absolute inequality index used for the measurement of income inequality can also be used for the measurement of health inequality. But there is an important caveat: because the health variable is not necessarily of the ratio-scale type, one should not take for granted that indicators developed for ratio-scale variables generate meaningful information when applied to other types of variables, such as nominal, ordinal, or cardinal health variables. Depending on the nature of the variable, different inequality indicators are called for. For instance, Abul Naga and Yalcin (2008) have derived a class of indicators tailored to measure inequality for ordinal health variables.

The situation also changes when the health variable has a finite upper bound, for example, a maximum value of 100%. In that case, one can look either at attainment levels, measured by the health variable h_i, or at shortfall levels, measured by the ill-heath variable $s_i = h^{\max} - h_i$. Recent publications (Erreygers, 2009b; Lambert and Zheng, 2011) have explored what this implies for health inequality measurement. These studies start from the idea that the attainment and shortfall indicators should be complementary, which in its strongest form imposes that attainment inequality is always equal to shortfall inequality. As far as the Gini family is concerned, the strong complementarity criterion leads to the following corrected version of the Gini indicator (Erreygers, 2009b):

$$CG(h) = \frac{4}{n}\sum_{i=1}^{n}\left[\frac{h_i}{\left(h^{\max} - h^{\min}\right)}\left(2R_i^h - 1\right)\right] \quad [3]$$

A similar correction can be applied to the coefficient of variation family. As shown by Lambert and Zheng (2011), the combination of a weak version of complementarity and decomposability points in the direction of the variance as a measure of inequality.

Measurement of Socioeconomic Health Inequality

The dominant strand in the health inequality literature deals with bivariate inequality, and focuses on the correlation between health and socioeconomic status. The most popular measure in this field is the concentration index. Suppose that y_i is a variable which measures the socioeconomic status of individuals; this variable can be occupation, education, income, wealth, etc. Let R_i^y be the fractional rank of an individual according to the chosen socioeconomic variable. The concentration index can be written as (Wagstaff *et al.*, 1991):

$$C(h;y) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_i}{\overline{h}}\left(2R_i^y - 1\right)\right] \quad [4]$$

Observe that the socioeconomic variable need not be of the ratio-scale type; the index only requires information on the socioeconomic rank, which can also be obtained from an ordinal variable.

Different variants of the standard concentration index $C(h;y)$ have been introduced over the years. If one wants to focus on absolute, rather than relative, health differences between individuals, one can use the generalized concentration index $GC(h;y) = \overline{h}C(h;y)$. It is also possible to express different degrees of sensitivity to inequality by using the extended concentration index (Wagstaff, 2002). Again, the authors have a different story when they are dealing with bounded health variables (Clarke *et al.*, 2002). The counterpart of the strong complementarity criterion, which the authors mentioned in the previous subsection, is the 'mirror' condition. This requires that the measured degree of socioeconomic inequality of health should be the reverse of the measured degree of socioeconomic inequality of ill health. Recently, two indicators, which satisfy the mirror condition, have been suggested. The first (Wagstaff, 2005) is defined as:

$$W(h;y) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_i\left(h^{\max} - h^{\min}\right)}{\left(h^{\max} - \overline{h}\right)\left(\overline{h} - h^{\min}\right)}\left(2R_i^y - 1\right)\right] \quad [5]$$

and the second (Erreygers, 2009a) as:

$$E(h;y) = \frac{4}{n}\sum_{i=1}^{n}\left[\frac{h_i}{\left(h^{\max} - h^{\min}\right)}\left(2R_i^y - 1\right)\right] \quad [6]$$

Because both have the mirror property, the level of socioeconomic inequality in health and ill health is identical, except for the sign, i.e., $W(h;\ y) = -W(s;\ y)$ and $E(h;\ y) = -E(s;\ y)$, but in other respects the indices are very different. Erreygers and Van Ourti (2011) provide an in-depth discussion of the properties of these two indicators, in the context of a more general examination of the applicability of rank-dependent indicators.

Decomposition Methods

In the previous sections Measurement of Inequality, the authors covered the most popular inequality indices in the health economics literature. These indices have frequently been used to compare inequality levels between countries or within countries over time but do not allow to infer what lies behind these differences in (socioeconomic) inequality. Decomposition methods – first developed in labor economics and in the income inequality literature – are a useful tool to align the analysis more with an explanatory approach.

Factor Decompositions

Wagstaff *et al.* (2003) were the first to highlight the usefulness of applying existing decomposition methods to the health domain, in particular to the concentration index. When health can be written as a linear function of K factors (e.g., socioeconomic status, demographics, lifestyles, …), one can express socioeconomic health inequality as a weighted sum of the socioeconomic inequalities in these factors. This is most easily seen from combining eqn [4] with a regression model that associates health linearly to K factors x_{ik}:

$$h_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i \qquad [7]$$

where α and $\beta_1,\beta_2,\ldots,\beta_K$ are coefficients and ε_i an error term with zero mean. After some algebra, the following result emerges:

$$C(h;\gamma) = \sum_{k=1}^{K} \left[\left(\beta_k \frac{\overline{x}_k}{\overline{h}} \right) C(x_k;\gamma) \right] + \frac{2\mathrm{cov}\left(\varepsilon_i, R_i^{\gamma}\right)}{\overline{h}} \qquad [8]$$

which shows that socioeconomic health inequality is affected (a) by the magnitudes of the impact of the K factors on health – measured by the average elasticities $\beta_k \overline{x}_k / \overline{h}$ – and (b) by the socioeconomic inequalities in each of the contributing factors – measured by the concentration indices $C(x_k; \gamma)$. There is also a residual term summarizing the covariance between the error term of eqn [7] and the fractional socioeconomic rank. Similarly, one can derive decompositions of the other univariate and bivariate indices discussed in the previous sections Measurement of Inequality. The authors refer interested readers to O'Donnell *et al.* (2006), Erreygers (2009a) and Van Doorslaer and Van Ourti (2011). Readers interested in subgroup decompositions should consult Clarke *et al.* (2003).

Longitudinal Decompositions

A factor decomposition unravels the link between (socioeconomic) health inequality and its associated factors, but in many occasions the authors are interested in the difference between two inequality indices. They now describe decompositions of the change of (socioeconomic) health inequality over time. Note that many of these methods can also be used to decompose differences between countries.

Wagstaff *et al.* (2003) describe an Oaxaca–Blinder decomposition of the change in the concentration index that starts from eqn [8]. It reveals whether changes in (socioeconomic) health inequality are mainly driven by changes in socioeconomic inequalities in the associated factors x_k or by changes in the associated elasticities η_k.

$$\Delta C = \sum_{k=1}^{K} \eta_{kt} [C(x_{kt};\gamma_t) - C(x_{kl};\gamma_l)] + \sum_{k=1}^{K} C(x_{kl};\gamma_l)[\eta_{kt} - \eta_{kl}] + REST \qquad [9]$$

where ΔC denotes the difference between two concentration indices in period t and l; and REST is a residual term.

Van Ourti *et al.* (2009) have adapted the Oaxaca–Blinder decomposition in eqn [9] in order to reveal the relation between the change in income-related health inequality, income growth, and the change in income inequality. This decomposition starts from eqn [7] but allows for a nonlinear association between income (included in x_k) and health. The health elasticity of income turns out to play a crucial role; if this elasticity is increasing with income, then proportional income growth will lead to higher income-related health inequality, and vice versa.

Allanson *et al.* (2010) have recently developed a related longitudinal decomposition that extends the work of Jones and López Nicolás (2004). Jones and López Nicolás (2004) study concentration indices based on short-run (cross-section) and long-run (panel averages) measures of health and socioeconomic status using insights from the literature on income mobility (Shorrocks, 1978), and show they diverge when there are systematic differences in health between those whose socioeconomic status is upwardly and downwardly mobile. An important trademark of their decomposition is that it allows to show whether socioeconomic health inequalities are persistent over time. However, it cannot illustrate whether health changes are more/less pronounced for those with high relative to low socioeconomic status. Allanson *et al.* (2010) show that the change in socioeconomic health inequalities can be written as the sum of 'socioeconomic health mobility' (i.e., the extent to which health changes accrue to those with an initial high relative to low socioeconomic status) and 'health-related socioeconomic mobility' (i.e., the extent to which socioeconomic status changes are larger/smaller for the initially healthy or unhealthy). The same authors have also studied the effects of deaths in longitudinal decompositions (Petrie *et al.*, 2011).

Measurement of Inequity

Until now, the discussion has been mainly confined to ways of measuring and decomposing (socioeconomic) health inequality. Although this is totally in line with having 'the numbers tell the tale', it is not clear whether society at large is concerned about all (socioeconomic) health inequalities. It seems highly plausible that people are concerned about some causes/drivers of inequalities, but less about others. The former is usually denoted as inequity and is the focal point of this section.

Measurement of Horizontal Inequity

The dominant inequity concept is that of horizontal inequity, which states that equals should be treated equally. The concept of vertical equity – which states how unequally unequals should be treated – is as important, but has received far less attention in the literature due to empirical difficulties to estimate the vertical equity norm (Sutton (2002) is a noteworthy exception).

When measuring horizontal socioeconomic inequity in health, one should start by defining whether variation in health attributable to certain factors is equitable or inequitable.

The typical stance in the literature is to consider the variation due to age and sex as equitable and all other variation as inequitable. This is much in line with the practice of standardizing health for age and sex that is popular in public health and epidemiology, but in principle the subdivision between equitable and inequitable health variation allows for a broad range of value judgments (including e.g., the case where equality of health outcomes is inequitable). Two procedures have become popular in the health economics literature (Wagstaff and Van Doorslaer, 2000; Gravelle, 2003; Fleurbaey and Schokkaert, 2009). The first, denoted 'direct standardization', boils down to calculating the predicted value of eqn [7] keeping those factors that lead to equitable health variation fixed (e.g., fixing age and sex at a specific value). The resulting index of socioeconomic inequity $HI^{dir}(h;\ \gamma)$ calculates the socioeconomic inequality in these predicted health values:

$$HI^{dir}(h;\gamma) = C\left(\hat{h}_{|\overline{a},\overline{s}};\gamma\right)$$

$$= C\left(\hat{\alpha} + \sum_{k=1}^{K-2}\widehat{\beta_k}x_{ik} + \widehat{\beta_a}\overline{a} + \widehat{\beta_s}\overline{s} + \widehat{\varepsilon_i};\gamma\right) \qquad [10]$$

where ^ denotes an estimate, and age and sex have been fixed at their average values $\overline{a}$ and $\overline{s}$. The second approach, 'indirect standardization', boils down to calculating the difference between the actual socioeconomic inequality in health and the hypothetical situation where socioeconomic inequality reflects only variation due to equitable variables (which is obtained by fixing the values of the variables that lead to inequitable health variation in eqn [7]):

$$HI^{ind}(h;\gamma) = C(h;\gamma) - C\left(\hat{h}_{|\overline{x_k},\overline{e}};\gamma\right)$$

$$= C(h;\gamma) - C\left(\hat{\alpha} + \sum_{k=1}^{K-2}\widehat{\beta_k}\overline{x_k} + \widehat{\beta_a}a_i + \widehat{\beta_s}s_i;\gamma\right) \qquad [11]$$

where the inequitable variables and the error term are fixed at their average values $\overline{x_k}$ and 0.

The horizontal inequities obtained from eqns [10] and [11] are similar because eqn [7] is linear. Owing to the linearity of eqn [7], it is also straightforward to see that there is an exact link between the factor decomposition of the concentration index and indices of horizontal inequity: in other words, by rearranging the decomposition in eqn [8] – i.e., moving the contributions of age and sex to the left hand side – eqns [10] and [11] are obtained. However, in many empirical applications a nonlinear functional form is preferred for eqn [7] due to the skewed distribution of health. In the latter case, the exact link with the decomposition in eqn [8] is lost, but as long as the variables leading to equitable and inequitable health variation are additively separable, eqns [10] and [11] are still similar. When additive separability no longer holds – which occurs, for example, when the health effect of medical supply (an inequitable variable in our example) depends on the age of the individual (an equitable variable) – eqns [10] and [11] will give different estimates of horizontal inequity. In the next section, the authors discuss this difference in a more general setting and highlight the ethical positions underlying the indirect and direct standardization procedures.

Methodology of Fleurbaey and Schokkaert (2009): Insights from Social Choice

In this section, the authors very shortly introduce a recent contribution to the literature on health equity measurement that is not based on the concentration index. Fleurbaey and Schokkaert (2009) have discussed how the theory of fair allocation (Fleurbaey, 2008) – a social choice theory – could be used to measure health and health care inequities. The most important difference with approaches based on the concentration index (or other related rank-dependent inequality indices) is that it consists of a two-step approach. In the first step, the sole and ultimate goal should be to estimate the 'best' empirical model that links health to its determinants. In a second and independent step the inequities in health are measured, and the procedure is similar whether the underlying equation linking health and its determinants is linear, nonlinear, or not additively separable in the equitable and nonequitable variables. It boils down to subdividing the list of variables into those causing equitable and inequitable health variation (much like before), and next calculates all inequities related to the variables that lead to inequitable health variation. This is different from the methods based on the concentration index that focus on socioeconomic inequity only; and hence the theory of fair allocation allows the measurement of inequities along a broader spectrum of ethical stances.

The first step consists of modeling how health relates to its determinants, i.e., an exercise in pure positive economics. Preferably, a structural econometrics model that disentangles how determinants affect health directly and indirectly (via other endogenous variables such as income, medical care, lifestyles, and so on) is used, but in this section the authors stick to a reduced form to illustrate the most basic version of the approach of Fleurbaey and Schokkaert (2009):

$$h_i = f(x_i) \qquad [12]$$

where $f(.)$ links health to the vector of regressors x_i.

Once $f(.)$ has been estimated, the researcher (or the outcome of a public debate) should subdivide the vector of regressors into a set of variables that lead to equitable (x_i^{eq}) and inequitable health variation (x_i^{in}). Although the description is based on the reduced form in eqn [12], it should be clear that a structural model might be extremely useful in guiding the subdivision, as it allows the distinction of the direct and indirect effects of explanatory variables (e.g., think of a case where the indirect impact of gender on health via unhealthy behavior is considered equitable, whereas the direct impact of gender on health might be considered inequitable). The subdivision allows to introduce two concepts that have been developed in the theory of fair allocation and that are closely related to the two standardization approaches the authors introduced in previous section Measurement of Horizontal Inequity. 'Direct unfairness' is in the same vein as 'direct standardization' and proceeds by fixing the value of x_i^{eq} at a reference value $h_i^{dir} = f\left(\tilde{x_i^{eq}}; x_i^{in}\right)$. The alternative procedure compares actual health with a 'fair' distribution of health where x_i^{in} is fixed, i.e., $h_i^{fg} = h_i - f(x_i^{eq}; (\tilde{x_i^{in}}))$. Next, one calculates inequity in health by measuring inequalities in h_i^{dir} or h_i^{fg}. Fleurbaey and Schokkaert (2009) argue in favor of using an absolute inequality index.

Several things are worth pointing out. First, if socioeconomic status is considered as the only determinant leading to inequitable health variation, these methods conceptually coincide with the approach based on the concentration index; but as soon as other choices are made with respect to the subdivision of factors leading to equitable and inequitable health variation, both approaches will diverge. Second, the approach translates an inherently multidimensional problem into a one-dimensional inequality problem. In contrast, approaches based on the framework of concentration indices are multidimensional in nature. Third and similarly to the discussion of the two standardization procedures in the previous section Measurement of Inequality, the functional form of $f(.)$ is crucial. When additive separability applies so that $h_i = f(x_i) = g(x_i^{ed}) + h(x_i^{in})$, inequalities in 'direct unfairness' and the 'fairness gap' are identical, but when this is not the case inequalities diverge. The theory of fair allocation can however guide the choice between 'direct unfairness' and the 'fairness gap'. 'Direct unfairness' imposes that health differences due to factors leading to equitable health variation are not reflected in estimates of inequity, whereas the 'fairness gap' imposes that absence of inequity in health coincides with an absence of inequitable health variation. Both requirements seem plausible but cannot be jointly true when the function linking health to its determinants is not additively separable in the factors leading to equitable and inequitable health variation. For more discussion, the authors refer to Fleurbaey and Schokkaert (2009).

Conclusion

This article gives a nonexhaustive overview of techniques to measure inequality and inequity in health and health care. The authors have focused on the most important health economics contributions since 2000, but they have also acknowledged that this literature is embedded in a long tradition of research on the socioeconomic health gradient that dates back to more than a hundred years. In fact, the recent research can be seen as revival interest in bivariate as opposed to univariate measures of inequalities.

The first part of the article has dealt with the measurement of univariate inequalities in health. Special attention was paid to bounded health variables; and the implications for health inequality measurement. Next, the authors covered the concentration index and related indices that have been popular to measure bivariate socioeconomic health inequalities.

The second part of the article introduced decomposition methods that are useful to align the analysis more with an explanatory approach. The authors subsequently covered factor decompositions and longitudinal decompositions. The first allows the contribution of separate health determinants to health inequalities to be disentangled; the second is useful to understand what drives changes in health inequalities over time (or between countries) and whether it is always the same people in poor or good health.

In the final part of the paper the authors have moved from inequalities to inequities, i.e., that share of total inequalities that is found to be inequitable. They have covered the traditional approach in health economics that focuses on horizontal socioeconomic-related inequities; and introduced a new and promising approach – derived from social choice theory – that allows to calculate health inequities along a broad set of ethical positions.

Acknowledgments

Tom Van Ourti is supported by the National Institute on Ageing, under grant R01AG037398, and also acknowledges support from the NETSPAR project 'Health and income, work and care across the life cycle II'. This article has benefited from the comments and suggestions of Ulf Gerdtham, Gustav Kjellson, and the editors. The usual caveats apply and all remaining errors are our responsibility.

See also: Dominance and the Measurement of Inequality. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Measuring Health Inequalities Using the Concentration Index Approach. Measuring Vertical Inequity in the Delivery of Healthcare

References

Abul Naga, R. H. and Yalcin, T. (2008). Inequality measurement for ordered response health data. *Journal of Health Economics* **27**, 1614–1625.

Allanson, P., Gerdtham, U. -G. and Petrie, D. (2010). Longitudinal analysis of income-related health inequality. *Journal of Health Economics* **29**, 78–86.

Chen, M. K. (1976). The K index: A proxy measure of health care quality. *Health Services Research* **11**, 452–463.

Clarke, P., Gerdtham, U., Johannesson, M., Bingefors, K. and Smith, L. (2002). On the measurement of relative and absolute income-related health inequality. *Social Science & Medicine* **55**, 1923–1928.

Clarke, P. M., Gerdtham, U. -G. and Connelly, L. B. (2003). A note on the decomposition of the health concentration index. *Health Economics* **12**, 511–516.

Erreygers, G. (2009a). Correcting the concentration index. *Journal of Health Economics* **28**, 504–515.

Erreygers, G. (2009b). Can a single indicator measure both attainment and shortfall inequality? *Journal of Health Economics* **28**, 885–893.

Erreygers, G. and Van Ourti, T. (2011). Measuring socioeconomic inequality in health, health care, and health financing by means of rank-dependent indices: A recipe for good practice. *Journal of Health Economics* **30**, 685–694.

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford: Oxford University Press.

Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics* **28**, 73–90.

Ghezelbash, A. (1963). The urban consumer survey and income elasticities in Iran. *Review of Income and Wealth* **1963**, 168–176.

Gravelle, H. (2003). Measuring income related inequality in health: Standardisation and the partial concentration index. *Health Economics* **12**, 803–819.

Hibbs, H. H. (1915). The influence of economic and industrial conditions on infant mortality. *Quarterly Journal of Economics* **30**, 127–151.

Iyengar, N. S. (1960). On a method of computing Engel elasticities from concentration curves. *Econometrica* **28**, 882–891.

Jevons, W. S. (1870). Opening address of the President of section F (Economic Science and Statistics), of the British Association for the Advancement of Science, at the fortieth meeting, at Liverpool. *Journal of the Statistical Society of London* **33**, 309–326.

Jones, A. M. and López-Nicolás, A. (2004). Measurement and explanation of socioeconomic inequality in health with longitudinal data. *Health Economics* **13**, 1015–1030.

Lambert, P. (2001). *The distribution and redistribution of income* (3rd ed.). Manchester: Manchester University Press.

Lambert, P. and Zheng, B. (2011). On the consistent measurement of attainment and shortfall inequality. *Journal of Health Economics* **30**, 214–219.
Le Grand, J. (1987). Inequalities in health: Some international comparisons. *European Economic Review* **31**, 182–191.
O'Donnell, O., van Doorslaer, E. and Wagstaff, A. (2006). Chapter 17: Decomposition of inequalities in health and health care. In Jones, A. (ed.) *The elgar companion to health economics*, pp. 179–192. Cheltenham: Edward Elgar.
Petrie, D., Allanson, P. and Gerdtham, U. (2011). Accounting for the dead in the longitudinal analysis of income-related health inequalities. *Journal of Health Economics* **30**, 1113–1123.
Shorrocks, A. (1978). Income inequality and income mobility. *Journal of Economic Theory* **19**, 376–393.
Sutton, M. (2002). Vertical and horizontal aspects of socio-economic inequity in general practitioner contacts in Scotland. *Health Economics* **11**, 537–549.
Van Doorslaer, E. and Van Ourti, T. (2011). Chapter 35: Measuring inequality and inequity in health and health care. In Smith, P. and Glied, S. (eds.) *The Oxford handbook of health economics*, pp. 837–869. Oxford: Oxford University Press.
Van Doorslaer, E., Wagstaff, A., Bleichrodt, H., et al. (1997). Income-related inequalities in health: Some international comparisons. *Journal of Health Economics* **16**, 93–112.
Van Ourti, T., van Doorslaer, E. and Koolman, X. (2009). The effect of income growth and inequality on health inequality: Theory and empirical evidence from the European panel. *Journal of Health Economics* **28**, 525–539.
Wagstaff, A. (2002). Inequality aversion, health inequalities, and health achievement. *Journal of Health Economics* **21**, 627–641.
Wagstaff, A. (2005). The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality. *Health Economics* **14**, 429–432.
Wagstaff, A. and Van Doorslaer, E. (2000). Measuring and testing for inequity in the delivery of health care. *Journal of Human Resources* **35**, 716–733.
Wagstaff, A., van Doorslaer, E. and Watanabe, N. (2003). On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *Journal of Econometrics* **112**, 207–223.
Wagstaff, A., Paci, P. and van Doorslaer, E. (1991). On the measurement of inequalities in health. *Social Science and Medicine* **33**, 545–557.
Wiśniewski, J. (1935). Demand in relation to the income curve. *Econometrica* **3**, 411–415.
Woodbury, R. M. (1924). Economic factors in infant mortality. *Journal of the American Statistical Association* **19**, 137–155.

Further Reading

Cutler, D. M., Lleras-Muney, A. and Vogl, T. (2011). Chapter 7: Socioeconomic status and health: Dimensions and mechanisms. In Smith, P. and Glied, S. (eds.) *The Oxford handbook of health economics*, pp. 124–163. Oxford: Oxford University Press.
Fleurbaey, M. and Schokkaert, E. (2011). Equity in health and health care. In Pauly, M. V., McGuire, T. and Barros, P. P. (eds.) *Handbook of health economics*, Vol. 2, pp. 1003–1092. Amsterdam: North Holland.
Gravelle, H., Morris, S. and Sutton, M. (2006). Economic studies of equity in the consumption of health care. In Jones, A. (ed.) *The elgar companion to health economics*, pp. 193–204. Cheltenham: Edward Elgar.
O'Donnell, O., van Doorslaer, E., Wagstaff, A. and Lindelöw, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington DC: The World Bank.
Wagstaff, A. and van Doorslaer, E. (2000). Equity in health care finance and delivery. In Culyer, A. and Newhouse, J. P. (eds.) *Handbook of health economics*, Vol. 1, pp. 1803–1862. Amsterdam: North Holland.

Measuring Health Inequalities Using the Concentration Index Approach

G Kjellsson and U-G Gerdtham, Lund University, Lund, Sweden

Introduction

Health inequality can be defined as variations in health status across individuals within a population. To compare inequalities between countries or over time periods, it may be, for example, interesting to know how much more healthy the healthier individuals are than the unhealthy individuals. However, it may be more interesting to know how health is distributed in relation to a socioeconomic variable. Any version of the concentration index (C) measures inequality in the distribution of a health variable in relation to a socioeconomic rank attached to each individual. Although there are other measures of socioeconomic-related health inequalities (e.g., epidemiologists frequently use absolute and relative range and the population attributable risk), health economists generally use the C. The popularity is probably due to the illustrative and intuitive interpretation. In addition, the C takes the whole population into account rather than only calculating differences between the extremes.

The remainder of this article is a short overview of the recent discussion on how to use different versions of the C to measure socioeconomic health inequalities. The next section defines and discusses the standard C and the related generalized C (GC). These indices are related to the (generalized) GINI coefficient, which is popular within the income inequality literature. Herein, lie parts of the problem of using the C as a measure of health inequalities: health is rarely measured on the same scale as income. Measurement Properties of Health Variables therefore considers the measurement properties of different health variables. Desirable Properties of Inequality Indices discusses desirable properties of an inequality index: the recent literature suggests that an index should be invariant to arbitrary transformations of the health variable. Recent Corrections of Concentration Index (a) presents the recent corrections of C that satisfies these properties and (b) discusses how one may relate to inequality indices for health variables that have different measurement properties from income. Guidelines for Practitioners compiles this literature into a guideline for practitioners and provides an illustration using European Survey of Health, Ageing and Retirement (SHARE)-data.

Concentration Index and the Generalized Concentration Index

Definitions

Just as the GINI coefficient is derived from the Lorenz curve, C is derived from the Concentration Curve (CC). Although the Lorenz curve plots the fraction of the total income concentrated in a fraction of the population ranked by income, CC plots the fraction of the total sum of a health variable that is concentrated in a fraction of the population ranked by a socioeconomic variable (e.g., income). For example, in Figure 1 the poorest 10% possess only 2.5% of the total health that is distributed within the society. As the line at 45° represents a perfectly equal distribution (i.e., the poorest 10% of the individuals possesses 10% of the total accumulated health), it is referred to as the line of equality (LE).

The GINI is equal to twice the area between LE and the Lorenz curve. However, as CC, in contrast to the Lorenz curve, can be both above and below LE, C is defined as twice the area below LE and above CC (i.e., area a in Figure 2) subtracted by twice the area above LE and below CC (i.e., area b). Equivalently, C may be expressed as a ratio between the area $a-b$ and total area below the LE (i.e., $a+c$). Thus, C attains values between -1 and 1. A negative value suggests that the health variable is concentrated among the poor, whereas a positive value suggests that the health variable is concentrated among the rich. Thus, if the health variable is expressed positively in terms of health, a positive (negative) index suggests a pro-rich (pro-poor) distribution. The opposite applies if the health variable is expressed negatively in terms of ill-health. In the former case, C attains its maximum value when all health is concentrated to the richest individual. In the remainder of the article, this will be referred to as the most pro-rich state.

In a finite sample, the C may be formally expressed as:

$$\mathrm{C} = \frac{2}{n\mu}\sum_{i=1}^{n} h_i(R_i - 1)$$

where n denotes the number of individuals, h_i is health of individual i, μ is the mean of h, and $R_i = n^{-1}(i - 0.5)$ is the fractional socioeconomic rank ranging from the poorest to the richest.

The related GC is analogously derived from the GCC, which plots the fraction of the mean of the health variable that is concentrated in a fraction of the population. As GC equals μC, it is not bounded between -1 and 1.

Absolute and Relative Value Judgment

C and GC are sensitive to different types of health changes. C is unaffected if health increases proportionally for all individuals, whereas GC is unaffected if health increases with an equal amount for all individuals. This difference relates to the clash in the income inequality literature between a relative and an absolute view of inequalities (e.g., Kolm, 1976). The degree of inequality can be preserved either if relative differences (ratios) or absolute differences remain the same. However, although income is always unbounded and measured on a ratio scale, health variables can be measured on different scales and can be either bounded or unbounded. It is therefore not appropriate to directly apply the value judgments from the income inequality literature to all health variables. Further elaborations on this question requires a discussion of the differences between health and income.

 doi:10.1016/B978-0-12-375678-7.00206-6

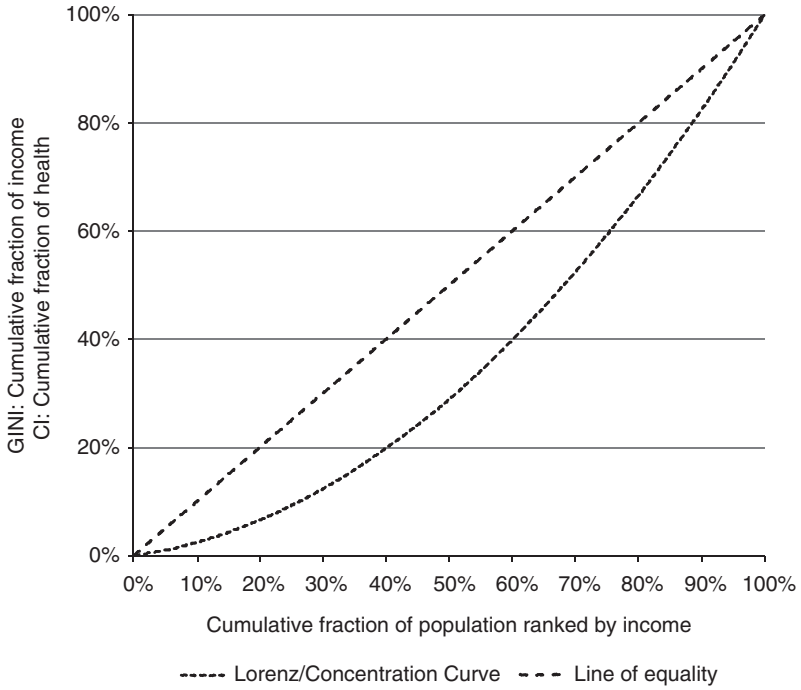


Figure 1 The Lorenz Curve and the CC.

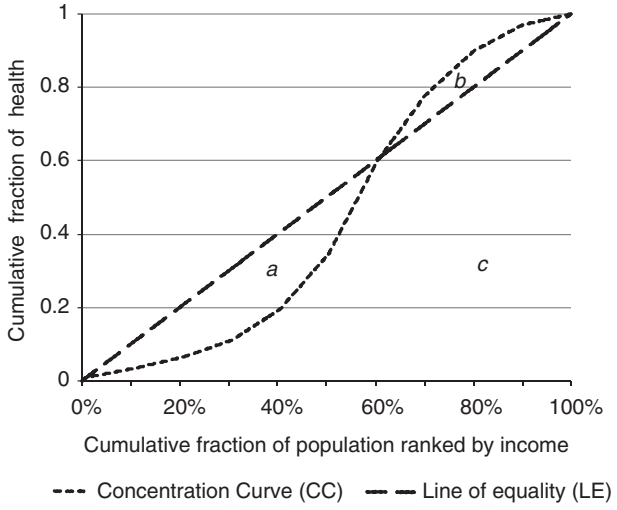


Figure 2 The concentration index. *Note*: Using a graphical representation, C may be defined as $2 \times (a - b)$ in **Figure 2** or as $(a - b)/(a + c)$ in either **Figure 2** or **Figure 3**.

Measurement Properties of Health Variables

Erreygers and van Ourti (2011a) categorize health variables by two dimensions: their measurement scale and boundedness. In principle, health variables can be measured on five different scales:

- Nominal, that is, a scale that allows for classifying, but not ordering, individuals (e.g., type of sickness).
- Ordinal, that is, a scale that allows for ordering individuals but not measuring differences between them (e.g., self-assessed health graded from very bad to excellent).
- Cardinal, that is, the zero point is fixed arbitrary and does not have an intuitive interpretation of total absence, one may meaningfully calculate differences but not ratios (e.g., body temperature or Health Utility Indices (HUIs); The HUI is in the quality-adjusted life-years literature generally anchored between 0 and 1, representing being dead and perfect health respectively, and interpreted as if it was ratio scaled).
- Ratio scale, that is, the zero point corresponds to complete absence and ratios can be meaningfully measured (e.g., health care expenditures).
- Unique, that is, the zero point corresponds to complete absence, and it is not possible to scale the variable (e.g., number of general practitioner (GP)-visits).

Unless one circumvents the meaningless ordinal or nominal differences by projecting the health indicator on a cardinal or ratio scale (e.g., binary variables may always be interpreted as a ratio-scaled variable of average prevalence at the level of deciles/percentiles), one cannot use the C approach for nominal or ordinal scales. Therefore, the remainder of this article only considers health variables that are cardinal, ratio scaled, or unique.

The other dimension in which health variables and income may differ is that while there is no upper bound on income, health variables can be either unbounded or bounded. A bounded variable ranges from a theoretical lower bound $h_{\min}$ to a theoretical upper bound $h_{\max}$. Therefore, one may – in

contrast to unbounded variables – measure both attainments h_i and shortfalls s_i of such a health variable (i.e., $s_i = h_{max} - h_i$). This has crucial implications for the desirable properties and value judgments of the indices, discussed in the sections Desirable Properties of Inequality Indices and Value Judgment for Bounded Variables.

Desirable Properties of Inequality Indices

The literature discusses several possibly desirable properties for health inequality indices. This section considers the most important ones. Although the transfer property and scale invariance are relevant, and indisputably desirable, for all health variables, the mirror condition is only relevant for bounded variables.

The transfer property suggests that if health is (hypothetically) transferred from a poorer to a richer individual, then the inequality index becomes more pro-rich and vice versa.

Scale invariance suggests that the inequality index is unaffected by the scale of the variable (e.g., Erreygers and van Ourti, 2011a). For example, it is desirable that the measured degree of inequality is the same if health spending is measured in Euros or Dollars. For the same reason, it is desirable that the measured degree of inequality remains the same for different cardinal scales.

The mirror condition requires that the measured degree of inequalities is the same for shortfalls and attainments, i.e., the inequality index of attainments should be equal to the inequality index of shortfalls but have the opposite sign. As there is no general consensus of whether it is appropriate to measure inequality in shortfall or attainment, Clarke *et al.* (2002) highlight that the mirror condition may be desirable as it assures that the ranking between populations is the same irrespective of the chosen perspective. However, although the first two properties are indisputable, the mirror condition implies an implicit value judgment that is only desirable if one truly considers inequalities in shortfalls and in attainments to be two measures of the same concept.

Both C and GC satisfy the transfer property as long as the health variable is nonnegative. Moreover, GC satisfies mirror but is not scale invariant for any measurement scale (other than for a unique scale). Conversely, C does not satisfy mirror but is scale invariant for ratio-scaled (but not cardinal) variables. As neither C nor GC satisfies all properties for all type of variables, further corrections of C have been proposed.

Recent Corrections of Concentration Index

Definitions

This section presents and discusses three corrections that have recently been suggested. The first correction applies for cardinal variables (bounded or unbounded). Erreygers and van Ourti (2011a) suggest modifying the C as:

$$\text{Modified C} = \frac{\mu}{(\mu - h_{min})}\text{C} = \frac{2}{n(\mu - h_{min})}\sum_{i=1}^{n} h_i(R_i - 1)$$

which is equivalent to computing C of a transformed health variable m_i for which the minimum value are set to zero (i.e., $m_i = h_i - h_{min}$). Thus, this modification of C satisfies scale invariance for cardinal variables (which indirectly also implies that the index satisfies the transfer property even if h_i attains negative values).

The other two corrections are specifically developed for bounded variables and satisfy the mirror condition as well as scale invariance for cardinal variables. Wagstaff (2005) corrects C as:

$$W = \frac{(h_{max} - h_{min})}{(h_{max} - \mu)(\mu - h_{min})}\text{C}$$

and Erreygers (2009a) corrects C as:

$$E = \frac{4}{(h_{max} - h_{min})}\mu\text{C}$$

Although these three corrections of C satisfy scale invariance for cardinal variables, one still cannot directly apply the value judgments from the income inequality literature. Therefore, the next section reviews the recent discussion in the literature of how one may relate to these inequality indices for bounded (cardinal) variables (**Table 1**).

Value Judgments for Bounded Variables

In the ongoing discussion on inequality indices for bounded variables, Erreygers and van Ourti (2011a,b) advocate a redefinition of the relative and absolute value judgments, whereas Wagstaff (2005, 2009, 2011a) suggests an approach that compares

Table 1 Properties of the indices

	Mirror	*Transfer*		*Scale invariance*		
		Nonnegative	*Possibly negative*	*Cardinal*	*Ratio*	*Unique*
C		✓			✓	✓
GC	✓	✓				✓
Mod C		✓	✓	✓	✓	✓
E	✓	✓	✓	✓	✓	✓
W	✓	✓	✓	✓	✓	✓

Abbreviations: C, concentration index; GC, generalized concentration index; Mod C, Modified concentration index; E, Erreygers' correction of C (Erreygers, 2009a); W, Wagstaff's normalization of C (Wagstaff, 2005).

Source: Reproduced from O'Donnell, O., van Doorslaer, E., Wagstaff, A. and Lindelöw, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington, DC: The World Bank.

how far the health distribution is from the most pro-rich state. This section presents the two views, starting with the former.

Scale invariance implies that, without changing the measured degree of inequalities, any bounded health variable can be represented by a standardized health variable h_i^* ranging from zero to one, that is, $h_i^* = (h_i - h_{\min})/(h_{\max} - h_{\min})$. As differences in such a standardized variable always represent real health differences and are not an effect of changing the unit of measurement, Erreygers and van Ourti (2011a) define the value judgment for bounded variables based on inequality preserving changes of this variable. Still, the bounds of the variable act as constraints for some inequality preserving changes, that is, for some health distributions it is technically impossible to add an equal amount of health or to proportionally increase the health for all individuals without exceeding the upper bound of the variable. Erreygers and van Ourti (2011a) therefore redefine the value judgments so that an index embodies a specific value judgment if it is invariant to the corresponding inequality preserving change given that such a change is feasible.

Following this definition, Erreygers' correction of C (E) captures an absolute value judgment; it is invariant to equal increments of the standardized health variable but not to proportional changes. However, for the relative value judgment, the transition to bounded variables is not as straightforward. As (the modified) C is invariant to equiproportionate changes of the standardized variable, it captures a relative value judgment. But C does not satisfy the mirror condition. In fact, Erreygers and van Ourti (2011a) show that it is impossible to combine the mirror condition with a relative value judgment.

Wagstaff's normalization of C (W) satisfies the mirror condition but captures neither a relative nor an absolute value judgment. For an equal increment, W increases if the mean of the standardized health variable is larger than 0.5 but decreases if the mean of the standardized health variable is smaller than 0.5. This seemingly strange and counterintuitive behavior is a result of Wagstaff's solution to what he refers to as 'the bounds issue.' For bounded variables, the maximum and minimum value of C depends on the mean of the health variable; as C tends to one when only one single individual is in possession of all the health available in the society, the most pro-rich society cannot be reached unless there is only one individual in full health. This issue complicates comparisons between populations with different mean health. As a solution, Wagstaff normalizes C by the maximum value of the index (i.e., C of the most pro-rich state possible) given the level of health in the society (see **Figure 3**). Thus, in a society with a population of n individuals where the sum of h_i^* is equal to m, W attains the value of one when the richest m individuals have full health, whereas everyone else has no health. How a health change affects W reflects whether the society moves closer or further away from the most pro-rich state and, consequently, W may be interpreted as the answer to the question of how far the society is from that state (Wagstaff, 2009, 2011a).

Kjellsson and Gerdtham (2013) points out that C and E may also be interpreted as answering a similar question. However, the indices differ in their definition of the most pro-rich state; C attains its maximum value when the richest individual has all the health, and E attains its maximum value when only the upper half of the income distribution have full health and the lower half has no health.

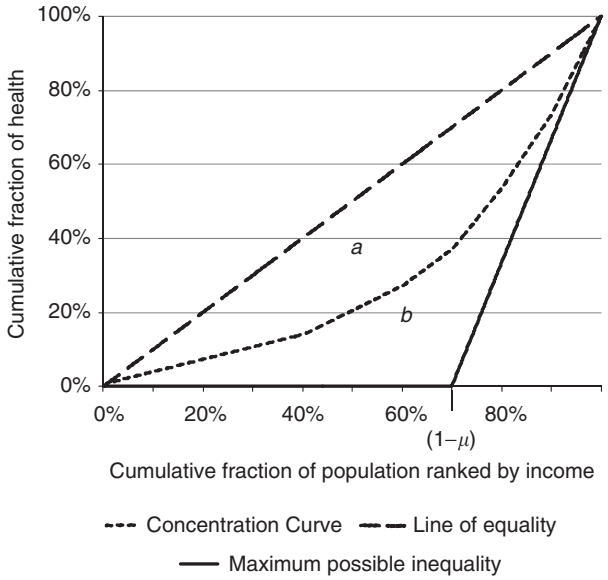


Figure 3 Wagstaff's normalization of C. *Note*: Using a graphical representation, we may define Wagstaff's normalization of C as: $W = C/(1-\mu) = a/(a+b)$.

Table 2 Appropriate indices

	Bounded	*Unbounded*
Unique	E	GC
	W	C
	C	
Ratio	E	C
	W	
	C	
Cardinal	E	Modified C
	W	
	Modified C	
Binary	E	
	W	
	Modified C	

Abbreviations: E, Erreygers' correction of C; GC, generalized concentration index; W, Wagstaff's normalization of C; C, concentration index.

Having reviewed the two approaches of how to measure health inequalities for variables of different measurement properties, the authors are now ready to compile the literature into a guideline for practitioners and follow the guideline in an empirical illustration.

Guidelines for Practitioners

Which Index to Use and When?

Simplifying the guidance from Erreygers and van Ourti (2011a), **Table 2** summarizes the possible choices for a researcher or practitioner depending on the measurement scale and the boundedness of the health variable. To be eligible, the index has to satisfy transfer and scale invariance.

As scale invariance is not an issue for unbounded variables measured on a unique scale, one may apply either GC or C depending on the value judgment that one wants to impose. However, for unbounded variables that are either ratio scaled or cardinal one is constrained to apply a relative value judgment as (the modified) C is the only index that satisfies scale invariance.

For bounded variables, the choice boils down to the following three alternatives. First, one may choose to impose the mirror condition and apply an absolute value judgment by using E to answer the question of how far a society is from a state where the upper half of the income distribution has full health and the lower part has no health. Second, one may take the level of health in the society into account, but depart from a pure relative judgment, by using W to answer the question of how far a society is from a state where the richest m individuals have full health and everyone else has zero health. Third, one may choose to relax the mirror condition, not address the bounds-issue that Wagstaff highlights, and apply a relative value judgment, that is, using (the modified) C. However, applying a relative value judgment for bounded variables requires a decision of whether it is appropriate to measure inequalities in shortfalls or attainments.

The current advice in the literature is to accept that the relative value judgment and mirror condition are incompatible and either use a relative or an absolute value judgment (Erreygers and van Ourti, 2011a,b; Wagstaff, 2011b). Erreygers and van Ourti (2011a,b) advocate the attractiveness of the mirror condition and, thus, prefer E. They also stress that E satisfies two additional possibly desirable properties. The first property is as follows: if starting with an unequal health distribution and gradually decreasing the health of all individuals toward zero (i.e., in the limit all the health of individuals is zero, which implies a perfectly equal distribution), then E tends to zero. Neither C nor W shows this tendency. The second property is: if the health of a rich individual, i.e., an individual from the upper half of the income distribution, increases, then E always increases. Neither C nor W satisfies this property.

Conversely, Kjellsson and Gerdtham (2013) and Wagstaff (2009, 2011a) claim that these two properties are a result of the absolute value judgment. In a recent note, Wagstaff (2011b) also advocates abandoning the mirror property (and thereby also his own correction) for the relative value judgment. However, this literature provides no guidance on the choice between attainments and shortfalls. The bottom line of this discussion is that any index inevitably enforces a value judgment that the researcher needs to be aware of and explicitly consider.

Empirical Illustration

For illustrational purposes, this section uses three health variables from the second wave of SHARE to compute inequality indices that, depending on the measurement properties of the variable, satisfy scale invariance and the transfer property. All the three variables, a health index, out-of-pocket payments, and GP-visits, differ in respect of their measurement properties. For a comparison between these results and the work of the ECuity group (e.g., van Doorslaer and Koolman, 2004; van Doorslaer *et al.*, 2004, 2006) horizontal inequity indices are calculated by indirectly standardizing for age, sex, and, when appropriate, health (see O'Donnell *et al.*, 2008).

The health index is a cardinal variable ranging from 0 (being dead) to 1 (perfect health) and is similar to the HUI in van Doorslaer and Koolman (2004) but is specifically developed for the SHARE-data (Jürges, 2007; Jürges, 2005). **Table 3** shows the value of the indices that satisfy scale invariance and the transfers property for bounded cardinal

Table 3 Socio-economic inequality in health among 13 European countries

Health index

Country	*Mean*	*C(h)*		*C(h)^{HI}*		*W(h)*		*W(h)^{HI}*		*E(h)*		*E(h)^{HI}*		*C(s)*		*C(s)^{HI}*	
		Index	*#*	*Index*	*#*	*Index*	*#*	*Index*	*#*	*Index*	*#*	*Index*	*#*	*Index*	*#*	*Index*	*#*
Austria	0.870	**0.010**	6	**0.008**	3	**0.074**	6	**0.058**	2	**0.034**	6	**0.026**	3	**−0.065**	5	**−0.009**	11
Belgium	0.878	**0.009**	7	**0.007**	6	**0.077**	4	**0.056**	3	**0.033**	7	**0.024**	5	**−0.068**	3	**−0.013**	7
Czech Republic	0.856	**0.011**	3	**0.008**	2	**0.077**	3	**0.054**	4	**0.038**	3	**0.027**	2	**−0.066**	4	**−0.017**	5
Denmark	0.864	**0.021**	1	**0.011**	1	**0.153**	1	**0.081**	1	**0.072**	1	**0.038**	1	**−0.132**	1	**−0.059**	1
France	0.876	**0.009**	8	**0.007**	7	**0.073**	8	**0.054**	5	**0.032**	8	**0.023**	7	**−0.064**	7	**−0.013**	9
Germany	0.869	**0.010**	5	**0.007**	4	**0.074**	5	**0.053**	6	**0.034**	5	**0.024**	4	**−0.065**	6	**−0.014**	6
Greece	0.877	**0.006**	10	**0.003**	11	**0.051**	10	**0.026**	11	**0.022**	10	**0.011**	11	**−0.045**	10	**−0.021**	3
Italy	0.845	**0.007**	9	**0.005**	9	**0.042**	11	**0.032**	10	**0.022**	9	**0.017**	9	**−0.035**	11	**−0.005**	13
Netherlands	0.886	**0.004**	13	0.003	13	**0.039**	12	0.024	12	**0.016**	13	0.009	13	**−0.035**	12	−0.008	12
Poland	0.834	**0.006**	11	0.003	12	**0.036**	13	0.019	13	**0.020**	12	0.010	12	**−0.030**	13	−0.010	10
Spain	0.853	**0.011**	4	**0.007**	5	**0.074**	7	**0.047**	8	**0.037**	4	**0.024**	6	**−0.063**	8	**−0.018**	4
Sweden	0.873	**0.013**	2	**0.006**	8	**0.101**	2	**0.048**	7	**0.045**	2	**0.021**	8	**−0.088**	2	**−0.042**	2
Switzerland	0.902	**0.006**	12	**0.004**	10	**0.059**	9	**0.039**	9	**0.021**	11	**0.014**	10	**−0.053**	9	**−0.013**	8

Notes: C(h), W(h), and E(h) all measure inequalities in attainments while C(s) measures inequalities in shortfalls. HI indicates that the index has been standardized for age and sex. The countries are ranked by the level of inequality (i.e., ranging from the most pro-rich to the most pro-poor). Bold figures indicate a significant result on the 5% level. Indices and standard errors are calculated using the convenient regression method (O'Donnell *et al.*, 2008) and the imputation methods developed for European Survey of Health, Ageing and Retirement (Christelis, 2011).

Source: Reproduced from O'Donnell, O., van Doorslaer, E., Wagstaff, A. and Lindelöw, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington, DC: The World Bank.

Table 4 Socio-economic inequality in health care use among 13 European countries

Country	*Out-of-pocket payment*					*General practitioner (GP)-visits*							
	Mean	*C*		C^{HI}		*Mean*	*C*		C^{HI}		*GC*		GC^{HI}	
		Index	#	*Index*	#		*Index*	#	*Index*	#	*Index*	#	*Index*	#
Austria	336.19	**0.133**	4	**0.202**	2	6.10	**−0.069**	5	−0.022	3	**−0.422**	10	−0.134	6
Belgium	507.71	0.014	9	**0.068**	8	5.92	**−0.075**	6	−0.026	6	**−0.442**	11	−0.154	7
Czech Republic	1809.34	−0.053	11	−0.005	10	4.78	**−0.082**	9	**−0.037**	9	**−0.394**	8	**−0.178**	10
Denmark	2347.76	0.031	6	**0.068**	7	3.20	**−0.101**	12	−0.013	2	**−0.322**	7	−0.042	2
France	107.99	**0.197**	1	**0.234**	1	4.70	**−0.068**	4	**−0.036**	7	**−0.319**	6	**−0.170**	8
Germany	227.83	0.027	7	0.075	6	4.92	**−0.080**	7	−0.037	8	**−0.396**	9	−0.184	11
Greece	366.23	−0.039	10	−0.015	11	3.37	**−0.055**	2	−0.024	4	**−0.186**	4	−0.081	4
Italy	448.40	**0.153**	3	**0.171**	4	7.43	**−0.082**	8	**−0.063**	12	**−0.608**	12	**−0.466**	12
Netherlands	115.68	0.106	5	**0.122**	5	2.62	**−0.060**	3	−0.040	10	**−0.157**	2	−0.105	5
Poland	1106.40	−0.075	12	−0.051	13	5.45	−0.012	1	0.005	1	−0.066	1	0.029	1
Spain	119.14	**0.169**	2	**0.174**	3	6.85	**−0.120**	13	**−0.071**	13	**−0.819**	13	**−0.486**	13
Sweden	3464.79	−0.105	13	−0.044	12	1.86	**−0.089**	10	−0.023	5	**−0.165**	3	−0.044	3
Switzerland	1088.80	0.020	8	0.041	9	2.88	**−0.096**	11	**−0.060**	11	**−0.276**	5	**−0.172**	9

Notes: HI indicates that the index has been standardized for age, sex, and health. The countries are ranked by the index value for either payments or GP-visits (i.e., highest value equals rank 1). Bold figures indicate a significant result on the 5% level. Indices and standard errors are calculated using the convenient regression method (O'Donnell *et al.*, 2008) and the imputation methods developed for SHARE (Christelis, 2011).

Source: Reproduced from O'Donnell, O., van Doorslaer, E., Wagstaff, A. and Lindelöw, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington, DC: The World Bank.

variables, i.e., (the modified) C of both shortfalls and attainments as well as W and E. As seen, the reranking of countries between the different inequality index is limited. Generally, the ranking varies less between the inequality indices when there is less variation in average health between countries. However, high average health generates a pattern where the ranking diverge into two groups: C of shortfalls and W; and C of attainments and E. A similar pattern would appear if one was to apply all four indices to 1996 European Community Household Panel in van Doorslaer and Koolman (2004). van Doorslaer and van Ourti (2011) confirm that the ranking is similar for C of attainments and E, whereas the authors encourage the reader to verify for oneself that using W or C of shortfalls will rerank countries in a similar manner as in their example. This reranking stresses on the importance of being aware of the value judgment that a particular index implies.

According to the nonstandardized indices, Denmark, Sweden, and the Czech Republic are ranked as the three most unequal countries, whereas the Netherlands, Poland, Greece, Switzerland, and Italy are the least unequal. However, when accounting for the demographics (i.e., standardize for age and sex), Sweden becomes relatively less unequal whereas Austria moves in the opposite direction. The positions of two of the extremes, Denmark and the Netherlands, are consistent with the findings in van Doorslaer and Koolman (2004).

As out-of-pocket payments, which is the sum of all the individuals' out-of-pocket health spending (excluding insurance premiums) is a ratio-scaled variable, C is the only index that satisfies scale invariance and the transfer property. Therefore, one can only apply a relative value judgment. The results (Table 4) show that, except for Sweden, Poland, Greece, and the Czech Republic, the richer individuals pay a larger fraction of the out-of-pocket payment. The standardization increases the index for all countries, that is, the fraction the poorer individuals pay decreases when controlling for need (i.e., standardizing for age, sex, and the health index).

As the number of GP-visits is measured on a unique scale, it implies that both C and GC may be applied as both indices satisfy the necessary properties. The inequality indices of a utilization variable such as GP-visits measure inequality in access to care. When standardizing for age, sex, and health, the interpretation of the index is a measure of horizontal inequity. The overall tendency of the results (Table 4) is that the indices are negative even after controlling for the above, that is, there is a pro-poor discrimination of access to care. The notable differences between the rankings of C and GC again stress the importance of considering the value judgment. However, regardless of the value judgment, the pro-poor discrimination appears to be strongest in Spain and weakest in Poland. Although the results overall differ to some extent, the finding of Spain having the strongest pro-poor discrimination is in line with the work of the ECuity group (van Doorslaer *et al.*, 2004, 2006).

Conclusion

This article reviews the recent literature on measuring socio-economic health inequalities using the concentration index approach. The authors have briefly discussed when the different corrections of C are appropriate to use depending on the measurement properties of the health variable and value judgment one wants to impose. For an in-depth discussion of the topic see the articles in the further reading list.

Acknowledgments

Financial support from the Swedish Council for Working Life and Social Research (FAS) (dnr 2007-0318) is gratefully

acknowledged. The Health Economics Program (HEP) at Lund University also receives core funding from FAS (dnr. 2006-1660), the Government Grant for Clinical Research ("ALF"), and Region Skåne (Gerdtham). This paper uses data from SHARE release 2.3.1, as of 29 July 2010. SHARE-data collection in 2004–07 was primarily funded by the European Commission through its 5th and 6th framework programs (project numbers QLK6-CT-2001-00360; RII-CT-2006-062193; CIT5-CT-2005-028857). Additional funding from the US National Institute on Aging (grant numbers U01 AG09740-13S2; P01 AG005842; P01 AG08291; P30 AG12815; Y1-AG-4553-01; OGHA 04-064; R21 AG025169) as well as by various national sources is gratefully acknowledged (see http://www.share-project.org for a full list of funding institutions).

See also: Measuring Equality and Equity in Health and Health Care. Measuring Vertical Inequity in the Delivery of Healthcare

References

Christelis, D. (2011). Imputation of missing data in waves 1 and 2 of SHARE. Available at: http://www.share-project.org/t3/share/fileadmin/pdf_documentation/Imputation_of_Missing_Data_in_Waves_1_and_2_of_SHARE.pdf (accessed 01.07.11).

Clarke, P., Gerdtham, U. G., Johannesson, M., Bingefors, K. and Smith, L. (2002). On the measurement of relative and absolute income-related health inequality. *Social Science and Medicine* **55**, 1923–1928.

van Doorslaer, E. and Koolman, X. (2004). Explaining the differences in income-related health inequalities across European countries. *Health Economics* **13**, 609–628.

van Doorslaer, E., Koolman, X. and Jones, A. M. (2004). Explaining income-related inequalities in doctor utilisation in Europe. *Health Economics* **13**, 629–647.

van Doorslaer, E., Masseria, C. and Koolman, X. (2006). Inequalities in access to medical care by income in developed countries. *Canadian Medical Association Journal* **174**, 177–183.

van Doorslaer, E. and van Ourti, T. (2011). Measuring inequality and inequity in health and health care. In Smith, P. and Glied, S. (eds.) *The Oxford handbook of health economics*, ch. 35, pp. 837–869. Oxford: Oxford University Press.

Erreygers, G. (2009a). Correcting the concentration index. *Journal of Health Economics* **28**, 504–515.

Erreygers, G. and van Ourti, T. (2011a). Measuring socioeconomic inequality in health, health care and health financing by means of rank-dependent indices: A recipe for good practice. *Journal of Health Economics*, doi:10.1016/j.jhealeco.2011.04.004.

Erreygers, G. and van Ourti, T. (2011b). Putting the cart before the horse. Comment on "The concentration index of a binary outcome revisited". *Health Economics* **20**, 1161–1165.

Jürges, H. (2005). Computing a comparable health index. In Börsch-Supan, A., Brugiavini, A., Jürges, H., et al. (eds.) Health, ageing and retirement in Europe – First results from the Survey of Health, Ageing and Retirement in Europe, p. 357. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Jürges, H. (2007). True health vs response styles: Exploring cross-country differences in self-reported health. *Health Economics* **16**, 163–178.

Kjellsson, G. and Gerdtham, U.-G. (2013). On correcting the concentration index for binary variables. *Journal of Health Economics*. Available at: http://dx.doi.org/10.1016/j.jhealeco.2012.10.012 (accessed 15.07.13).

Kolm, S. C. (1976). Unequal inequalities II. *Journal of Economic Theory* **12**, 416–442.

O'Donnell, O., van Doorslaer, E., Wagstaff, A. and Lindelöw, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington DC: The World Bank.

Wagstaff, A. (2005). The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality. *Health Economics* **14**, 429–432.

Wagstaff, A. (2009). Correcting the concentration index: A comment. *Journal of Health Economics* **28**, 516–520.

Wagstaff, A. (2011a). The concentration index of binary outcome revisited. *Health Economics* **20**, 1155–1160.

Wagstaff, A. (2011b). Reply to Guido Erreygers and Tom Van Ourti's comment on "The concentration index of a binary outcome revisited". *Health Economics* **20**, 1166–1168.

Further Reading

Erreygers, G. (2009b). Correcting the concentration index: A reply to Wagstaff. *Journal of Health Economics* **28**, 521–524.

Measuring Vertical Inequity in the Delivery of Healthcare

L Vallejo-Torres and S Morris, University College London, London, UK

Glossary

Concentration index A measure of the degree of income-related inequality in health. Where there is no income-related inequality, the concentration index is zero. A negative value indicates a disproportionate concentration of ill-health among the poor.

Equity It relates in general to ethical judgments about the fairness of the distribution of things such as income and wealth, cost and benefit, access to health services, exposure to health-threatening hazards and so on, not necessarily to be identified with equality or egalitarianism. Although not the same as 'equality', for some people, equity frequently involves the equality of something (such as opportunity, health, access).

Horizontal equity It refers to treating equally those who are equal in some morally relevant sense. Horizontal equity principles include 'equal treatment for equal need' and 'equal treatment for equal deservingness'. Applied to insurance, the notion that two individuals facing the same risks should have access to the same coverage at the same premium.

Inequity It refers to treating unequally those who are equal in some morally relevant sense or treating equally those who are unequal in some morally relevant sense.

Need The most frequently met practical measures of need at the community level are morbidity and mortality data. They plainly imply a need for health though not necessarily a need for health care (which may not be effective in altering either for the better). Other concepts include capacity to benefit from health care and the resources that are necessary to reduce capacity to benefit to zero (i.e. to the point at which the marginal benefit falls to zero).

Resource allocation It refers to societal or individual decisions about the equitable distribution of available resources.

Socioeconomic status A description of a person or group of people having a similar social, political and economic position in society.

Vertical equity It refers to treating unequally those who are unequal in some morally relevant sense. Vertical equity principles include 'higher contributions from those with greater ability to pay', 'more resources for those with greater need'.

Introduction

Equity in the delivery of health care is an important policy objective in many countries, and some, such as Australia, Canada, Sweden, and the UK, distribute healthcare resources on the basis of explicit equity objectives. Such objectives often subscribe to egalitarian goals, which suggest that health care should be distributed according to need and financed according to ability to pay.

Egalitarian goals can include horizontal and/or vertical equity principles. The horizontal equity principle requires that individuals with the same needs receive the same treatment. The vertical equity principle requires that those with different needs receive appropriately different treatment. Taken together, these principles suggest not only that patients with the same needs should receive the same treatment irrespective of, for instance, their social class or place of residence, but also that those with greater needs should be appropriately prioritized in receiving health care.

In the literature, little attention has been paid to vertical equity in the delivery of health care. This is probably because measuring vertical equity requires strong value judgments regarding the way healthcare delivery ought to vary amongst individuals with different levels of need. Most empirical work considers horizontal equity, though the importance of vertical equity is increasingly being emphasized. When considering the role that health care might play in reducing inequalities in health, some authors have argued that accounting for vertical equity in the delivery of health care addresses health inequalities that will not be addressed by focusing only on horizontal equity. For example, it has been suggested that horizontal equity is not relevant when dealing with individuals with substantial differences in health status. The Marmot review (Marmot, 2010) concluded that in order to reduce the steepness of the social gradient in health, ''actions must be universal, but with a scale and intensity that is proportionate to the level of disadvantage.'' The review named this principle 'proportionate universalism,' which is related to the principle of vertical equity.

Crucial to the measurement of vertical equity are measures of 'healthcare delivery' and 'need.' 'Healthcare delivery' is a broad term that can be used to refer to the receipt of treatment, the use of, or access to, healthcare services, or to the allocation of healthcare resources between individuals or areas. In economic studies of equity, healthcare delivery typically refers to the use of healthcare services by individuals or to the allocation of healthcare resources to areas.

Although a wide variety of definitions of need have been developed, economists often define need in terms of 'capacity to benefit' – the ability to benefit from healthcare provision. In empirical studies, however, need is usually defined in terms of ill-health, where people who are ill are deemed to have greater need than those who are not. This is an imperfect measure, because unlike capacity to benefit, it does not account for the

instrumentality of need, i.e., that needs for health care when ill only exist if there is health care available that can improve health. Although it is a limitation, this definition is used for pragmatic reasons in that measures of ill-health are often directly available in datasets used to measure equity, whereas measures of capacity to benefit are not available.

Another limitation, also commonly adopted for pragmatic reasons, is that empirical studies usually measure ill-health and healthcare delivery contemporaneously, when ideally need would be measured prior to utilization so that its causal impact on utilization could be assessed.

When investigating inequity in healthcare delivery, it is common to distinguish between 'need variables' which ought to affect healthcare delivery and 'nonneed variables' which ought not to. Inequality in healthcare delivery can be associated with both need and nonneed variables. There is horizontal inequity when healthcare delivery is affected by nonneed variables, so that individuals with the same needs consume different amounts of care. There is vertical equity when individuals with different levels of need consume appropriately different amounts of healthcare. The categorization of variables, such as age, gender, income, and ill-health, as need or nonneed variables is crucial to testing for horizontal and vertical inequity. The measurement of horizontal and vertical equity rests on value judgments as to what are need and nonneed variables and what constitutes an appropriate level of health care.

Notably, there are methodological challenges associated with measuring vertical equity, and as a result few studies have investigated this issue. Among these studies, there is considerable variation in the methods used and in the assumptions underpinning the analyses. In this article, these methods are reviewed and appraised; approaches taken outside the field of health care to measure vertical equity are also examined.

A Simple Model of Vertical Equity

Gravelle *et al.* (2006) have described the conditions for the identification of and distinction between vertical and horizontal inequity in healthcare delivery, and have highlighted the main challenges faced by economic studies of equity in health care. They place equity analyses in the context of welfare maximization. For example, let v be the individual welfare accruing to individual i from the consumption of health care q, from k need characteristics N, and from the cost of accessing health care, c:

$$v_i = v(q_i, N_{ki}, c_i) \quad [1]$$

The aim of the policy maker is to enable individuals to choose utilization levels q that maximize an aggregate welfare function W subject to the constraint that total utilization cannot exceed total supply S:

$$W = \sum_i v(q_i, N_{ki}, c) \quad s.t. \sum_i q_i \leq S \quad [2]$$

In the optimal allocation, which meets the horizontal and vertical equity principles, healthcare consumption is not affected by nonneed variables, and the effects of the need variables on consumption reflect the appropriate difference in treatment that individuals with different levels of need ought to receive. Consider a model of actual health care use given by

$$q_i = \alpha + \sum_k \beta_k N_{ik} + \sum_j \delta_j Y_{ij} + u_i \quad [3]$$

where Y_j denotes a set of j nonneed variables that ought not to affect healthcare use. The condition for horizontal equity is $\delta_j = 0$, i.e., use is not affected by nonneed variables. The conditions for vertical equity are $\beta_k = \beta_k^*$ and $\alpha = \alpha^*$, where β_k^* and α^* denote the appropriate effect of the need variables on healthcare use, however, defined, and the optimum base level of consumption from the optimal healthcare allocation, respectively.

Vertical Equity in Healthcare Delivery

In this section, different approaches that have been taken to measure vertical equity in healthcare delivery are reviewed. Following Gravelle *et al.* (2006), the criteria for assessing the different approaches are based on: distinguishing between need and nonneed variables; testing the potential impact of omitted variables; disentangling horizontal and vertical equity; and measuring the extent of vertical inequity. The different approaches to defining the appropriate way in which healthcare consumption ought to vary for individuals with different levels of needs is also considered.

Separation between need and nonneed variables depends largely on value judgments on what counts as a need variable and what counts as a nonneed variable. It is commonly accepted that measures of health status and morbidity ought to affect healthcare use. In individual-level analyses with health and morbidity data, socioeconomic indicators are generally considered to be nonneed indicators. In the case that needs are not comprehensively measured, socioeconomic indicators may be picking up the effects of unobserved need factors, such as unmeasured severity levels. In that case, the analysis would be affected by an omitted variable problem. Although the extent to which needs are captured depends largely on the availability of the data, one criteria for assessing the different approaches is the ability of such approaches to account for needs comprehensively.

The exploration of vertical equity requires estimating the appropriate way in which healthcare consumption ought to vary for individuals with different levels of needs. Without the knowledge about the optimal effect of needs on healthcare delivery, conclusions about whether individuals with different needs are being appropriately treated cannot be made. In addition, the separation between vertical and horizontal aspects is not straightforward. This is because both need and nonneed variables are likely to be related. For instance, if on average, healthy individuals consume more health care than they ought to (i.e., there is evidence of vertical inequity favoring the relatively healthy – prohealthy vertical inequity) and there is a positive correlation between health and income, then probably on average, richer individuals will consume more health care than they ought to; there is horizontal inequity favoring the rich (prorich horizontal inequity). More generally, if health and income

Table 1 Approaches used to measure vertical equity in the delivery of health care and in other fields

No	*Methods*	*Metric*	*Control for nonneed variables?*	*Disentangle HI and VI aspects?*	*Allow quantification of VI?*
Healthcare delivery					
1.	Test the association between SES and healthcare delivery	Unadjusted odd ratio Ratio analysis Concentration curves Adjusted odd ratios Correlation coefficient	No	No	No
2.	Compare ranking of observations according to both needs and healthcare delivery	Coefficient of concordance	No	No	No
3.	Test the association of need and healthcare delivery after controlling for SES	Regression coefficient	Yes	Yes	No
4.	Test the association between a nonneed factor and healthcare delivery at different levels of needs	Interaction term	Yes	No	No
5.	Test the association between health outcomes and healthcare delivery across a nonneed factor	Adjusted odd ratios	Yes	No	No
6.	Compare actual and target effect of the need indicators on healthcare delivery	NA	Yes	Yes	No
7.	Compute healthcare gaps between actual and target healthcare delivery	Poverty index	Yes	No	No
8.	Measure the difference in the allocation based on the target healthcare delivery and on the need-expected healthcare delivery with respect to SES	Concentration index	Yes	Yes	Yes
Other fields					
1.	Healthcare finance: Measure the difference between the concentration of payments with respect to prepayment income and the Gini coefficient of prepayment income	Kakwani index	Yes	Yes	No
2.	School funding: Observations are weighted by the inverse of a need characteristic and variations in per pupil revenues are compared before and after the weighting is applied	Weighted dispersion index	Yes	Yes	No

Abbreviations: HI, horizontal inequity; NA, not applicable; SES, socioeconomic status; VI, vertical inequity.

are positively correlated, prohealthy vertical inequity will tend to benefit those on higher incomes, leading to prorich horizontal inequity. Conversely, propoor horizontal inequity will tend to mean that the sick have higher than expected levels of use, leading to prosick vertical inequity. Therefore, separation of vertical and horizontal inequity aspects is an important challenge in measuring vertical equity in healthcare delivery.

Finally, the simple model described in Section A Simple Model of Vertical Equity can be used to identify vertical and horizontal inequities in healthcare delivery, but it does not allow measurement of the extent of inequity; it is not possible to measure this purely on the basis of the size of α and β in eqn [3]. However, this is of interest because it permits comparisons over time and between areas, which is helpful both to identify trends and for policy evaluation.

Approaches to Measuring Vertical Equity in Healthcare Delivery

The following eight approaches, summarized in Table 1, have been used to measure vertical equity in healthcare delivery. Each of them has been explained, moving from the simplest to the more complex, and the limitations of each approach are assessed.

Approach 1: The association between socioeconomic status and healthcare delivery

This approach assumes that vertical equity is identified by a positive association between socioeconomic status (SES) and the delivery of health care. The assumption underpinning this approach is that individuals in lower socioeconomic groups

have higher needs and they should therefore receive more health care in order to meet the vertical equity principle.

Let q_i denote the quantity of healthcare delivery and Y_i the SES for individual i. Higher values of q and Y denote greater healthcare use or resources and higher SES, respectively. Note that in this approach Y is used as a proxy for need and not as a nonneed variable. The test for vertical equity is based on the relationship between q_i and Y_i:

$$q_i = \alpha + \delta Y_i + \varepsilon_i \quad [4]$$

$$\frac{\partial q_i}{\partial Y_i} = \delta < 0 \quad [5]$$

There is vertical equity if SES is negatively correlated with healthcare use (i.e., $\delta < 0$). Among the studies that have used this approach, the relationship between SES and healthcare delivery has been explored by different bivariate measures such as unadjusted odd ratios, ratio analysis, concentration curves, and correlation coefficients (Table 1).

In the absence of good epidemiological data, area-level analyses often rely on socioeconomic indicators as need variables. However, the choice of SES as a need variable is contested, and in many other studies, SES is defined as a nonneed variable. Although the correlation between SES and health is well documented, it does not imply that differences in SES will only be reflecting differences in needs. Moreover, there may be needs that are not correlated with SES that will not be picked up by an analysis of this kind. Therefore, the interpretation of the association between SES and healthcare delivery is ambiguous. These analyses are also not able to distinguish between vertical or horizontal inequity, as they cannot judge whether individuals receive different amounts of health care because of their different needs or because of same needs but different nonneed variables (e.g., SES). Even if SES was an appropriate need variable, this type of analysis cannot identify whether or not the differences in treatment received by those in lower SES is appropriate to meet their relatively higher needs. Nor can it measure the extent of vertical equity. Therefore, this approach is of limited use for analyzing vertical equity in healthcare delivery.

Approach 2: Comparison of the ranking of observations according to both needs and healthcare delivery

This approach assumes that vertical equity is identified by a positive correlation between the ranking of healthcare delivery and that of need. The method is thus based on a comparison of the hierarchy of observations when ranked according to both needs and the delivery of health care received. For example, this approach involves creating a need index based on health status, ranking observations with respect to need, and comparing this ranking with another ranking based on some measure of healthcare delivery. A measure such as Kendall's coefficient of concordance between the two rankings could be used to indicate vertical inequity.

Inappropriate ranking in the allocation of health care with respect to needs could be due to deviation from either the horizontal or the vertical equity principles. Therefore, this method cannot distinguish between horizontal and vertical aspects of inequity, nor does it control for nonneed factors. Furthermore, although the method provides a framework for testing if the delivery of health care is ordinally appropriate, it fails to account for whether or not the allocation is cardinally appropriate. The measurement of vertical equity requires that the size of the differences in healthcare delivery between observations is sufficient to account for their relative differences in needs. Hence, this method is also of limited use for measuring vertical equity in healthcare delivery.

Approach 3: The association between need and healthcare delivery

This approach assumes that vertical equity is identified by a positive correlation between healthcare delivery and need variables, usually measured by one or more health measures, after controlling for nonneed variables, commonly measured in terms of SES. It focuses on the idea that a positive association between healthcare delivery and ill-health is a necessary condition for vertical equity. Following from eqn [4], let N denote a measure of ill-health for individuals i, so that the test for vertical equity in healthcare delivery is that $\beta > 0$, i.e., the need variable has a positive association with use:

$$q_i = \alpha + \beta N_i + \delta Y_i + \varepsilon_i \quad [6]$$

$$\frac{\partial q_i}{\partial N_i} = \beta > 0 \quad [7]$$

Most studies using this approach have tried to incorporate some assessment of vertical equity by looking at the coefficients of the need variables that have been included in their regression models testing for horizontal inequity.

The approach is a simplified version of the test for vertical equity presented in eqn [3]. The main limitation of this approach is that it cannot discern whether or not the higher levels of use by those with higher levels of need adequately meets their relative need when compared with the healthy; hence the condition described by eqn [7] is at best a necessary but not a sufficient condition for vertical equity. Moreover, it is not possible to measure the extent of vertical inequity using this approach.

Approach 4: The effect of socioeconomic status on healthcare use at different level of needs

This approach assumes that vertical equity is identified by differences in the impact of nonneed variables on healthcare use in groups with different needs. One approach that has been used is to explore whether or not nonneed variables affect healthcare delivery at different levels of health. This is achieved by the interaction between need and nonneed variables, for example, by extending eqn [6] to

$$q_i = \alpha + \beta N_i + \delta Y_i + \sigma N_i \times Y_i + \varepsilon_i \quad [8]$$

In this approach, there is said to be vertical equity if $\beta > 0$ and $\sigma = 0$. An alternative approach would be to run separate models for groups with different levels of nonneed variables and testing if the impact (β) of the need variables on healthcare delivery is the same for every group. This is a test for vertical equity because differences in healthcare delivery between groups with different levels of need cannot be regarded as appropriate as long as they are affected by differences in nonneed characteristics, such as income. An extension to this

approach involves including an additional condition that in order to meet the vertical equity principle, the response of utilization to need in every SES group should be the same as that observed in a predefined reference group, which is thought to be (more) vertically equitable.

This approach to measuring vertical inequity is problematic because it has also been proposed in the literature as a means of testing for horizontal inequity, on the grounds that, for example, if sick individuals when they are rich receive more health care than the sick when they are poor, the horizontal equity principle is not met. Significant interactions between need and nonneed variables cannot be separated into horizontal and vertical aspects, so this approach cannot be used to test for either vertical equity or horizontal equity in isolation.

Approach 5: Health outcomes derived from unequal treatment across nonneed groups

This approach assumes that vertical equity is identified by a significant association between healthcare use and a nonneed indicator, but with the same health outcomes in different nonneed groups. The idea behind the approach is that there is vertically equity if different groups, defined in terms of one or more nonneed variable, receive different levels of healthcare delivery and so are treated unequally, but achieve the same health outcomes. In the general framework, where H_i stands for the health outcome of individual i:

$$q_i = \alpha_0 + \delta_0 Y_i + \varepsilon_i \quad [9]$$

$$H_i = \alpha_1 + \delta_1 Y_i + \varepsilon_i \quad [10]$$

the delivery of health care is considered to be vertically equitable if $\delta_0 \neq 0$ in eqn [9], and $\delta_1 = 0$ in eqn [10]; therefore, different SES groups are treated differently, but their outcomes are the same.

One limitation of this approach is that differences in health outcomes are assumed to be a result of differences in the treatment received. There may be a range of reasons why individuals receiving different treatment end up having same health outcomes that do not relate to their treatment but to other factors such as inefficiencies in the provision of health care or to nonhealth-care factors, for example, other social determinants of health. This could be tested directly by including q as an additional covariate on the right hand side of eqn [10]. In addition, and similarly to previous methods, this method is not able to quantify the extent of vertical inequity.

Approach 6: Comparing the actual effect of need indicators on use with the target effect of need indicators

This approach assumes that vertical equity is identified by need variables having the appropriate effect on healthcare use. Studies using this approach search for the appropriate effect of the need variables such that individuals with unequal needs receive appropriately unequal treatment. Once the appropriate effects, for example, in terms of regression coefficients, have been derived, they can be compared with the actual effects in order to assess whether the allocation was vertically equitable. Therefore, based on a model such as the one described in eqn [6], this approach tests whether or not the estimated effect of the need variables on healthcare delivery equals the target effect, i.e., if $\hat{\beta} = \beta^*$, where β^* is the target effect of the need variable.

Approaches to identifying values of β^* could include eliciting society's preferences, calculating the actual effect of the need variables on healthcare delivery in different subgroups of the population, and using the largest value as the target effect.

This approach provides the basis for a test of vertical inequity as described by Gravelle *et al.* (2006) in eqn [3]. However, it is not capable of measuring the extent of vertical equity. It therefore precludes the quantification of inequity over time and/or between areas and does not assist in monitoring efforts to reduce vertical inequity.

Approach 7: Healthcare gap between actual and target health care

This approach assumes that vertical equity is identified by the distribution of healthcare gaps (HCG), which are defined as the distance between the target level of healthcare delivery and the actual level of healthcare delivery. The target level of healthcare delivery might be exogenously set by policy makers as the minimum level of healthcare delivery that individuals or areas should receive given their levels of need. For example, HCGs x for each individual i are given by:

$$x_i = \max(q_i^* - q_i; 0) \quad [11]$$

where, q_i^* is target healthcare use and q_i is actual healthcare use. $x = 0$ when $q \geq q*$, i.e., the HCG has a zero value when individuals receive more health care than the targeted level (the focus of the analysis needs to be reversed to consider the distribution of individuals receiving more than targeted level of health care). The HCGs can be combined across individuals or areas, for example, using poverty indices or standard inequality measures, making value judgments regarding the relative weight of the HCGs in different groups, to provide an aggregate measure of deviations from target levels of healthcare delivery.

The main limitation of this approach, in terms of measuring vertical inequity, is that it is not capable of disentangling horizontal and vertical inequity aspects because the difference between the target level of healthcare delivery and the actual level is affected by deviations from both vertical and horizontal inequity principles. Moreover, this measure only captures healthcare inequity among individuals receiving less than the target level of health care unless we reverse the focus and consider individuals receiving more than the target level. This implies that situations in which individuals receive more than their target level of health care would not be deemed inequitable. Therefore, it is not possible to derive a measure that considers both sides simultaneously and provide a meaningful estimate of the extent and direction of inequity.

Approach 8: Measuring the difference between target and need-expected healthcare delivery across socioeconomic status

This approach assumes that vertical equity is identified and measured by the difference between the distribution of the target and the need-predicted health care use with respect to

SES. The approach applies similar methods to those now widely used to measure horizontal inequity using concentration indices. Let $\hat{q}_i$ denote the predicted value of healthcare delivery from eqn [6] based on the estimated effect of the need variables ($\hat{\beta}$); and $\hat{q}_i^*$ the predicted values of healthcare delivery based on the target effect of the need variables (β^*), however defined. In both equations, SES is set equal to the mean value $\overline{Y}$ in order to neutralize its effect (as a nonneed variable) in the prediction.

$$\hat{q}_i = \hat{\alpha} + \hat{\beta} N_i + \hat{\delta}\overline{Y} \quad [12]$$

$$\hat{q}_i^* = \alpha^* + \beta^* N_i + \hat{\delta}\overline{Y} \quad [13]$$

Equation [12] gives the need-expected (also referred to as need-predicted) allocation of health care; eqn [13] gives the target allocation of health care based on the optimal effect of the need variables and the intercept; α^*, β^*. Sutton (2002) has proposed this methodology to measure the extent to which the gap between the target and the need-expected allocation falls disproportionally on specific SES groups. The target allocation of health care is created by imposing across the whole need distribution the (strictly positive) level of β found in the subpopulation with the lowest levels of need. The method involves computing the concentration index (CI) of the need-predicted and target allocation of health care with respect to SES. The estimate of vertical inequity is the difference between the two. Following Wagstaff (2002), the formula for the CI of socioeconomic inequality can be written as follows:

$$CI = 1 - (2*(1 - R_i)) \sum_{i=1}^{n} \frac{q_i}{Q} \quad [14]$$

where Q is the total healthcare use across the sample and $R_i = i/n$ is the fractional rank in the income distribution of the ith person, with $i = 1$ for the poorest and $i = n$ for the richest. Therefore, the CI is one minus the weighted sum of the share of the healthcare variable of each observation, where the weight is given by the position of the individual in the SES distribution of that population. The CI provides a summary measure of the magnitude of socioeconomic-related inequality in a health variable of interest, and by comparing a set of indices one can derive a clearer ranking when trying to compare inequality across a number of countries, regions or time periods.

Vertical equity is then measured as the difference between the CI of the need-predicted healthcare allocation and the CI of the target allocation, i.e., the divergence in the allocation of health care that relates only to the difference between the actual effect and the appropriate effect of the need variables:

$$VI = CI_{\hat{q}_i} - CI_{q_i^*} \quad [15]$$

These methods control for nonneed indicators in order to appropriately separate the effect of need factors; they provide the comparison between the actual and the target effect of the need variables; and, in particular, they allow for the measurement of vertical inequity by looking at the distributional consequences across the income distribution. However, the focus of this approach is on the measurement of socioeconomic-related vertical equity in healthcare delivery, which although of interest, may be only part of the vertical inequity which is present in a healthcare system. The reason for this is that vertical inequity arises when individuals with unequal needs do not receive appropriately unequal treatment, and this definition does not rely on the inequity being identified with respect to the socioeconomic dimension solely. Thus, this approach measures what Gravelle *et al.* (2006) described as the consequences of vertical equity for the groups identified by horizontal inequity, i.e., across the socioeconomic distribution.

As mentioned above, Sutton (2002) derived the target allocation by imposing the value of β found in one part of the health distribution (among the healthy) on to respondents across the whole health distribution. This assumes that the relationship between changes in ill-health and changes in use among the unhealthy ought to be the same as this relationship among the healthy. The underlying requirement for choosing this target was that the effect of the need variable ought to be positive across the full range of the health distribution. The imposition of a strictly positive effect of need on utilization may not be appropriate for specific types of services or patients.

Adapting Measures Used in Other Fields

Approaches to measuring vertical equity have also been considered in other fields, including: healthcare financing; poverty alleviation programs; the transport sector; aid allocation; and, education funding programs.

Most of these methods are of limited usefulness for measuring vertical equity in the delivery of health care. In the case of poverty alleviation programs, transport sector policies and part of the methodology for measuring vertical inequity in finance, the focus is on relative measures, assessing whether or not a variable was redistributed in a more or less vertically equitable way following a policy change. These methods could be applied to assess the relative impact of policies to reduce vertical inequity in healthcare delivery, but do not provide a static measure. A potential difficulty with adapting these approaches to our context is that they require an assessment of the extent to which a particular healthcare policy contributes to the observed redistribution of healthcare delivery.

There are two measures that could be used to capture vertical equity in healthcare delivery – Kakwani's progressivity index and the ratio of the estimated coefficient to the optimal coefficient of the need indicator. The details of these approaches are summarized in Table 1. Kakwani's progressivity index, which is widely used in the tax and healthcare finance literature, focuses on the measurement of how far health care is financed according to ability to pay. It is defined as the difference between the CI of payments with respect to income, and the Gini coefficient for prepayment income:

$$\Pi_k = CI_{\text{payments}_i} - G_{\text{preincome}} \quad [16]$$

where the CI is defined as in eqn [14], but now q_i represents healthcare payments and R_i is the fractional rank in the prepayment income distribution of the ith person. The Gini coefficient G is analogous to this index where q_i stands for the prepayment income. The Kakwani index equals zero if

payments as a proportion of income are constant across the income distribution; if payments as a proportion of income increase with income, the index is positive, and the finance source is considered to be vertically equitable or progressive.

This index could be used to measure the extent to which healthcare delivery as a proportion of need increases with needs. However, it could only discern whether or not healthcare delivery is 'progressive.' It is not capable of assessing whether the system is 'responsive enough' or whether it 'overmeets' the needs of the population being served. In this sense, it is similar to Approach 3 described above, with the same limitations.

Applying Toutkoushian's and Michael's (2007) method developed for the context of school funding, vertical equity in healthcare delivery could be measured as the ratio of the estimated coefficient, $\hat{\beta}$, to the optimal coefficient of the need factors, β^*, in healthcare utilization equations, such as eqns [12] and [13]. Vertical equity would be achieved when VE defined below equals 100%.

$$VE = \frac{\hat{\beta}}{\beta^*} \times 100\% \qquad [17]$$

Toutkoushian and Michael do not provide a method for estimating the target effect of the need variables, other than suggesting (in an education funding context) to use the monetary amounts prescribed by the state's foundation program in the setting of per pupil revenues. In the context of the delivery of health care, this would be similar to obtaining policy makers' or medical experts' opinion about how health service use ought to increase with needs. If this information were available, this method would permit a summary in one measure of how far the estimated coefficient is from the optimal effect. The ratio could also be compared over time and across different geographies. As with Approach 6 described above, this ratio does not measure the redistributive impact of the difference between the estimated and target allocation across the whole distribution, because it is focused on only what happens on average in a population.

Conclusions and Implications

In this article, different approaches to measuring vertical equity in healthcare delivery have been described. At the outset, it has been noted that although vertical equity considerations seem to be gaining momentum in the context of addressing inequalities and inequities in health and health care, vertical inequity analysis are rarely undertaken. Therefore, existing techniques to investigate vertical inequity have been appraised and assessed, and areas suitable for further work have been identified.

Methods were classified into different approaches and the validity of each was assessed. Approaches used to measure vertical equity outside the field of healthcare delivery were also explored, though none of these approaches was considered to provide any advantage over the methods already used in the field.

Of all the methods considered, Sutton's (2002) approach using CI techniques was found to provide the most comprehensive analysis of vertical equity in the delivery of health care. However, this approach was developed to measure socioeconomic-related vertical equity only. Emphasis on the socioeconomic dimension of inequity has been the norm in the analyses of horizontal inequity in health care, which is appropriate given that the aim is usually to identify systematic variations in the treatment of those with equal needs but are from different socioeconomic backgrounds. However, in the context of vertical inequity, this approach is possibly rather restrictive. Vertical inequity arises when healthcare delivery is not allocated appropriately according to differences in needs. This definition therefore does not require inequity to be measured with respect to a socioeconomic dimension, but emphasizes the need dimension. Further work is necessary to extend this methodology for ensuring that the consequences across the need distribution have been accounted for. This could be accomplished, for example, by computing the concentration indices with respect to the needs variable rather than with respect to SES.

Quantifying the extent to which healthcare delivery ought to vary at different levels of needs is a major challenge when investigating vertical inequity. In the absence of information from policy makers regarding how much more high need individuals ought to receive as compared with those with lower needs, the literature on vertical equity has suggested different approaches. These include asking the community, identifying areas within a country that are more responsive to the needs of their populations, and imposing the positive effect of ill-health on use as found in one part of the health distribution (where the effect has been found to be strictly positive) onto respondents across the whole health distribution. One alternative approach could involve using external evidence regarding subgroups of the population more likely to achieve a vertically equitable allocation (i.e., subgroups more likely to receive the health care they ought to, given their needs, or geographical areas where performance indicators suggest an allocation of resources that align resources appropriately with needs). Another alternative could be to identify the use of services for a set of clearly defined needs that are based, for example, on clinical guidelines for a particular medical condition. This may be of limited use, especially for studies looking at vertical inequity across healthcare services generally, but could be appropriate if considering a specific service.

In conclusion, the importance of incorporating vertical equity considerations in the way healthcare resources are allocated is being increasingly emphasized. The few empirical studies that have investigated vertical equity in healthcare delivery have used a variety of methods with different underlying assumptions. In this article, some light has been shed on the differences and the validity of the approaches being used has been discussed. Also, areas for further work have been highlighted with a view to improving methods for measuring vertical equity in the delivery of health care.

See also: Health and Health Care, Need for. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach

References

Gravelle, H., Morris, S. and Sutton, M. (2006). Economic studies of equity in the consumption of health care. In Jones, A. (ed.) *The Elgar companion to health economics*, pp 193–204. Cheltenham: Edward Elgar.

Marmot, M. (2010). *Fair society, healthy lives: The Marmot review.* London: University College London.

Sutton, M. (2002). Vertical and horizontal aspects of socio-economic inequity in general practitioner contacts in Scotland. *Health Economics* **11**, 537–549.

Toutkoushian, R. K. and Michael, R. S. (2007). An alternative approach to measuring horizontal and vertical equity in school funding. *Journal of Education Finance* **32**, 395–421.

Wagstaff, A. (2002). Inequality aversion, health inequalities and health achievement. *Journal of Health Economics* **21**, 627–641.

Further Reading

Abásolo, I., Manning, R. and Jones, A. M. (2001). Equity in utilization of and access to public-sector GPs in Spain. *Applied Economics* **33**, 349–364.

Alberts, J. F., Sanderman, R., Eimers, J. M. and Van Den Heuvel, W. J. A. (1997). Socioeconomic inequity in health care: A study of services utilization in Curacao. *Social Science and Medicine* **45**, 262–270.

Laudicella, M., Cookson, R., Jones, M. J. and Rice, N. (2009). Health care deprivation profiles in the measurement of inequality and inequity: An application to GP fundholding in the English NHS. *Journal of Health Economics* **28**, 1048–1061.

Raine, R., Goldfrad, C., Rowan, K. and Black, N. (2002). Influence of patient gender on admission to intensive care. *Journal of Epidemiology and Community Health* **56**, 418–423.

Raine, R., Hutchings, A. and Black, N. (2004). Is publicly funded health care really distributed according to need? The example of cardiac rehabilitation in the UK. *Health policy* **67**, 227–235.

Rocha, G. M. N., Martínez, A. M. S., Ríos, E. V. and Elizondo, M. E. G. (2004). Resource allocation equity in northeastern Mexico. *Health Policy* **70**, 271–279.

Sutton, M. (2002). Vertical and horizontal aspects of socio-economic inequity in general practitioner contacts in Scotland. *Health Economics* **11**, 537–549.

Sutton, M. and Lock, P. (2000). Regional differences in health care delivery: Implications for a national resource allocation formula. *Health Economics* **9**, 547–559.

Medical Decision Making and Demand

S Felder, Universität Basel, Switzerland
A Schmid and V Ulrich, Universität Bayreuth, Germany

Introduction

Ever since medicine has been practiced, medical decision making has been conceptualized. Informal and formal rules have been in place to guide physicians in the therapeutic process – the Hippocratic Oath is a very early example. Probably up until the late nineteenth century the patient had little say in the physician's decision making, and a rather paternalistic relationship between the two prevailed.

The twentieth century brought two important developments. First, the recognition of informed consent as a fundamental ethical requirement strengthened the patient's role in the decision making process. Patients today are frequently characterized as responsible clients and customers who make informed decisions. Second, medical decision making was formalized and its scope extended beyond the direct patient–physician relationship, for example, to questions of public health and reimbursement.

Nowadays, the term medical decision making refers to a wide range of decisions in health care, with a special focus on the methods applied in the decision making process. The point of view from which such analyses are conducted can vary and includes the perspective of an individual physician or patient as well as the perspective of the policy maker. As the concern here is with the relationship between decision making and demand, the natural perspective is the patient's.

Depending on one's point of view, the criteria for assessing decision outcomes may differ and different theories and analytical methods may be applied. The field of medical decision making is characterized by a strong quantitative focus and covers a range of aspects from health economic evaluation to the analysis of cognitive processes and psychological factors. This article adheres to a very narrow definition of medical decision making, focusing on the role of formal decision analysis. This is well-suited to the fundamental concepts of health economics, which rely strongly on utilitarian principles.

Medical decision making and the analysis thereof is driven by the fact that decisions in medical care are in most cases characterized by two attributes: A varying, but frequently high degree of uncertainty and the availability of more than one alternative. Uncertainty arises from a number of sources. First of all, the onset of illness is a seemingly random event and can hardly be predicted. Furthermore, the presence of a sickness in a patient is uncertain; hence the decision maker faces a diagnostic risk. Then there is the therapeutic risk, as there is uncertainty over the effects of a specific therapy on a particular patient. At the same time, the patient himself knows neither how his body will react to the medication prescribed nor how the individual risk factors his genetics expose him to play into it. Another source of uncertainty is the fallibility of diagnostic tests, introducing the possibility of error at this stage of the treatment process.

Medical Decision Making

Basic Model

A formal analysis of medical decision making can help structure very complex problems. The structure of decision processes are typically represented by decision trees in which probabilities, outcomes, and sometimes other parameters are formalized and quantified. This makes all assumptions explicit and enables decision makers to draw informed conclusions. The aim is to identify the treatment strategy which yields the highest expected utility. Decision analysis is thus merely a tool for making informed decisions – the decision itself is still up to the decision maker, who decides how utility is defined and which components are taken into account. This means that legal, ethical, and professional considerations as well as patient preferences and other factors can have an influence. The contribution of formal decision analysis is to provide structure, explicitness, and quantitative measures in this process and thereby facilitate an informed discussion (cf. Weinstein *et al.*, 1980).

A basic model of medical decision making is developed, beginning with the decision over medical treatment under diagnostic risk. For simplicity, there is no differentiation between physicians and patients as decision makers in this model, but rather the assumption that physicians decide purely in the interest of their patients. The patient's health state H_i is unknown; one can be sick or healthy, $i = \mathrm{s}, h$. p describes the *a priori* probability of the sick state. The decision maker must decide whether to treat the patient ($j = +$) or not ($j = -$). $U(H_i^j)$ is the utility function which values health in its different possible states. Expected utility as a function of the treatment decision j becomes

$$EU^j(p) = pU\left(H_{\mathrm{s}}^{\,j}\right) + (1-p)U\left(H_{\mathrm{h}}^{\,j}\right) \qquad [1]$$

The decision maker can also use diagnostic tools and make the treatment decision dependent on the test outcome, i.e., treat if the test is positive and not treat if it is negative. Therefore, j, $j = +, -$, is used, for the test outcome as well as the treatment decision. It follows that $H_{\mathrm{h}}{}^{-}$ indicates the health state after a true negative test result and $H_{\mathrm{h}}{}^{+}$ the health state after a false positive test result, whereas $H_{\mathrm{s}}{}^{+}$ stands for the health consequences of a true positive test result and $H_{\mathrm{s}}{}^{-}$ for those of a false negative test result.

The expected utility of a diagnostic test, $EU^{Dx}(p)$, can then be written as follows:

$$\begin{aligned} EU^{Dx}(p) = {} & p\left[Se \cdot U\left(H_{\mathrm{s}}{}^{+}\right) + (1-Se)U\left(H_{\mathrm{s}}{}^{-}\right)\right] \\ & + (1-p)\left[Sp \cdot U\left(H_{\mathrm{h}}{}^{-}\right) + (1-Sp)U\left(H_{\mathrm{h}}{}^{+}\right)\right] \end{aligned} \qquad [2]$$

where 'Se' is the sensitivity or true positive rate, $1-Se$ is the false negative rate, 'Sp' is the specificity or true negative rate, and $1-Sp$ is the false positive rate of the test.

 doi:10.1016/B978-0-12-375678-7.00808-7

The decision tree in Figure 1 illustrates the decision maker's choice. Squares represent decision nodes and circles indicate chance nodes.

It is useful to define:

$$G = U(H_s^+) - U(H_s^-) > 0 \quad \text{and}$$
$$L = U(H_h^+) - U(H_h^-) < 0 \qquad [3]$$

as the decision maker's utility gain from treatment in the sick state and one's utility loss from treatment in the healthy state.

Three threshold probabilities can now be derived at which the decision maker is indifferent between two actions, as introduced by Pauker and Kassirer (1975, 1980):

$$\tilde{p} = \frac{1}{1 - G/L} \qquad [4]$$

$$\tilde{p}^{Dx} = \frac{1}{1 - LR^+ G/L} \qquad [5]$$

and

$$\tilde{p}^{Rx} = \frac{1}{1 - LR^- G/L} \qquad [6]$$

where $LR^+ = Se/(1 - Sp)$ denotes the positive and $LR^- = (1 - Se)/Sp$ the negative likelihood ratio of the test. $LR^+ > 1$ and $0 \le LR^- < 1$ holds for useful tests (i.e., $Se + Sp > 1$).

The first threshold $\tilde{p}$ is called the therapeutic threshold, at which the decision maker is indifferent between treating and not treating the patient in a situation where no diagnostic test is available. At the test threshold, $\tilde{p}^{Dx}$, one is indifferent between not treating and testing (followed by the treatment decision depending on the test outcome). At the treatment threshold, $\tilde{p}^{Rx}$, the decision maker is indifferent between testing and treating without prior testing. Because $LR^+ > 1 > LR^- > 0$ and $-G/L > 0$, $\tilde{p}^{Dx}$ is located below the therapeutic threshold and $\tilde{p}^{Rx}$ above it: $\tilde{p}^{Dx} < \tilde{p} < \tilde{p}^{Rx}$.

These thresholds allow us to characterize the decision maker's optimal test and treatment strategy. At low *a priori* probabilities of sickness, $0 \le p \le \tilde{p}^{Dx}$, the optimal decision is not to use the test and not to treat. The dominant aspect in this situation is the utility loss resulting from a false positive test result. At intermediate *a priori* probabilities of sickness, $\tilde{p}^{Dx} < p < \tilde{p}^{Rx}$, the optimal strategy is to test and then to treat if the test outcome is positive. With a negative test outcome not treating is indicated. Finally, at high *a priori* probabilities, $\tilde{p}^{Rx} < p < 1$, not testing is optimal and immediate treatment is indicated. The utility loss stemming from a false negative test outcome is the dominant factor here.

The Value of Information

Medical decisions often involve the use of diagnostics. The optimal use of the information gleaned can be conceptualized using the value of information as introduced by Gould (1974). The value of diagnostic information is defined as the additional expected utility resulting from the use of the test. This requires considering the optimal decision when no test is available. This reference situation is characterized by the expected utility of not treating at $p < \tilde{p}$ and the expected utility of treating at $p \ge \tilde{p}$. Given eqns [1]–[3], we can solve for the value of information of a diagnostic test and find

$$VI(p) = \begin{cases} \max(p \cdot Se \cdot G - (1-p)(1-Sp)L, 0) & \text{for } 0 \le p < \tilde{p} \\ \max(-p(1-Se)G + (1-p)Sp \cdot L, 0) & \text{for } \tilde{p} \le p \le 1 \end{cases} \qquad [7]$$

Figure 2 shows the value of information and the three thresholds discussed above. The value of information is positive between the test and treatment thresholds and reaches its maximum at the therapeutic threshold. At *a priori* probabilities of sickness below the test threshold or above the treatment threshold the value of information is zero: in these ranges the information technology should not be used in the treatment decision.

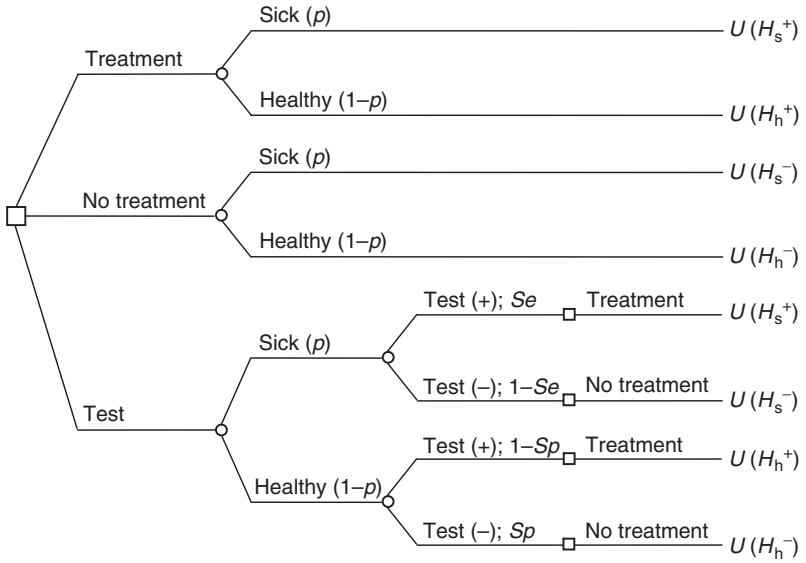


Figure 1 Decision tree for the test treatment decision.

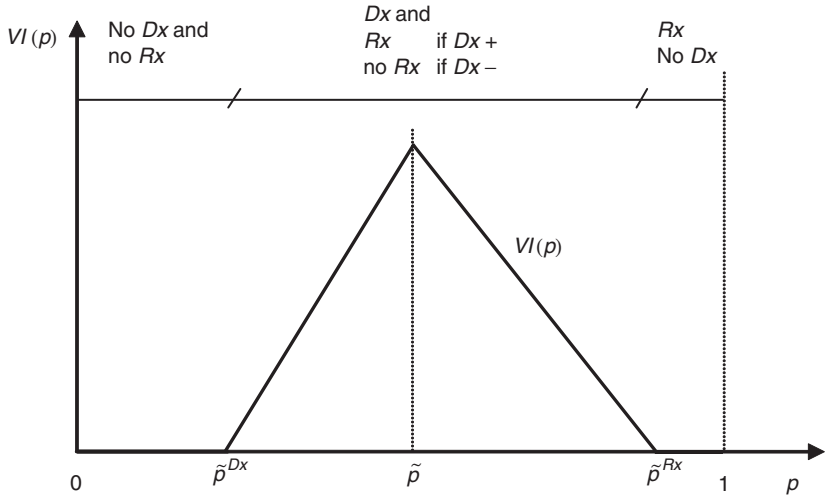


Figure 2 The value of information of a test.

The value of information concept can be applied to situations in which several tests are available and the decision maker has to choose which one to use. It allows the evaluation of the decision maker's marginal rate of substitution between sensitivity and specificity. The decision maker chooses the optimal test dependent on his preferences and the given *a priori* probability of sickness.

Equation [7] leads to

$$dVI(p) = 0 \text{ implies } p \cdot dSe \cdot L = -(1-p) \cdot d(1-Sp) \cdot G \text{ or}$$

$$\left.\frac{dSe}{dSp}\right|_{dVI=0} = \frac{1}{p/(1-p) \cdot (G/L)} \qquad [8]$$

This equation was derived by McNeil *et al.* (1975) and Metz (1978), though without reference to the value of information concept. The marginal rate of substitution between sensitivity and specificity is decreasing in p. In other words, at high *a priori* probabilities of sickness the decision maker chooses a test with high sensitivity, whereas at low *a priori* probabilities one favors tests with high specificity. In the first situation, the benefits of treating the patient dominate, whereas in the second situation, the benefits of not treating become more relevant. The marginal rate of substitution is also decreasing in G/L. Hence, if either G increases or L decreases (in absolute terms), tests with high sensitivity are more attractive. Inversely, if either G decreases or L increases (in absolute terms), tests with high specificity are favorable.

Risk Aversion

Medical decision analysis often implicitly assumes a decision maker to be risk neutral. This is consistent with the practice of measuring intervention outcomes in terms of changes in mortality. New research also investigates the effect of higher order risk preferences on medical decision making. Consider the effect of risk aversion on the test and treatment thresholds. For a decision maker with a concave utility function, i.e., marginal utility of health decreasing in the level of health, it is easy to show that G, the gain from treatment in the sick state, increases and L, the loss from treatment in the healthy state, decreases in absolute terms as compared to a decision maker with linear utility. With this, the decision maker's test and treatment thresholds decrease (all $\tilde{p}$ are decreasing in $(-G/L)$). Intuitively, if either the potential benefit from treatment increases or the potential harm decreases, the treatment option becomes more attractive. The risk averse decision maker uses his test and treatment strategy as an insurance device (it reduces the spread between the possible health states) by opting for testing and treatment at lower prevalence rates than a risk neutral decision maker. In situation where the decision maker can choose between several tests he will tend to favor those with high sensitivity if he is risk averse. This can be seen in eqn [8] because the marginal rate of substitution between sensitivity and specificity is increasing in $(-G/L)$.

The decision maker's risk preferences are only partially described by risk aversion. Higher-order attitudes such as prudence and temperance also appear to play a role in decision making under uncertainty (Kimball, 1990; Eeckhoudt and Schlesinger, 2006). Felder and Mayrhofer (2013) analyze higher-order risk preference in a medical setting by adding a comorbidity risk to a specific index condition, or primary illness. They show that the effects of risk aversion on the test and treatment thresholds are reinforced when prudence and temperance are taken into account. Risk attitudes may explain why screening activities in very low *a priori* probability ranges are observed (for instance, the prostate-specific antigen (PSA)-test for prostate cancer in asymptomatic middle-aged men) where under risk neutral decision making would not be expected.

The above example covers some of the basic concepts of formal decision analysis in the context of medical decision making. The focus is exclusively on the utility derived from the patient's health status. (The utility of health states is often quantified and thus made comparable across illnesses by means of the quality-adjusted life-years concept.) Depending on the aim of the analysis the model will incorporate further information, for example, relating to treatment costs. The analysis can also pertain to a supply rather than a demand

decision. In the following, we elaborate on a model which puts a stronger focus on the patient's perspective and his demand for medical care.

Health Care Demand

Medical decision making can be integrated into a conventional demand framework. So far the gain from treatment, i.e., $h = H_s^+ - H_s^-$, was exogenous to the decision model. However, the effect of a treatment depends on the type and amount of medical inputs M that are used. A health production function $h = f(M)$ describes this relationship.

The utility maximizing decision maker faces a limited budget, which he can spend on medical inputs M or on other consumption goods C. The utility function is thus $U(C,H)$, which is strictly positive and strictly increasing in both M and C. Marginal utility is decreasing, implying that the individual is risk averse. To ensure that a change in consumption does not have an effect on the marginal utility of medical inputs and vice versa, the mixed derivatives must be zero.

To include h – the health gain produced by medical inputs – explicitly in the utility function, $H = H_s^- + h$ is noted. (To keep this illustration simple, healthy patients and the potential issues they raise are disregarded.) The utility function can then be written as

$$U = U(C, H_s^- + h) = U(C, H_s^- + f(M)) \quad [9]$$

Demand for health services now becomes a derived demand, as it is driven by the underlying demand for health. All prices except for medical inputs are set equal to one. With income Y, the budget constraint is

$$C + p_M M = Y \quad [10]$$

Solving the Lagrangean function $L(C,M,\lambda) = U(C, H_s^- + f(M)) + \lambda(C + p_M M - Y)$, where λ indicates the marginal utility of income, a combined first-order condition is derived for the utility maximizing demand for consumption and health care:

$$U_H / U_C = p_M / f'(M) \quad [11]$$

The right side of the equation can be interpreted as the marginal cost of investing in health. By rewriting this equation the following equation is obtained:

$$U_H f'(M) / p_M = U_C \quad [12]$$

which can be interpreted intuitively: In the optimum, the last monetary unit spent must generate the same marginal utility whether it is spent on medical care or on consumption.

Conducting a comparative static analysis can investigate the effects of changes in the exogenous variables on the demand for health. Let us first look at technological innovation and assume that $h = \varphi M^{\alpha}$, where $\varphi > 0$ reflects the productivity parameter. If innovation increases the productivity of medical inputs, this is reflected in an increase in the marginal product of medical inputs $f'(M) = \varphi \alpha M^{\alpha - 1}$. In other words, the marginal cost of producing health diminishes. This results in higher demand for health and medical inputs. The opposite effect is caused by higher prices for medical inputs. Although higher income also results in a higher level of demand, a higher initial health level does not have a clear positive or negative effect. In this model, the probability of survival, which is included in the expected utility function, has no influence on demand.

This model can be extended to include the aspect that health care demand is often directed toward preventing early death. A simple approach to modeling survival is to introduce an initial probability of survival π_0 which can be increased by medical inputs: $\pi(M)\pi_0 + g(M)$ with $g'(M) > 0$ and $g''(M) < 0$. This extension permits the evaluation of the demand for health as a function of initial survival. It can be shown (see Felder and Mayrhofer (2011)) that the demand for health care is inversely related to the initial survival rate, which is tantamount to the famous dead-anyway effect (Pratt and Zeckhauser, 1996).

Double Moral Hazard under Two-Sided Asymmetric Information

This article began by outlining a basic model of medical decision making which assumes that the physician acts as a perfect agent for the patient. In the second model the focus was on the patient's view. In practice, both perspectives have to be combined. A primary reason is that patients have become increasingly involved in treatment decisions. They want to participate in the decision making process and physicians want them to understand the implications of the decisions that have to be taken. Second, even disregarding an ideal-type shared decision making process, the outcomes for physicians and patients are interlinked.

A model that characterizes this relationship must also map out the underlying two-way moral hazard structure. A simplified model would contain only one stage, in which a utility maximizing physician makes a treatment decision based on factors such as the patient's health status, coinsurance, and his remuneration. However, the health outcome depends not only on the treatment, but also on the patient's health-related behavior, even beyond the narrowly defined notion of compliance. This creates strategic interdependence between the levels of utility realized by the physician and the patient. Both sides suffer from information deficits. The patient either does not know whether the physician is sharing all relevant information or simply lacks the capability to understand all the information that is given. Furthermore, they cannot easily verify the quality of the treatment. On the other side, the physician does not know everything about the patient's health-related behavior. This behavior, however, affects the health outcome. Thus, health outcomes can be interpreted as the result of both the physician's medical services and the patient's behavior.

The physician's treatment decision and the patient's health-related behavior can interact in three key ways: Medical services and health behavior can be strategically independent, implying that the level of medical services does not affect the marginal productivity of the patient's compliance and vice versa. Alternatively, the two components can be strategic complements, i.e., if one factor is increased this has a positive effect on the other factor. Finally, they might be strategic

substitutes, so that a lower level of health care services can be compensated – at least so some extent – by better health-related behavior. These three types of interaction between physician decision and patient behavior yield different results. There is some evidence that there is indeed a strong mutual influence of the demand for health and the provision of health care services (Schneider and Ulrich, 2008).

Concluding Remarks

The expected utility theory of von Neumann and Morgenstern is the fundamental building block of most models in medical decision making under uncertainty. However, it has been criticized for its failure to predict individual behavior. Alternative non-expected utility theories such as rank-dependent choice models have been suggested to reflect actual behavior more precisely. One aspect, for instance, is that decision makers tend to overweight small probabilities and underweight large probabilities, which leads to an inverse S-shaped probability transformation which has been confirmed in empirical studies (Abdellaoui, 2000; Bleichrodt and Pinto, 2000).

Recent research, in turn, has challenged the validity of rank-dependent theory. Among others, List (2004) showed that individuals with extensive experience behave largely rationally, or in accordance with the expected utility theory. Physicians take decisions on tests and treatments as a matter of routine – and they are expected to make unbiased estimations of probabilities and take coherent decisions.

The use of the expected utility theory is also warranted in the prescriptive realm of medical decision making. If an optimal policy has to be chosen or recommended, "the expected utility is the best theory to determine which decisions to undertake" (Wakker, 2008, p. 687).

See also: Adoption of New Technologies, Using Economic Evaluation. Economic Evaluation, Uncertainty in. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Physician-Induced Demand. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Pricing and User Fees. Problem Structuring for Health Economic Model Development. Willingness to Pay for Health

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Sciences* **46**(11), 1497–1512.

Bleichrodt, H. and Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting Functions. *Management Sciences* **46**(11), 1485–1496.

Eeckhoudt, L. and Schlesinger, H. (2006). Putting risk in its proper place. *American Economic Review* **96**, 280–289.

Felder, S. and Mayrhofer, T. (2011). *Medical decision making, a health economic primer*. Heidelberg: Springer.

Felder S. and Mayrhofer T. (2013). Higher-order risk preferences: Consequences for test and treatment thresholds and optimal cutoffs. Medical Decision Making (in press).

Gould, J. P. (1974). Risk, stochastic preference, and the value of information. *Journal of Economic Theory* **8**(1), 64–84.

Kimball, M. S. (1990). Precautionary saving in the small and in the large. *Econometrica* **58**(1), 53–73.

List, J. (2004). Neoclassical theory versus prospect theory: Evidence from the market place. *Econometrica* **72**(2), 615–625.

McNeil, B. J., Keeler, M. and Adelstein, S. M. (1975). Primer on certain elements of medical decision making. *New England Journal of Medicine* **293**(5), 211–215.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**(4), 283–298.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: A cost benefit analysis. *New England Journal of Medicine* **293**(5), 229–234.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine* **302**(20), 1109–1117.

Pratt, J. W. and Zeckhauser, R. J. (1996). Willingness to pay and the distribution of risk and wealth. *The Journal of Political Economy* **104**(4), 747–763.

Schneider, U. and Ulrich, V. (2008). The physician-patient relationship revisited: The patient's view. *International Journal of Health Care Finance and Economics* **8**(4), 279–300.

Wakker, P. P. (2008). Lessons learned by (from?) an economist working in medical decision making. *Medical Decision Making* **208**, 690–698.

Weinstein, M. C., Fineberg, H. V., Elstein, A. S., et al. (1980). *Clinical decision analysis*. Philadelphia: W.B. Saunders.

Medical Malpractice, Defensive Medicine, and Physician Supply

DP Kessler, Stanford University, Stanford, CA, USA

Introduction

An efficient system of medical malpractice liability law should induce physicians to supply precautionary medical treatments as long as the benefits exceed the costs. In practice, the US malpractice system may deviate from this ideal along two dimensions. First, it may create incentives to supply cost-ineffective treatments based on fear of legal liability – to practice 'defensive medicine' (Kessler, 2011). Defensive medicine can take many forms: too many diagnostic tests, specialist office visits, and even unnecessary surgeries. One recent paper estimates the cost of defensive medicine in the US to be 2–3% of health spending, or over US$50 billion per year (Mello *et al.*, 2010).

Second, it may create incentives to decline to supply cost-effective treatments. This phenomenon is sometimes described as 'negative defensive medicine,' to distinguish it from its counterpart above. Negative defensive medicine can also take many forms, including physician avoidance of high-risk patients or procedures, reduction of hours of work, relocation, or exit from the profession altogether. All of these involve a restriction in the supply of physicians' services that is not in society's interests.

Thus, the impact of the US malpractice system on physician supply is an important policy issue. This article summarizes the empirical research on this topic. In general, the research finds that higher levels of malpractice liability lead to lower levels of supply. This finding is strongest and most robust for specialists, other physicians who are most likely to be at high risk for a malpractice claim, and for physicians in rural areas. However, the consequences of malpractice-induced reductions in supply on the cost of care and patient health outcomes – and hence on social welfare – remains largely an open question.

The article begins with an overview of the operation of the US malpractice system and a theoretical framework in which its effects on supply can be evaluated. It then summarizes the empirical evidence. It concludes with a discussion of the implications of these findings for social welfare and suggestions for future research.

The US Malpractice System and Physician Supply

In general, malpractice claims are adjudicated in state courts according to state tort laws. (The text in this section borrows heavily from Kessler (2011).) These laws generally require three elements for a successful claim. First, the claimant must show that the patient actually suffered an adverse event. Second, a successful malpractice claimant must establish that the provider caused the event: the claimant must attribute the injury to the action or inaction of the provider, as opposed to nature. Third, a successful claimant must show that the provider was negligent. Stated simply, this entails showing that the provider took less care than that which is customarily practiced by the average member of profession in good standing, given the circumstances of the doctor and the patient (Keeton *et al.*, 1984). Collectively, this three-part test of the validity of a malpractice claim is known as the 'negligence rule' (see Budetti and Waters (2005), for a layperson's explanation).

Even though they share the same basic structure, states' liability laws differ in terms of the level of liability they impose on providers. In particular, several states have changed their laws in ways that reduce liability relative to its historical levels – to adopt 'tort reforms.'

The consequences of malpractice law and tort reforms for the supply of physician services are theoretically indeterminate. An early model by Danzon *et al.* (1990) illustrated why it is so difficult to assess the impact of malpractice law on supply a priori. Suppose that physician markets are monopolistically competitive, and states' liability regimes impose both fixed and variable costs on providers. In this context, the model shows that the extent to which decreases in liability costs lead to increases in supply depends on the magnitude of costs, the extent to which they are fixed or variable, and the effect of variable costs on physicians' profits. If liability costs are primarily fixed, then decreases in costs would lead to higher profits in the short run and ultimately to increased supply. If liability costs are primarily variable, however, and the market-wide elasticity of demand for physician services is low, then tort reform would lead to a decrease in price with a minimal change in quantity, and hence little change in profits and ultimately supply.

As Matsa (2007) pointed out, extensions of this model suggested that malpractice law may have very different effects across specialties and geographic areas. Malpractice insurance premiums, perhaps the most important cost imposed on providers by the liability system, differ dramatically along these dimensions. For example, in 2009, premiums in Suffolk County, New York, for specialists in internal medicine and obstetrics were US$33 000 and US$178 000, respectively, whereas premiums in Colorado were approximately one-third as much (Medical Liability Monitor 2009). In the Danzon *et al.* (1990) framework, such differences could imply very different supply responses to similar tort reforms.

Empirical Assessment of the Effects of the Malpractice System and Tort Reforms

The extent to which changes in malpractice law lead to changes in supply is, thus, an empirical question. Empirical research on the effects of malpractice law on supply is of three types. The first arm of the literature surveys physicians about their opinion of the role of the malpractice system on their scope of practice, hours of work, or likely future labor force participation (e.g., Mello *et al.*, 2005). Although opinion surveys indicate that physicians believe that the malpractice system has a significant effect on supply, this approach only

 doi:10.1016/B978-0-12-375678-7.01106-8

provides information about physicians' perceptions, which may or may not be closely related to their economic decisions.

A second arm examines the correlation between supply and measures of malpractice costs such as insurance premiums, claims rates, or average payments per claim. Baicker and Chandra (2005) estimated the relationship between the change in the number of physicians per capita from 1993–2001 by specialty, age, and rural location and the change in these measures of costs across US states. They found that the overall size of the physician workforce does not respond to increases in costs, although in rural areas, the response of the size of the workforce is small but statistically significant. Dranove and Gron (2005) found that the incidence of a high-risk procedure (craniotomy) and women's travel time for a high-risk delivery did not change in Florida contemporaneous with that state's dramatic increase in premiums in the early 2000s. Mello *et al.* (2007) reported similar findings about physicians scope of practice in Pennsylvania contemporaneous with that state's dramatic increase in premiums, but documented a decline in the number of practicing obstetricians there. In a recent working paper, Reyes (2010) found that increases in malpractice premiums lead to increased specialization among US obstetrician–gynecologists, with some physicians concentrating more in obstetrics and others in gynecological surgery.

However, the possibility that unobserved determinants of supply are correlated with premiums, claims rates, or payments qualifies the results of all these studies. The malpractice costs in a particular area may be increasing because the patients are particularly sick (and hence prone to adverse outcomes and malpractice claims), because the patients have more 'taste' for medical interventions (and hence more likely to disagree with their provider about management decisions), or because of many other factors. To the extent that these factors are not captured fully in observational data, estimates of the impact of malpractice costs in the studies above would tend to understate the magnitude of the true effect.

The third arm of the literature addresses this concern by identifying the effect of malpractice costs on supply with variation in tort reforms across states and over time. As Kessler (2011) pointed out, this technique yields unbiased assessments of the impact of the malpractice system under the assumption that the adoption of reforms is uncorrelated with unobserved determinants of supply (see US Congress, Congressional Budget Office (2006) for a criticism of this assumption).

Kessler *et al.* (2005) estimated the effect of reforms on the number of physicians using individual–physician-level data from the American Medical Association's Physician Masterfile, matched with data on US states' tort laws and state demographic, political, population, and health care market characteristics. They grouped reforms into two types – 'direct' and 'indirect' reforms. Direct reforms directly reduced malpractice awards. The most important reforms of this type are caps on damages that limit a defendant's financial liability (or some element of liability, like pain-and-suffering or punitive damages) in a successful lawsuit. As the research discussed in Kessler (2011) showed, direct reforms reduce the frequency and size of claims, and insurance premiums. Other reforms that only affect awards indirectly, such as reforms imposing mandatory periodic payments (which require damages in certain cases to be disbursed in the form of annuity that pays out over time), limits on joint and several liability, or limits on the contingent fees that plaintiffs' attorneys can charge have had a less consistent impact on malpractice costs. They find that 3 years after adoption, direct reforms increase physician supply in the US by 3.3%, all else held constant. They also find that direct reforms had a larger effect on the supply of most (but not all) specialties with high malpractice insurance premiums, on states with high levels of managed care, and on supply through retirements and entries than through the propensity of physicians to move.

Newer work examines the supply response to reforms for different subgroups. Using county-level data from 1985–2000, Encinosa and Hellinger (2005) showed that caps on pain-and-suffering damages lead to increases in supply, especially in rural areas. Using county-level data from 1970–2000, Matsa (2007) found no aggregate effect of caps on damages, but large effects for rural physicians and especially rural specialists; he hypothesized that this is because rural doctors face greater uninsured liability costs and a more elastic demand for medical services. Using state-level data from 1980–2001, Klick and Stratmann (2007) found that caps have a significant effect on the number of doctors per capita, and that this effect is concentrated among the specialties that face the greatest exposure to malpractice risk.

Two other studies assess the extent to which reforms affect physician avoidance of high-risk patients and hours of work. In the only study of its kind, Dubay *et al.* (2001) examined the effect of tort reforms on the supply of prenatal care for pregnant women. They found that reforms result in prenatal care beginning earlier in pregnancy, especially for women with low socioeconomic status, who may be more likely to file a malpractice claim. Helland and Showalter (2009) showed that reform-induced reductions in liability costs lead to increases in hours, especially for physicians aged 55 years and older.

Conclusions

The small but growing literature on the effect of malpractice law on physician supply reaches two main conclusions. First, tort reforms that directly reduce the costs imposed by the US malpractice system have a small, statistically significant positive effect on the number of physicians per capita, on the order of 2–4%. Second, although there is some debate about the robustness of this result for the overall physician population, there is almost universal agreement that reforms increase supply of certain subgroups that are likely to be sensitive to malpractice incentives. These subgroups include specialists who face substantial exposure to malpractice risk; rural physicians who may be less able to increase their markups to recover the costs of malpractice from their patients; and physicians who serve patient populations that are more likely to file a claim.

If markets for physician services functioned perfectly, these results would imply that tort reforms improve social welfare. Suppose that patients had perfect information about the risks of treatment, and physician services were priced at their marginal cost. In this case, as long as the liability system imposed no transactions costs, tort reforms should not affect supply.

Patients and their physicians would make optimal decisions regardless of the level of liability, and any reductions in awards for malpractice would be completely offset by increases in prices. Reductions in liability would only increase supply if the liability system consumed real resources, which would mean that it was, by definition, inefficient.

Of course, there are many reasons why physician markets might not reflect this ideal. Most important, if patients don't have perfect information about risks, then liability-induced reductions in supply might be optimal. Physicians might not take appropriate precautions in the absence of liability, which could lead to an equilibrium quantity of physician services that was too large. Liability-induced reductions in supply might be optimal for another reason: the widespread prevalence of health insurance, which means that neither patients nor physicians bear the full costs of care in any particular case. Such moral hazard could also lead to a socially excessive supply of services, which might be mitigated by liability costs. The fact that markets for physician services are unlikely to be perfectly competitive adds a third source of indeterminacy. Even simple models of monopolistic competition showed that the free-entry outcome can involve socially too few or socially too many suppliers (Spence, 1976; Tirole, 1988).

Any assessment of the welfare implications of liability-induced reductions in physician supply must therefore examine their consequences for health care costs and health outcomes. To date, only two studies have attempted to do so, and their findings are inconclusive. Dubay *et al*. (2001) found that tort reforms lead to prenatal care beginning earlier in pregnancy, although they fail to reject that this increase in supply led to improved infant health. Along the same lines, Klick and Stratmann (2007) found that reforms lead to an increase in the supply of high-risk specialties, but reported that the effects of reform on infant mortality were mixed.

Investigation of the links between physician supply, costs, and outcomes is therefore an important topic for future research. Future research might also investigate the effect of other, nontraditional reforms to the liability system on physician supply. Kessler (2011) discussed several of these, including restricting the legal discoverability of information gathered as part of private, voluntary efforts to reduce medical errors; allowing evidence of compliance with clinical practice guidelines as an affirmative defense to negligence; and expanding the use of alternative dispute resolution, no-fault, and administrative compensation systems.

See also: Health Care Demand, Empirical Determinants of. Moral Hazard. Physician Management of Demand at the Point of Care

References

Baicker, K. and Chandra, A. (2005). The effect of malpractice liability on the delivery of health care. *Forum for Health Economics and Policy: Frontiers in Health Policy Research*. Available at: http://www.bepress.com/fhep/8/4 (accessed 22.09.10).

Budetti, P. P. and Waters, T. M. (2005). Medical malpractice law in the United States. *Kaiser Family Foundation Report*. Available at: http://www.kff.org/insurance/upload/Medical-Malpractice-Law-in-the-United-States-Report.pdf (accessed 22.09.10).

Danzon, P. M., Pauly, M. V. and Kington, R. S. (1990). The effects of malpractice litigation on physicians' fees and incomes. *American Economic Review* **80**(2), 122–127.

Dranove, D. and Gron, A. (2005). Effects of the malpractice crisis on access to and incidence of high-risk procedures: Evidence from Florida. *Health Affairs* **24**(3), 802–810.

Dubay, L., Kaestner, R. and Waidmann, T. (2001). Medical malpractice liability and its effect on prenatal care utilization and infant health. *Journal of Health Economics* **20**, 591–611.

Encinosa, W. E. and Hellinger, F. J. (2005). Have state caps on malpractice awards increased the supply of physicians? *Health Affairs Web Exclusive* W5-250–258.

Helland, E. and Showalter, M. (2009). The impact of liability on the physician labor market. *Journal of Law and Economics* **52**(4), 635–663.

Keeton, W. P., Dobbs, D. B., Keeton, R. E. and Owen, D. G. (1984). *Prosser and Keeton on Torts*, 5th ed. St. Paul, MN: West Publishing Co.

Kessler, D. P. (2011). Evaluating the medical malpractice system and options for reform. *Journal of Economic Perspectives* **25**(2), 93–110.

Kessler, D. P., Sage, W. and Becker, D. (2005). The impact of malpractice reforms on the supply of physician services. *Journal of the American Medical Association* **293**, 261–825.

Klick, J. and Stratmann, T. (2007). Medical malpractice reform and physicians in high-risk specialties. *Journal of Legal Studies* **36**(Part 2), S121–S142.

Matsa, D. A. (2007). Does malpractice liability keep the doctors away? Evidence from tort reform damage caps. *Journal of Legal Studies* **36**(Part 2), S143–S182.

Mello, M. M., Chandra, A., Gawande, A. A. and Studdert, D. M. (2010). National costs of the medical liability system. *Health Affairs* **29**(9), 1569–1577.

Mello, M. M., Studdert, D. M., DesRoches, C. M., et al. (2005). Effects of a malpractice crisis on specialist supply and patient access to care. *Annals of Surgery* **242**(5), 621–627.

Mello, M. M., Studdert, D. M., Schumi, J., Brennan, T. A. and Sage, W. M. (2007). Changes in physician supply and scope of practice during a malpractice crisis: Evidence from Pennsylvania. *Health Affairs* **26**(3), w425–w435.

Reyes, J. W. (2010). The effect of malpractice liability on the specialty of obstetrics and gynecology. *NBER Working Paper* 15841. Cambridge, MA: National Bureau of Economic Research.

Spence, M. (1976). Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* **43**, 217–235.

Tirole, J. (1988). *The theory of industrial organization*. Cambridge, MA: MIT Press.

US Congress, Congressional Budget Office (2006). Medical malpractice tort limits and health care spending, *Background Paper No. 2668*. Available at: http://www.cbo.gov/ftpdocs/71xx/doc7174/04–28-MedicalMalpractice.pdf (accessed 22.09.10).

Medical Tourism

N Lunt and D Horsfall, University of York, Heslington, York, UK

Introduction

The notion of traveling abroad for the purposes of health and well-being is well established. The spas of Hungary, baths of Turkey, and geysers of Sweden have long been popular destinations for those seeking convalescence. Where surgical care was required, the direction of travel generally saw the wealthy citizens of poorer nations traveling to the richer, more medically advanced countries. However, international travel for the purposes of medical treatment is no longer the preserve of political and economic elites. Contemporaneously the whole spectrum of medical treatment is offered in destinations around the world as part of a global market in health care. This treatment spans the full range of medical services, but most commonly includes dental care, cosmetic surgery, elective surgery, and fertility treatment. Medical value travel or 'medical tourism' as a term has come to represent situations where consumers elect to travel across international borders with the intention of receiving some form of medical treatment. Differences between medical and health tourism focus on the type of intervention, setting, and particular inputs involved. Setting the boundary of what is health and counts as medical tourism for the purposes of trade accounts is not straightforward. Within this range of treatments, not all would be included within health trade. Cosmetic surgery for esthetic rather than reconstructive reasons, for example, would be considered outside the health boundary.

At first sight medical and tourism are curious terms to run together. Medical and surgical treatment can involve risk, pain, and discomfort; tourism is associated with relaxation and pleasure albeit associated with travel. Although some treatment destinations are those associated with sun, scenery, and sightseeing, it is not clear the extent to which such local attractions are important in the patient's decision.

A range of nomenclature is used in the health services literature, including international medical travel, outsourcing, and refugees. Although the term medical tourism is increasingly being employed, there are a number of commentators who are critical of its use. Their criticisms center on the notion somehow devaluing the rather serious procedures that such patients are often undertaking.

Arguably, however the concept of medical tourism does have analytical purchase – capturing the health sector element as well as the wider economic impact of such travel. Although it must be acknowledged that medical tourism may have little to do with general tourism rationale, the term is useful because it points toward the commodification and commercialization of health travel. Hence, the role of the industry, issues of advertising, and supplier-induced demand are brought to the fore.

The Nature and Scope of Medical Tourism

Although there is a general consensus that the medical tourism industry has burgeoned over the past decade, there remains disagreement as to the current size of the industry. Figures that are regularly reproduced in the literature draw on data collected and projections made by Deloitte, which places the number of US citizens leaving the country in search of treatment at 750 000 in 2007. The main objection to Deloitte's figures come from McKinsey and Co who suggests that, although the potential for such large numbers exist, a more accurate worldwide figure would be between 60 000 and 85 000 medical tourists per year. In large part, very significant disparity may be due to different definitions of medical tourism.

The numbers of medical tourists proffered by McKinsey still appear rather small, particularly in the context of a US population of 360 million, and even the 50 million uninsured. Given that even the most conservative estimates of inward medical tourism to India place the number of tourists at 200 000, alongside figures between 200 000 and 350 000 for Singapore, and 200 000 for Cuba, it would seem that McKinsey's numbers are understated.

This lack of clarity extends beyond not only the numbers of medical tourists but also their profile, the process of becoming a medical tourist, and the aftereffects of medical tourism. It is assumed that different drivers exist for higher and lower income patient groups traveling from North America and Western Europe. But relatively little is known about sociodemographic profile, age, gender, existing health conditions, and status in attempting to map the composition of the medical tourism market.

Such numbers and insights are important to quantify economic impact and also to assess potential risk to source health systems. Given this gap in knowledge, the discussion within the article is inevitably limited in terms of the literature base from which it can draw upon.

Why do People Travel?

As would be expected, globalization has played a significant role in the development of medical tourism. Developments in medical tourism mirror the expansion of markets in health care and embedding of neoliberalism on the world stage. Precipitating the rise of what is seen as a consumerist age marked by lower levels of social solidarity, globalization has advanced the commodification of health care. At the most basic level the advent of the Internet and lowering of travel costs has undoubtedly played a vital role in not only raising awareness of the opportunities for surgery abroad but also making the pursuit economically viable. The freeing of medical services from their traditional territorial boundaries reflects a more transnational and international role for health policy development, with partnerships developing between organizations in established and developing medical markets, transnational companies with an increasing stake in health care in multiple countries, and an emerging role developing for supranational bodies.

In terms of the medical tourism market, the free movement of goods and services under the auspices of the World Trade Organization and its General Agreement on Trade in Services has accelerated the liberalization of the trade in health services, as have developments with regard to the use of regional and bilateral trade agreements. As health care is predominantly a service industry, this has made health services more tradable, global commodities. But that it is easier to travel abroad for care does not fully explain why patients become medical tourists.

The most common explanatory factor cited is that of cost. Indeed it is clear that for those in the US and Western Europe who feel the need to 'go private,' the potential cost savings of traveling abroad are huge, with any review of prices showing a potential saving between 30% and 90% depending on the treatment sought (see **Boxes 1–4**). In terms of familiarity, expatriates often have medical care on their visits back to their 'home' country, which would also show up as medical tourism; for example, the large Indian Diaspora in the UK, and the 2nd Generation Mexicans living in the US (see **Boxes 4** and **5**). In addition, some treatments may not be available or may be subject to a wait in the home country, including the latest technology and techniques. Moreover, some treatments may not be legal in the country of origin. The desire for privacy and wish to combine traditional tourist attractions, hotels, climate, food, cultural visits with medical procedures are also thought to be key contributing factors to the growth in this market. Each of these factors, on their own or in combination, has shifted the direction of medical travel. In the present day not only are India and Thailand top destinations for complex elective procedures but also the very tourists visiting their hospitals, along with those in Poland, South Africa and beyond, are traveling from countries with established and often championed health care systems. Conversely, these championed health systems continue to treat medical tourists at facilities with long-standing reputations (see **Box 5**).

Marketing Medical Tourism

At the most basic level, prospective medical tourists are faced with a level of information that is at once overwhelming and also relatively unhelpful. Internet sites marketing destination and providers are relatively cheap to set up and run, and contributors may post information without being subject to clear quality controls or advertising standards. Medical tourist sites promote benefits and downplay the risks, and the lack of clear regulation regarding what information can be presented on the Internet to prospective medical tourists is then compounded by deeper issues of credibility, trust, and perceptions of risk. As with all medical treatments, an element of risk exists to the patient's health, which is supposedly outweighed by the potential benefits resulting from the treatment. What can be gleaned from the literature concerning risk and safety-related incidents for medical tourism is limited. Although there is evidence regarding, for example, the occurrence of adverse events in the UK hospitals, there is little similar overseas/international data for medical tourist destinations.

Evidence of clinical outcomes for medical tourist treatments is limited and reports are difficult to obtain and verify. Little is known about the relative clinical effectiveness and outcomes for particular treatments, institutions, clinicians, and organizations. There is scant evidence on long- or short-term follow-up of patients dispersing to home countries following treatments at the range of destinations. That positive treatment outcome should result is important not least because typically the patient's local health care takes on the responsibility and funding for postoperative care including treatment for complications and to remedy side effects.

Two particularly interesting stakeholders in the medical tourism industry are brokers/facilitators and providers. There has been a steady rise in the number of companies and consultancies offering brokerage arrangements for services and providing web-based information for prospective patients

Box 1 Case study 1

Brazil

Estimated size of industry	50 000 Inward medical tourists per year
Key procedures offered	Cosmetic, reproductive, and bypass
Estimated cost saving – for the consumer	45–60%
Key pull factors	Cost, quality of care, and range of procedures
Role of state	Limited

The Brazilian medical tourism market centers on cosmetic surgery. Indeed there are more plastic surgeons per capita in Brazil than any other country in the world. Studies suggest that the patient-centered nature of care in Brazil, the willingness to offer procedures not available elsewhere, and the reputation for quality are key pull factors in addition to cost. This perceived high level of quality is undoubtedly aided by the proportion of accredited clinics (more facilities accredited than any country other than the US) and partnerships with organizations such as the International Hospital Corporation, which is based in the US.

The Brazilian government plays only a limited role in promoting the industry. Indeed, although health tourism was recognized in the Brazilian National Tourism Plan (2007–10), it was not identified as an area of focus or potential revenue. This undoubtedly reflects the fact that most medical tourists are catered for in private clinics and that as yet, it is estimated that the financial benefits of even the long-established cosmetic industry are relatively modest.

Source: Reproduced from Edmonds, A. (2011). 'Almost invisible scars': Medical tourism to Brazil. *Signs: Journal of Women in Culture and Society* **36**, 297–302; Herrick, D. M. (2007). Medical tourism: Global competition in health care. *NCPA Policy Reports*. Dallas: National Center for Policy Analysis; and Ramírez de Arellano, A. B. (2007). Patients without borders: The emergence of medical tourism. *International Journal of Health Services* **37**, 193–198.

Box 2 Case study 2

Hungary

Estimated size of market	Estimated 40% share of European dentistry tourism market. Has been estimated between 400 000 and 1 million medical tourists
Average potential cost saving for the consumer	45%
Major clinical areas	Dentistry
Key pull factors	Cost, quality, reputation, and location
Role of government	Keen to promote the industry and reduce regulatory road blocks

Hungary is an exporter country within the medical tourism market. Its reputation for high quality and low costs makes it, along with countries such as Poland, a genuinely accessible and affordable option close to Western European countries such as Germany and Austria. It has been labeled the dental capital of the world, with a much higher dentist per capita ratio than can be found in neighboring Austria, Germany, or the UK. Indeed it is estimated that between one in two and one in three of all Austrians travel to Hungary to meet their dentistry needs. The Hungarian government has offered much in the way of support for the medical tourism industry, expanding domestic incentives to travel to wellness locations to the foreign tourist market, proclaiming 2003 as The Year of Medical Tourism, and even using its premiership of the EU to promote the medical tourism industry.

Recent years have seen an expansion into the cosmetic surgery market as well as the general medical surgery market. The expansion into other clinical areas is still at a relatively early stage.

Source: Reproduced from Herrick, D. M. (2007). Medical tourism: Global competition in health care. *NCPA Policy Reports*. Dallas: National Center for Policy Analysis; and Terry, N. P. (2007). Under-regulated health care phenomena in a flat World: Medical tourism and outsourcing. *Western New England Law Review* **29**, 421.

Box 3 Case study 3

Thailand

Estimated size of the market	Between 350 000 and 1 million visitors per year
Average potential cost savings for the consumer	From 30% to 90% depending on the procedure
Main clinical areas	Cosmetic, fertility, dental, gender reassignment, and cardiac
Major pull factors	Cost and quality
Role of state	Key role in promoting the industry. Legislative role in reducing barriers to the industry

The Thai medical tourism industry is part of the first wave of Asian medical tourism markets to open up and along with Singapore and India represents one of the biggest exporters of medical tourism services. As with most other Asian countries, cost is a key driver of medical tourism with cardiac, cosmetic, and gender reassignment procedures between 60% and 90% cheaper than US prices.

Both business and the government have played major roles in marketing Thailand as a cheap, safe, and relaxing country in which to get the highest standard of treatment. On a practical level, the Thai government has eased visa restrictions on medical tourists and provided funds to support the development of a medical hub in and around Bangkok. Thai hospitals meanwhile have sought to market themselves to the wider world by placing a heavy emphasis on high rates of Thai Joint Commission International accreditation and the reputations of established hospitals such as the Bumrungrad and Bangkok Hospitals.

Source: Reproduced from Chee, H. L. (2007). Medical tourism in Malaysia: International movement of healthcare consumers and the commodification of healthcare. *ARI Working Paper* (Online) **83**. Available at: http://www.ari.nus.edu.sg/docs/wps/wps07_083.pdf; Fedorov, G., Tata, S., Raveslooy, B., et al. (2009). Medical travel in Asia and the Pacific: Challenges and opportunities. Bangkok: UN ESCAP; Herrick, D. M. (2007). Medical tourism: Global competition in health care. *NCPA Policy Reports*. Dallas: National Center for Policy Analysis; and Whittaker, A. (2008). Pleasure and pain: Medical travel in Asia. *Global Public Health: An International Journal for Research, Policy and Practice* **3**, 271–290.

about available services and choices, which can be attributed to the transaction costs associated with medical tourism, where individuals have to assemble their own information and negotiate any treatment. Typically brokers and their web sites tailor surgical packages to individual requirements: flights, treatment, hotel, and recuperation. Brokers may specialize in particular target markets or procedures (treatments such as dentistry or cosmetic surgery), or destination countries (e.g., Poland and Hungary). Medical tourist facilities will often target particular cultural groups – Bumrungrad in Thailand, for example, has a wing for the Middle East patients. Within the wide picture of medical tourism there is a diversity of participating providers. Relatively small clinical providers may include solo practices or dual partnerships, offering a full range of treatments. At the other end of the scale are extremely large medical tourism facilities (e.g., Bumrungrad, Raffles in Singapore, and Yonsei Severance Hospital in South Korea) where clinical specialism is the order of the day. Providers are primarily from the private sector, but are also drawn from some public sectors (e.g., Singapore and within Cuba).

Box 4 Case study 4

India

Estimated size of the market	Conservative estimates suggest an inflow of at least 200 000 medical tourists, though much higher estimates are available
Average potential cost savings for the consumer	From 30% to 90% depending on the procedure
Main clinical areas	Cosmetic, fertility, orthopedic, and cardiac
Major pull factors	Cost and quality
Role of state	Key role in promoting the industry. Legislative role in reducing barriers to the industry

The Indian medical tourism industry is a major exporter of medical tourist services. Initially built around treating medical tourists from West Asia and the Middle East, India is beginning to attract an ever-increasing number of European and American medical tourists.

As with most other lower-middle income countries provider countries, cost is a key driver of inward medical travel – cardiac, cosmetic, and even dental procedures can be found between 60% and 90% cheaper than US prices.

Central to the expansion and dominance of the Indian medical tourism market have been both private and public endeavors. The Indian government has relaxed visa laws, introducing a special visa category – an M visa – to cater for the growing number of medical tourists as well as allowing tax breaks to providers. In the private domain, the Apollo Group controls some 50 hospitals and markets these rigorously on the basis of their quality, safety, and in many cases accreditation or partnership with US hospitals. One particular advantage the Indian market has over its Thai and Singaporean counterparts is the large Indian Diaspora, which is increasingly being targeted by marketing campaigns.

Source: Reproduced from Chee, H. L. (2007). Medical tourism in Malaysia: International movement of healthcare consumers and the commodification of healthcare. *ARI Working Paper* (Online) **83**. Available at: http://www.ari.nus.edu.sg/docs/wps/wps07_083.pdf; Ramírez de Arellano, A. B. (2007). Patients without borders: The emergence of medical tourism. *International Journal of Health Services* **37**, 193–198; and Whittaker, A. (2008). Pleasure and pain: Medical travel in Asia. *Global Public Health: An International Journal for Research, Policy and Practice* **3**, 271–290.

Hospitals may be part of large corporations (the Apollo Group, e.g., has 50 hospitals within and outside India), and ownership itself may lie primarily in the higher income countries from where patients mostly originate.

Countries seeking to develop medical tourism have the options of growing their own health service or inviting partnerships with large multinational players. Individual hospitals may develop relations with travel agencies or wider brokerage companies. Securing accreditation from international programs may be a part of the development of services. In addition to accreditation, other approaches to raising the profile of countries and their health facilities have been used. For example, partnerships and oversight by overseas hospitals and universities, most often from the American private sector, can fulfill a similar role. Formalized linkages with widely recognized medical providers and educators (such as Harvard Medical International and Johns Hopkins Hospital) are becoming increasingly popular among hospitals in middle-income countries catering for medical travelers. A long-standing approach of the Cleveland Clinic, is to train foreign physicians as house staff and fellows to encourage later patient referrals back to the US once they are practicing medicine in their home countries.

The Role of State Support

A range of national government agencies and policy initiatives have sought to stimulate and promote medical tourism in their countries. Many countries see significant economic development potential in the emergent field of medical tourism. Thai, Indian, Singaporean, Malaysian, Hungarian, and Polish governments have all sought to promote their comparative advantage as medical tourism destinations at large international trade fairs, via advertising within the overseas press, and official support for activities as part of their economic development and tourism policy (see **Boxes 2–4**).

Government support does manifest itself slightly differently across the medical tourism map. However, common features are the relaxation of visa regulation, promotion of medical tourism within the central ministries of tourism, support for hospitals to achieve accreditation from bodies such as the Joint Commission International (JCI), and a willingness to provide funding. Useful examples can be found in both Singapore and India. Since 2003, SingaporeMedicine has been a multiagency government–industry partnership aiming to promote Singapore as a medical hub and destination for advanced patient care. It is led by the Ministry of Health, and has the support of the Development Board (new investments and health care industry capabilities); International Enterprise Singapore (growth and expansion of Singapore's health care interests overseas); Singapore Tourism Board (branding and marketing of its health care services) (see **Box 4** for details on India).

As the case studies illustrate, some places may be simultaneously acting as countries of origin and destination in the medical tourism marketplace (e.g., the US). High income countries may service overseas elites, whereas at the same time their citizens choose to travel as medical tourists to lower and middle-income countries for treatments (e.g., India and Thailand). Thus, Harley Street in the UK and facilities including the Mayo Clinic and Cleveland Clinic in the US have long-standing reputations in the international provision of health care. Conversely, the emergence of lower cost

Box 5 Case study 5

USA

Estimated size of the market	An estimated 43 000–103 000 foreigners travel to the US and 50 000–121 000 US residents travel out. Estimates are as high as 500 000 for outward flows
Average potential cost saving for the consumer	Costs in the US are among the highest in the world
Main clinical areas	Cosmetic, fertility, orthopedic, and cardiac
Major pull factors	Quality
Major push factors	Cost and access

The US medical tourism market is more complex than any other, owing much to the fact that it is both an importer and exporter of medical tourists. There are three main import drivers. The first is access to health care within the US. With an estimated 46 million uninsured Americans and one of the most expensive health care systems in the world, many Americans can simply not afford surgery at home. Estimates are as high as half-a-million Americans leave the US as medical tourists every year, largely to Central and South America, but also increasingly to India and the East Asia.

Second, the large number of immigrants in the US, many of whom have demonstrated a preference to return to their home country for care. In particular, the large Mexican Diaspora is an increasing source of outward medical tourists, with some US-based companies even extending their health care cover to procedures undertaken in Mexico.

Third, changing attitudes as seen with the increasing outsourcing of the US medical industry. Although the use of foreign-based radiologists to provide overnight cover for US radiologists may not be directly bound to burgeoning medical tourism market, it is indicative of the changes occurring within the US market and shifting perceptions. There is an increasing acceptance that lower costs elsewhere do not necessarily signal lower quality. A range of employers are sanctioning its employees to seek medical care abroad within the coverage of their health care policies. Many of the facilities treating American medical tourists abroad are controlled by or in partnership with US corporations, many of which are providers of domestic health care.

The motivations for non-US citizens traveling to the US vary, however the high quality of care, international reputation of flagship health care facilities, and high waiting times in neighboring Canada are thought to be key drivers.

Source: Reproduced from Herrick, D. M. (2007). Medical tourism: Global competition in health care. *NCPA Policy Reports.* Dallas: National Center for Policy Analysis; Keckley, P. H. and Underwood, H. R. (2008). Medical tourism: Consumers in search of value. Washington: Deloitte Center for Health Solutions; Johnson, T. J. and Garman, A. N. (2010). Impact of medical travel on imports and exports of medical services. *Health Policy* **98**, 171–177; Terry, N. P. (2007). Under-regulated health care phenomena in a flat World: Medical tourism and outsourcing. *Western New England Law Review* **29**, 421; and Whittaker, A. (2008). Pleasure and pain: Medical travel in Asia. *Global Public Health: An International Journal for Research, Policy and Practice* **3**, 271–290.

treatments in Thailand, India, or parts of Europe will attract individuals from higher income countries who pursue treatments on the basis of cost. The remainder of the article discusses the impacts on health systems for exporter and importer countries. A country imports if their patients go overseas to receive care, and exports if they themselves provide care to inward medical tourists.

Implications of Importing Medical Tourist Services

There are a range of potential financial impacts for publicly funded health care in countries importing medical treatments. Costs may result from overseas cosmetic surgery or dental work that requires emergency or remedial treatment within home countries. Infection outbreaks resulting from travel will also bear upon the public health system. Similarly, there may be health and social care costs that arise from multiple births, as a result of overseas fertility treatments, particularly if facilities use more 'risky' procedures. Domestic private health activity may also experience costs, given that they potentially lose business to overseas providers, for example, cosmetic surgery and fertility treatment. National regulators may incur associated costs of patients traveling overseas caused by monitoring advertising and providing detailed information and advice to support potential or actual medical tourists. But overall, there has been little systemic collection of evidence or attempts to estimate system costs and knowledge is fragmented.

Large numbers of medical tourists traveling overseas will impact on the source country's own health system, perhaps increasing trends that are encouraged by the current domestic private provision. For example, outflows of high-income patients from low- and middle-income countries will reduce revenue and dilute political pressure for investment in particular facilities and technology. Indeed, outflows of medical tourists for treatments that could be provided locally could signal a failure of policy and delivery in sender countries. There are suggestions that the target market for South Africa's breast cancer treatment is a growing pool of middle-class women drawn from across the African region with financial means, but who experience failed domestic policy. Regarding travel from higher income countries, if eligibility for services such as fertility or dental work is tightened, then those with private resources may choose to travel overseas to maintain access. The ability to circumvent waiting times raises issues of equity. However, travel overseas for treatments that are not provided or are illegal within the source country may normalize such treatments and generate debate about the importance of providing them locally (e.g., latest fertility treatments, gender reassignment, organ transplantation, or even euthanasia services).

In countries where Third Party insurers are exploring medical tourism as a provider option, outflows of patients

may benefit employers and employees contributing to health plans, and the public insurance system itself. Opportunities for financial benefit may be consolidated if medical tourism becomes an outsourcing option. For the US, research has estimated that 15 treatments would show savings of US$1.4 billion annually if one in ten US patients chose to undergo treatment abroad. Similarly, a recent study looking at possible bilateral medical tourism trade between the UK and India demonstrated substantial savings could accrue to the UK National Health Service from sending its patients to India, both financially and in alleviating waiting lists. If one takes the waiting lists for a selected number of procedures suitable for medical tourism, and compares the cost of sending those patients (plus an accompanying adult) to India, with the costs of getting treatment in the UK, the savings would be of £120 million. Some subsets of the population, such the Indian Diaspora, may prefer to go back 'home' for treatment, and may be happy to cross-subsidize some of the costs.

There are arguments that some medical systems are inefficient and face restrictive barriers to entry. A development such as medical tourism can potentially exert competitive pressure on systems importing health care and help drive down the costs and prices offered in domestic systems. Medical tourism may encourage economies to maximize their comparative advantage across labor costs, utilization of technology, and spare capacity. Indeed the US employers are said to be encouraging workers to travel domestically for medical care – a development prompted by deals struck with overseas providers being used as leverage. The possibility of medical tourism resulting in underused capacity in American hospitals has also been raised.

One of the implications of globalization is the increased flow of clinical and ancillary staff around the globe. Individuals may fully or part-train in their home country and move overseas to continue their training and gain experience with a particular specialism. Of major concern has been the flow from low to high income countries. Medical tourism may provide opportunities for professional migrants to return home – so-called 'reverse brain drain' or 'brain circulation.' This may be a disbenefit for developed countries which have long relied on such expertise to underpin their health system.

System Implications for Exporting Countries

The main exporting countries (those who provide the services to medical tourists) are located across all continents, including Latin America, Eastern Europe, Africa, and Asia. Countries have specialized in certain procedures. For instance, Thailand and India specialize in orthopedic and cardiac surgery (Boxes 3 and 4), Brazil is famous for cosmetic surgery (as outlined in Box 1), and Hungary (Box 2) and Poland are hotspots for dental surgery. As the US case illustrates, all countries may possibly be source and destination countries for medical tourists. However, here we frame low- and middle-income countries as the destination, and high-income countries as source. The magnitude of the possible effects being discussed is largely unknown – typically the potential or actual occurrence of these effects has been observed, but the scale of effect, and how this scale may differ between countries is an unknown quantity.

Economic Impacts

Delivering care to medical tourists will likely increase the level of direct foreign exchange earnings coming into a country and improve the balance of payments position. There are suggestions that Thailand benefits between US$1.5 and 2 billion from medical services and approximately US$0.5 billion from related tourism – overall total value added is 0.4% of gross domestic product. Income from foreign patients can be used within hospitals and national systems to cross-subsidize care for domestic patients, or could be used to help fund capital investment for use by all patients within the hospital or health system. Similarly, there are suggestions that the Cuban experience is to reinvest income from foreign patients into the national system for broader public good. International patients will have multiplier effects – a RAND study of Cleveland's metropolitan economy highlighted the economic benefits that the Cleveland Clinic added to the local economy.

Economic implications vary depending if international patients are simply using spare capacity or competing with domestic patients. For instance, the push by Thailand to be a hub for medical tourists in the 1990s was a result of the economic crisis in Asia generating a fall in domestic private patients and hence spare capacity in their private sector. In this case, increasing foreign patients entailed a net benefit to the private health system with substantial income and little real opportunity cost. However, where capacity has to be developed, there are substantial potential costs not only in financial terms but also in the wider context of concerns around equity, access, and human resources.

Although medical tourism generates income for the health sector (physicians hospitals, medications, and medical devices), general increases in tourist income (airfares, food, hotels, and souvenirs) are also important. There is a substantial level of expenditure by medical tourists, and their companions, that is not related to medical care. For example, it is estimated that companions would spend approximately twice as much on hotels and tourism as the patient. As discussed earlier, the promise of these earnings often drives the government involvement in investing directly or indirectly (tax incentives) in private hospitals and actively promoting medical tourism. Sectors other than medical care – especially those associated with hospitality and travel – may benefit to some degree from increased medical tourism, as will the government more centrally through increased taxation revenue. However, global business models and the involvement of Transnational Corporations may result in profits from medical tourism and ancillary activities being remitted overseas.

In many instances, medical tourists are either Diaspora or patients who have previously visited the country and are likely to visit again (an estimated 2.2% of foreign travelers and 10% of nonresident Indians visited India with the objective of health treatment). Thus, they are 'regular' visitors who on one trip incorporate an element of medical care. In this situation clearly the additional income generated by the 'medical' element of medical tourism is far more limited.

There are financial costs associated with promoting medical tourism – including upgraded infrastructure, both within the health sector (e.g., hospital facilities) and beyond (roads, airports, and telecommunications). There are also likely to be costs concerned with the appropriate staffing of facilities (including taxpayer's subsidized education and training), and possible accreditation schemes. For instance, 48 countries have been granted accreditation from the US-based JCI, the international arm of the Joint Commission, which accredits US hospitals. India has already sought and obtained JCI accreditation for 17 hospitals, and Thailand for 14. Other international accreditation bodies include the Australian Council for Healthcare Standards, the Canadian Council on Health Services, and QHA Trent Accreditation. However, there are costs associated with ensuring compliance with these various criteria, maintenance of these accreditations, and the processing costs themselves.

Trickle Down Benefits

There are arguments around 'trickle down' of best practice and technological diffusion as benefitting countries providing medical tourism. The increased ability to purchase the latest technology, for example, and treating foreign patients may broaden the case-mix for staff, or increase throughput to enable them to become more skilled. Medical tourism may be linked to temporary secondments to overseas facilities, which may lead to enhancement of human capital. Increased quality may result through ensuring compliance with (higher) international standards for care.

However, there is the possibility of resources being diverted from the domestic population and invested into private hospitals; such as driving investment toward urban tertiary care rather than rural primary care centers, which more appropriately reflect domestic population needs. It is argued that a number of young professionals are drawn to specialist facilities catering for medical tourists. The focus of resources on high technology orthopedic, dental, and reproductive care, rather than more basic public health measures to tackle infectious disease may be disadvantageous for the local population.

Human Resource Implications

There are arguments that medical tourism provides exporting countries the opportunity to attract back to their home country health workers who had emigrated, thus reversing the 'brain drain' of professional mobility. Hospitals treating medical tourists can offer higher salaries and wider opportunities more comparable with overseas institutions. International patients are more likely to trust doctors who have trained or practiced in their countries of origin, as well as ensuring that human resources are brought back to the country or are less persuaded to leave. The empirical veracity of this effect remains unclear however.

There are concerns that medical tourism will cause an internal brain drain, with health professionals abandoning the public health system to work for the hospitals that attract medical tourists, lured by better salaries and work opportunities. There are longer consultation times for foreigners – so generating additional demand for physicians (mostly specialists). Suggestions are that for India there is a shortage of 600 000 doctors, 1 million nurses, and 200 000 dental surgeons. Thus medical tourism would decrease the quality of the public health system and doctor-to-patient ratio. (Given that medical education in countries such as Thailand is heavily tax subsidized, the medical tourism market also presupposes state activity and investment.) As with other aspects of medical tourism, there is little empirical evidence of whether this is happening, and to what extent; and what there is unclear.

Two-Tier System

Do foreign patients benefit from sophisticated private hospitals with a high staff-to-patient ratio and expensive, state-of-the-art medical equipment, whereas the local population only has access to basic, under-resourced health facilities? Certainly there is the potential for medical tourism to have effects in terms of the distribution of health care resources for the less well-off local population. There have been various accusations that in some countries private sector medical tourists may be accumulating medical resources and taking health care services and personnel away from the local population and by driving up prices in the private sector. State regulation can mitigate impacts; however, there is potentially incoherence between trade and health policy that promotes both medical tourism and universal coverage.

Although private hospitals in India may have a responsibility under the Public Trust Act to provide free health care to the extent of 20% of resources, there are no checks undertaken to ensure that this occurs and others have suggested that Indian hospitals renege on promises to provide free health care. Nonetheless, as with much in this area, there is no strong evidence that medical tourism exacerbates a two-tier system.

Conclusion

Medical tourism for source countries may alleviate waiting lists and reduce health care costs, but there is a quality of care risk. The impact of medical tourism on health care systems is not well understood for destination countries: finance, delivery, organization, and regulation. Evidence is limited and there is a necessity to develop a robust empirical base on a number of these issues. In compiling data we must scrutinize sources and surveys used to provide numbers, including the role of national agencies and private facilities. Extrapolating from a country to a more global perspective is difficult, as is ensuring 'the count' is appropriate (do we count patients or treatment episodes; day treatments or in-stay treatment; expatriates and those funded by their multinational employers; only large and accredited providers?).

Beyond health policy and management there are also key legal and ethical issues for medical care abroad – informed consent, liability, and legislating for clinical malpractice. Choosing an overseas treatment center brings a number of challenges – difficulties in assessing comparative quality and

performance of alternative providers, differences in legal liability and knowledge concerning the processes of how to pursue complaints and receive redress. There are complexities regarding who or what could be subject to legal proceedings – product advertising, initial Internet consultation, a brokerage service, surgery itself, and various mixes therein – the jurisdiction of hearing any case, and the country's law that should govern any case.

Research and evaluation has not kept pace with the development of medical tourism. The lack of data is significant because countries face difficulties keeping fully informed about the significance (potential or actual) of medical tourism for their health systems. Mechanisms are needed that help us track the balance of trade around medical tourism on a regular basis. This would allow us to add to the evidence-base by assessing who benefits and loses out at the level of system, program, organization, and treatment.

See also: Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity. International Movement of Capital in Health Services. International Trade in Health Services and Health Impacts. International Trade in Health Workers. Understanding Medical Tourism

Further reading

Carabello, L. (2008). A medical tourism primer for U.S. physicians. *Medical Practice Management* **23**, 291–294.

Connell, J. (2006). Medical tourism: Sea, sun, sand and... surgery. *Tourism Management* **27**, 1093–1100.

Ehrbeck, T., Guevara, C. and Mango, P. D. (2008). Mapping the market for medical travel. *The McKinsey Quarterly* (Online). Available at: https://www.mckinseyquarterly.com/Mapping_the_market_for_travel_2134 (accessed 18.12.12).

Lunt, N. and Carrera, P. (2011). Advice for prospective medical tourists: Systematic review of consumer sites. *Tourism Review* **66**, 57–67.

Lunt, N., Exworthy, M., Green, S., et al. (2011). *Medical tourism: Treatments, markets and health system implications: A scoping review.* Paris: OECD.

Smith, R., Martínez Álvarez, M. and Chanda, R. (2011). Medical tourism: A review of the literature and analysis of a role for bi-lateral trade. *Health Policy* **103**, 276–282.

Smith, R. D., Lee, K. and Drager, N. (2009). Trade and health: An agenda for action. *Lancet* **373**, 768–773.

Smith, R. D., Rupa, C. and Viroj, T. (2009). Trade in health-related services. *Lancet* **373**, 593–601.

Medicare

B Dowd, University of Minnesota, Minneapolis, MN, USA

What is Medicare?

The US Medicare program began in 1965 to address issues of access to care for three groups of Americans: the aged, the disabled, and people with end-stage renal disease (ESRD) or amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease). (The Medicaid program, also begun in 1965, covers low-income Americans.) Enrollees in Medicare are referred to as beneficiaries. Some beneficiaries are eligible for both Medicare and Medicaid and are referred to as dually eligible beneficiaries.

Beneficiaries become aged eligible for Medicare at the age of 65 years. Those under the age of 65 years can qualify for Medicare if they are entitled to disability benefits under the Social Security Disability Insurance (SSDI) or Railroad Retirement. ESRD eligibility is open to Americans who have ESRD or ALS and who have met the required work credits for, or are receiving, Social Security or Railroad Retirement benefits, or are the spouse or dependent child of a person receiving those benefits.

The Medicare program is divided into four different 'Parts,' due largely to historical and political influences. Part A covers primarily hospital charges, but also home health, hospice care, and care in skilled nursing facilities. Part B covers primarily medical expenses (including many physician charges incurred during hospitalization). Part C covers the payment of private health plans in Medicare – originally referred to as Tax Equalization and Fiscal Responsibility (TEFRA) risk plans, then Medicare + Choice, and now Medicare Advantage (MA) plans. For many years, Medicare did not cover outpatient prescription drugs, except through some MA plans. That changed with the introduction of Medicare Part D in the 2003 Medicare Modernization Act (MMA) legislation. Part D coverage was first offered in 2006. Long-term nursing home stays are not covered by any part of Medicare.

The basic Medicare entitlement benefit package reflects health insurance coverage in the USA in 1965. In 2013, beneficiaries face a deductible of $1184 per hospital stay during a 'benefit period,' plus $296 per day for days 61 through 90, $592 per 'lifetime reserve day' over day 90 for each benefit period up to a maximum of 60 lifetime reserve days. Part D has a complex coverage structure that includes a coverage gap (referred to as a donut hole) that extends from total spending equal to $2970 to $6733.75 in 2013.

The entitlement benefit package, which reflects typical insurance benefits in 1965 when the program was established, has been criticized for its meager level of coverage. Medicare beneficiaries face the possibility of unlimited out-of-pocket expenses. As a result, approximately 90% of beneficiaries in fee-for-service (FFS) Medicare purchase private supplementary insurance or 'Medigap' policies that cover the coinsurance and deductibles for Medicare-covered services. These policies add to total Medicare costs because the effect of supplementary insurance is to reduce point-of-purchase cost sharing – in some cases to zero. The result is increased demand for services, and the Medicare program picks up approximately 80% of that additional cost, whereas the premiums for supplementary insurance reflect only the 20% paid by the supplementary insurer. Reform of the Medigap market is a topic of perennial policy interest, but there have been no substantial changes till date.

Although Medicare often is thought of as a 'public' insurance plan, its administration and delivery of services occur primarily through private health-care providers, health plans, and claims-processing firms under government contract. The traditional Medicare program contracts with private firms to pay health-care providers on a FFS basis and thus is referred to as FFS Medicare. Beneficiaries enrolled in MA plans are referred to as MA enrollees. Part D coverage is offered either through MA plans or by private stand-alone companies selling outpatient prescription drug coverage.

MA enrollees pay the same Part B premiums as beneficiaries in FFS Medicare, and MA plans receive a per-capita rate for each beneficiary they enroll. The level of capitation payment has been a source of controversy over the years. MA and Part D plan switching is limited to annual open enrollment periods.

How is the Medicare Program Financed?

There are two trust funds through which Medicare funds flow: the Hospital Insurance (HI) and Supplementary Medicare Insurance (SMI) trust funds. Part A expenses and the costs of program administration appear in the HI Trust Fund, whereas expenses for Parts B and D appear in the SMI Trust Fund. Administrative costs of running the Medicare program are represented in trust funds. The trust funds are administered by a Board of Trustees for Medicare that issues annual reports describing the state of the funds.

The trust funds are used solely for accounting purposes. Revenue and expenses do not actually flow through the trust funds. However, the HI Trust Fund balance is 'real' in the sense that Part A Medicare payments cannot be made if there are not adequate funds credited to the HI Trust Fund.

Part A is financed primarily through payroll taxes. The current payroll tax rate is 2.9%. Parts B and D are financed primarily through general taxes and beneficiary premiums. Beneficiary premiums are set at 25% of Part B and Part D costs. Part B premiums are means tested (increase with income) for beneficiaries whose incomes exceed $85 000 or $170 000 for a couple. The base-level monthly Part B premium in 2013 was $104.90 but rose to $335.70 for beneficiaries with incomes greater than $214 000, or couple with incomes greater than $428 000. The Part B deductible in 2013 was $147 per year.

Because Part A revenue is limited by the total level of taxable payroll income for the nation's workers, it is technically possible for the Part A trust fund to become insolvent. Parts B

and D have access to general tax revenue and thus they cannot become insolvent in the same sense as Part A, although increased Part B and Part D spending can limit other types of government spending. Thus, the most closely watched portion of the annual trustee report is the projected number of years until Part A payments will exceed the amount of money credited to the HI Trust Fund. That time span has varied from as little as 2 years in the early 1970s to the high 20s in early 2000s.

Since the introduction of capitated private health plans in the mid-1980s, payments to MA plans have been based roughly on a percentage of the cost of caring for beneficiaries in traditional FFS Medicare in the same county, adjusted for a set of risk factors. Because the average cost of caring for beneficiaries in FFS Medicare varies dramatically from one county to another, even when the counties are geographically proximate and after adjusting for the health risk of beneficiaries, government payments to MA plans also have varied dramatically.

MA plans must provide the entitlement benefit package, but are free to use any excess revenue to offer additional supplementary benefits to their enrollees. As a result, the amount of supplementary benefits offered by MA plans also varies dramatically from one county to another. In counties with low FFS spending levels, beneficiaries must pay an additional out-of-pocket premium above and beyond the Part B premium for the same benefits received for free by beneficiaries in counties with high FFS spending. The geographic disparity in government-financed benefits is a source concern due in part to issues of equity and in part to the lack of evidence that higher FFS spending is associated with either higher health risk or better health outcomes among FFS Medicare beneficiaries.

Expenditures and Cost Trends in Medicare

In 2010, approximately one-third of Medicare spending was for services covered only by Part A, whereas one-fifth was for services covered only by Part B. Payments to MA plans accounted for 23% of spending and Part D expenditures were approximately 11% of spending. Areas of particularly high growth in recent years have been home health-care and imaging services.

Cost concerns surfaced in the earliest days of the Medicare program. In 2012 Medicare's unfunded obligation over the next 75 years was $27.2 trillion. The US federal debt currently exceeds the US gross domestic product. As the mass of Americans born after World War II ('baby boomers') begin to reach the age of 65 years, the Medicare population will roughly double. Taxable income, particularly payroll income, will fall dramatically as the 'boomers' retire, as will the number of workers per Medicare beneficiary.

The geographic variation in FFS Medicare costs remains a controversial topic. Some analysts have found that much of the variation in cost is unrelated to either the health status or health outcomes of beneficiaries, whereas others find that sociodemographic characteristics explain a large proportion of the variation.

There have been several important responses to cost concerns over the life of the Medicare program. During the early 1980s, Medicare switched from cost-based reimbursement for inpatient care to a prospective rate per admission for Part A expenses. The Resource-Based Relative Value Scale introduced in the late 1980s was an attempt to reweight the fee schedule toward the work performed by primary care physicians.

Payments to providers and MA plans were reduced significantly in the Balanced Budget Act (BBA) of 1997, but some of the cuts were restored in subsequent legislation. The 1997 BBA legislation introduced the sustainable growth rate (SGR) legislation that mandated across-the-board reductions in physician fees if expenditures on physician fees grew too rapidly. Physician fees actually were cut by 5.4% in 2002, but since then, Congress has found ways to circumvent the mandated cuts. As a result, the SGR legislation now requires roughly a 30% cut in physician fees. A fee cut of that size likely would reduce physicians' willingness to see Medicare patients. Thus, each round of budget negotiations includes discussion of ways that the SGR 'problem' could be 'fixed.' Doing so, of course, would add to Medicare's projected unfunded deficit.

There was some hope in the 1980s that the introduction of capitated private plans would reduce the rate of cost growth in Medicare. Instead, paying MA plans a percentage of FFS Medicare costs, coupled with an inadequate risk adjustment system, resulted in private plans costing the program more money – a problem that continued through the 2000s, when private plans in low-cost areas were given supplementary payments in response to the geographic disparity in supplementary benefits offered by MA plans.

Options for cost reduction are more limited in FFS Medicare than in private health plans. FFS Medicare lacks the statutory authority to intervene in the patient care process, for example, by installing disease management programs except as demonstration experiments. Medicare does have an approval process for new technology, but continues to pay for many treatments that have been shown to provide no medical benefit to beneficiaries. Early attempts to design preferred provider systems in FFS Medicare – systems that featured varying coinsurance rates favoring lower cost or higher quality providers – were stymied by pervasive Medigap insurance that protected beneficiaries from the effect of coinsurance. Medigap also exacerbates FFS Medicare's cost problems by eliminating point-of-purchase cost sharing, thereby increasing the demand for services, and FFS Medicare pays approximately 80% of the cost of those additional services rather than having them reflected in Medigap premiums paid by beneficiaries.

The MMA of 2003 required an annual computation of the percentage of Medicare program revenues that come from general tax revenues. If that percentage was projected to be greater than 45% over the subsequent 7-year period, the Board of Trustees must issue a warning. If such a warning is issued for two consecutive years, the President must propose legislation to reduce the projected percentage to 45%. Although such warnings have been issued every year since 2007, Congress has taken no action.

Future Policy Options

Medicare faces daunting fiscal challenges in the coming years due to inefficiencies in the way the program is administered

and inadequate accumulation of revenue to cover the expenses of the mass of Americans who are beginning to become aged eligible for benefits. Medicare's unfunded obligation is an important part of the entitlement problem in the USA and a major contributor to long-run growth in the federal debt.

Current policy proposals cover the full range of revenue-increasing and cost-reducing options. Tax increases are a contentious alternative during a recession. However, the payroll tax was increased in January 2013, and further increases plus increases in beneficiary out-of-pocket premiums and cost sharing, as well as more aggressive means testing, remain viable but controversial options.

Cost-reducing options must operate through a reduction in the number of beneficiaries or the cost per beneficiary, which in turn is a function of the services delivered to each beneficiary times the unit prices of those services. One way to reduce the number of beneficiaries is to raise the age of eligibility. Current estimates suggest that the savings would be modest. Current legislation calls for significant reductions in provider fees or the rate of increase in provider fees. Hospitals will be penalized for excessive readmissions and hospital-acquired conditions. Payments to 'disproportionate share hospitals' will be cut, as will payments for graduate medical education. Official assessments of the fiscal health of the Medicare trust funds are required to assume that Congress and the administration will adhere to the laws that Congress has passed, even though few analysts actually believe that is the case.

Medicare fees currently average approximately 80% of fees paid by private insurers in the commercial (private) insurance market, whereas fees under the Medicaid program for low-income Americans typically are lower than Medicare fees. The 2010 Patient Protection and Affordable Care Act is estimated to add 33 million Americans to the ranks of the insured, and many of the newly insured will be purchasing private insurance. At the same time, there is no significant change to the supply of services by health-care professionals. Thus, a dramatic increase in the privately insured could result in severely reduced access for Medicaid beneficiaries, followed by difficulties for Medicare beneficiaries.

Although large across-the-board fee cuts in the Medicare program may prove infeasible, there are three other alternatives. The first is increased use of 'bundled payments'; for example, bundling postacute care into the prospective payment for hospital admissions is an approach to reduce excessive use of services. The implementation of prospective hospital payment into diagnosis-related groups adjusted for medical severity (MS-DRGs) provides an encouraging precedent. Of course, the ultimate 'bundle' is capitation payment, as implemented in the MA program. Accountable care organizations (ACOs) are another attempt to place providers at greater risk for the cost of care. One difficulty with the ACO approach in the traditional FFS Medicare program is that beneficiaries remain free to see any provider they like, despite the ACO provider being at some risk for the cost of their care.

A second alternative payment reform is competitive bidding for some health-care services. The Medicare program has had some demonstrated success with competitive bidding for durable medical equipment.

The third alternative is 'value-based purchasing.' The PPACA legislation requires the Medicare program to design value-based purchasing programs for hospital and physician services. Limiting fee increases to providers who meet certain cost and quality benchmarks could provide a way to reduce average fees without compromising access or quality. Assessment of physician quality may improve with full implementation of the Physician Quality Reporting System (PQRS) that allows physicians to report clinical outcomes such as the patient's blood pressure on a standard claims form. Participation in the PQRS system has been voluntary since its inception in mid-2007, and physicians were paid a bonus for participating. Beginning in 2015, however, nonparticipating physicians will face a financial penalty.

The PPACA legislation also establishes the Independent Payment Advisory Board (IPAB), a federal board charged with making recommendations to limit the growth in Medicare spending. Although IPAB has considerable power and discretion – either its recommendations must be implemented by the Department of Health and Human Services, or uHCongress must substitute equivalent cost saving – the scope of measures that it can recommend is severely limited. For example, it cannot recommend denial of care, or changes in eligibility standards, taxes, or beneficiary cost sharing.

There are numerous proposals that involve more basic restructuring of the Medicare program. Medicare beneficiaries currently are entitled to both a specific package of benefits (coverage) and the traditional FFS delivery system with out-of-pocket premiums limited to the national Part B premium, regardless of the actual level of FFS costs in their local market area. Conversely, MA enrollees face an additional out-of-pocket premium in areas with low FFS spending levels and subsequently low MA payment rates. One policy option is to retain the entitlement benefit package but to have both MA plans and FFS Medicare submit bids for that package and set the government's contribution to premiums equal to the lowest or second lowest bid in each county. That would mean that beneficiaries might have to pay an additional premium above and beyond the Part B premium for FFS Medicare, as they do now for MA plans in some market areas. In such a 'competitive pricing' or 'premium support' system, MA plans might have an advantage in areas where high FFS costs were due to inappropriate overuse of services that could be reduced through more aggressive selection of efficient providers or more careful management of care. However, FFS Medicare might have an advantage in areas with higher provider market concentration and subsequent market pricing power, if FFS Medicare's administratively determined fee levels were substantially lower than the prevailing rates charged to private insurers.

Although the list of policy options is broad, its ability to produce a significant reduction in Medicare spending per capita is limited by the resolve of elected officials to make difficult decisions. The history of congressional self-discipline through initiatives such as the SGR and '45 percent rule' is not encouraging. In fairness, however, congressional resolve is limited by the voter's preferences. The US Medicare program faces many of the same long-term challenges as government-provided benefits in other countries – a discrepancy between public demand for more generous benefits and the public's

willingness to pay for those benefits. The problem in the USA is exacerbated by the high unit prices paid by the government and the government's inability to control utilization. The result, to date, has a massive projected unfunded deficit that, in the absence of serious reform, will have to be paid by future generations, born and unborn, who are not able to vote on current policy.

See also: Demand for Insurance That Nudges Demand. Health Insurance and Health. Health Insurance in the United States, History of. Health Insurance Systems in Developed Countries, Comparisons of. Health-Insurer Market Power: Theory and Evidence. Markets in Health Care. Quality Reporting and Demand. Rationing of Demand. Social Health Insurance – Theory and Evidence. Supplementary Private Insurance in National Systems and the USA

Further Reading

Atherly, A., Dowd, B. E. and Feldman, R. (2004). The effect of benefits, premiums, and health risk on health plan choice in the Medicare program. *Health Services Research* **39**(4), 847–864, Part 1.

Bhattacharya, J. and Darius, L. (2006). Does medicare benefit the poor? *Journal of Public Economics* **90**(1–2), 277–292.

Coulam, R., Feldman, R. and Dowd, B. E. (2009). *Bring market prices to medicare: essential reform at a time of fiscal crisis*. Washington, DC: American Enterprise Institute.

Coulam, R., Feldman, R. and Dowd, B. E. (2011). Competitive pricing in Medicare: Can we overcome congressional micromanagement and provider self-interest? *Journal of Health Politics, Policy and Law* **36**(4), 649–689.

Cutler, D. M. and Louise, S. (1999). The geography of medicare. *American Economic Review* **89**(2), 228–233.

Dowd, B. E., Maciejewski, M. L., O'Connor, H., Riley, G. and Geng, Y. (2011). Health plan enrollment and mortality in the medicare program. *Health Economics* **20**(6), 645–659.

Finkelstein, A. (2007). The aggregate effects of health insurance: Evidence from the introduction of medicare. *Quarterly Journal of Economics* **122**(3), 1–37.

Fisher, E. S., Wennberg, D. E., Stukel, T. A., et al. (2003). The implications of regional variations in Medicare spending. Part 1: The content, quality and accessibility of care. *Annals of Internal Medicine* **138**(4), 273–288.

Fisher, E. S., Wennberg, D. E., Stukel, T. A., et al. (2003). The implications of regional variations in Medicare spending. Part 2: Health outcomes and satisfaction with care. *Annals of Internal Medicine* **138**(4), 288–299.

McClellan, M. (2000). Medicare reform: Fundamental problems, incremental steps. *Journal of Economic Perspectives* **14**(2), 21–44, Spring.

McClellan, M. and Skinner, J. (2006). The incidence of Medicare. *Journal of Public Economics* **90**, 257–276.

McGuire, T. G., Newhouse, J. P. and Sinaiko, A. D. (2011). An economic history of Medicare, Part C. *Milbank Quarterly* **89**(2), 289–332.

Miller, R. H. and Luft, H. S. (2002). HMO plan performance update: An analysis of the literature, 1997–2001. *Health Affairs* **21**(4), 63–86.

Pizer, S. D. and Frakt, A. B. (2002). Payment policy and competition in the Medicare + Choice program. *Health Care Financing Review* **24**(1), 83–94, Fall.

The Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds (2012). *Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds*. Washington, DC: The Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds.

Thorpe, K. E. and Atherly, A. (2002). Medicare + Choice: Current role and near-term prospects. *Health Affairs* W242–W252, Web exclusive.

Mental Health, Determinants of

E Golberstein, University of Minnesota School of Public Health, Minneapolis, MN, USA
SH Busch, Yale School of Public Health, New Haven, CT, USA

Glossary

Dysregulation of mood An emotional response that is not within the conventionally accepted range of responses.
Perceived stress The degree to which general situations in life are judged by individuals to be stressful. This is in contrast with measures of the frequency or types of stressful events experienced by individuals.
Psychotic disorder A severe mental disorder that causes abnormal thinking and perceptions. Symptoms include delusions and hallucinations.
Unipolar depressive disorder A mental disorder characterized by episodes of low mood with a loss of interest or pleasure in normally enjoyable activities.

Introduction

Mental illness is a common occurrence. Epidemiological evidence reveals that mental disorders are prevalent across more- and less-economically developed countries (WHO World Mental Health Consortium, 2004), and some mental health problems have been understood as being an illness since the time of Hippocrates. Mental disorders are known to have major consequences for longevity, quality of life, and productivity. For instance, the World Health Organization estimates that unipolar depressive disorders account for the third-largest share of lost disability-adjusted life years worldwide. There is a growing recognition that mental health disorders are also associated with reduced life expectancy. Recent estimates from the US suggest that 26% of nonelderly adults experienced a diagnosable mental disorder in the past 12 months (Kessler *et al.*, 2005a). In the US, individuals with a mental health disorder die on average approximately 8 years younger than individuals with no mental health disorder, and 95% of these deaths were from internal causes (i.e., cardiovascular disease, cancer, pulmonary disease). Less than 5% of deaths were from external causes of suicide, homicide, or accidents, similar to the rate in the population reporting no mental health disorder.

A comprehensive multidisciplinary overview of the determinants of mental health, including genetic and other neurological bases for disease, is covered in the landmark Surgeon General's report on mental health (US DHHS, 1999). Although the authors briefly review some of these concepts, their focus in this article is on the contributions of economists to this literature.

Mental illness includes a broad range of specific disorders, which are distinct from many physical health disorders in that there is no definitive diagnostic test for mental illness. Rather, mental illness is defined by the existence and severity of a set of symptoms that may include inappropriate anxiety, disturbances of thought and perception, dysregulation of mood, and cognitive dysfunction. The authors briefly review some of the more common and prominent examples of mental illness.

Anxiety disorders (including phobias, panic attacks, and generalized anxiety) are characterized by an individual's anxiety responses to a given situation substantially exceeding what is emotionally and/or physiologically appropriate. Psychotic disorders (such as schizophrenia) involve serious disruptions of perception and thought process, which manifest in symptoms of hallucination, delusion, disorganized thoughts, flat affect, and inability to think abstractly. Mood disorders include depression, which is characterized by persistent symptom of sadness that is often associated with physical symptoms of insomnia, decreased appetite, and low-energy, and bipolar disorder, which is characterized by extreme fluctuation between depressed mood and elated mood. Impulse-control disorders include attention-deficit and hyperactivity disorder, and are frequently associated with childhood and adolescence. Cognitive disorders affect the ability to organize, process, and recall information. Perhaps the most prominent example is Alzheimer's disease which is progressive and generally is associated with aging (often more so than with broader mental illness).

There is an important difference between mental illnesses such as schizophrenia, where there is a clear binary categorization of having versus not having the disease; from other types of mental disorders such as depression, anxiety, or attention-deficit disorder, where the symptoms lie along a continuum from manageable and self-limiting to profoundly and persistently disabling. Analogous physical health conditions are, respectively, cancer, where there is a clear difference between people with and without the disease, and back pain, where many people experience symptoms, but only relatively few have their functioning disrupted by those symptoms. Nevertheless, all categories of mental disorders include conditions that range in severity, from relatively mild to profoundly severe.

Some view mental health status and happiness as lying along a common continuum. This article treats mental illness as a disease. This conceptual distinction is especially important in the context of studying economic issues in mental illness, as the idea of happiness is closely linked with the notion of utility that underlies classical consumer theory. This distinction is made clear by explicitly considering the idea of a mental health production function.

The pioneering work on health production functions (Grossman, 1972) was formulated from the perspective of physical health conditions but one needs to consider how it

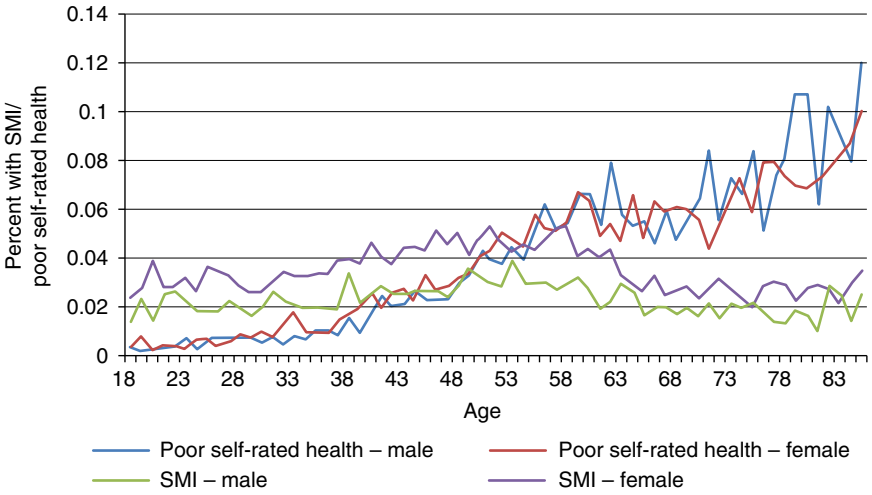


Figure 1 Percent with SMI/poor self-rated health, by age and gender. National Health Interview Survey, 2000–09. SMI determined by K6 score.

can be applied to mental health. Individuals are posited to have been born with an initial endowment of health capital. Health capital is a durable capital stock with two types of value: as a consumption good, where good health directly improves current utility, and an investment good, where health status affects other economic activity (such as employment). The stock of health capital at a point in time is determined by the stock of health capital at the previous point in time, health investments (including both behaviors and medical care), random shocks, and a depreciation term.

Physical health and mental health share some similarities in the context of this model, yet have some differences. Similar to the case of physical health, there is evidence of heterogeneity in initial endowment of mental health. For example, economists are beginning to contribute to an existing body of psychology and neuroscience research on the effects of stress *in utero*. Research finds that stressful conditions during pregnancy (particularly the first trimester) lead to significant increases in subsequent mental health problems. One study finds that maternal stress during pregnancy doubles the risk of schizophrenia in offspring (Malaspina *et al.*, 2008). In addition, similar to physical health, a range of treatment technologies exist which can improve mental health status between time periods.

Nevertheless, mental health contrasts in some important ways from physical health in the context of this model. The notion of intertemporal depreciation does not fit neatly with mental health. The Grossman model assumes that health depreciates at an increasing rate with age. This assumption fits the data fairly well for the case of physical health, as illustrated in Figure 1 (measured here as the proportion of the population reporting poor self-rated health). However, mental health (measured here as the proportion of the population with severe mental distress) exhibits a relatively flat, inverse U-shaped pattern in age, consistent with epidemiological studies that find that most mental illnesses first occur early in life. On average mental health depreciates only moderately until the mid-50s, and then improves moderately at older ages.

Another way that mental health departs from the classic health capital model is in the type of health investment inputs. Similar to general health, health care can affect mental health in ways that are broadly similar to physical health with many evidence-based treatments available. For example, pharmaceutical treatments can improve symptoms of anxiety, major depression, and schizophrenia as well as other mental health disorders. Brief psychotherapy is also effective in treating acute cases, as well as extending periods of remission. Also, like general health conditions, health behaviors (e.g., exercise) can affect mental health. Yet, it is believed that psychosocial stress plays a relatively larger role in the production of mental health than of physical health. Psychosocial stress has been studied extensively by research in psychology and sociology, and a review of this literature is outside of the scope of this article. But in brief, these fields have produced striking evidence on the effects of psychosocial stress on mental disorders, and have identified moderators and mediators of this relationship. More recently, some sources of psychosocial stress have been studied by economists, as described below in section Employment.

Finally, it is noteworthy that any economic model of health status that is derived from classical consumer theory is faced with the challenge that in many cases, mental illness represents a break from 'rational' behavior. Indeed, the departure from an individual's normal capacity for decision making (e.g., the compromised perception and thought processing that are common in psychosis) and changes to an individual's preferences (e.g., not caring about the future is a symptom of major depressive disorder and can be interpreted by an economist as a change to one's discount rate) are hallmarks of mental illness and can violate the axioms of expected utility theory.

Economists have focused most of their interest on three specific (related) determinants of mental health: income, labor market participation, and macroeconomic conditions. In the rest of this article the authors discuss findings related to these inputs. Like other health disorders, although risk factors have been documented, much of the heterogeneity in mental health disorders and outcomes remains unexplained.

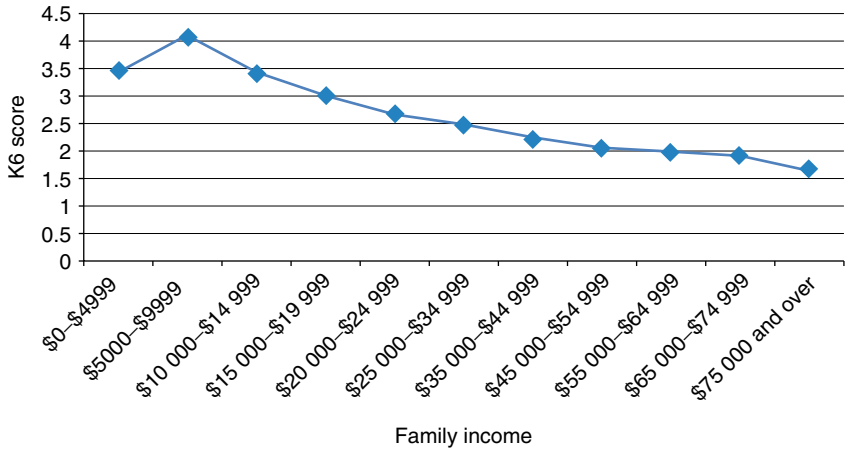


Figure 2 K6 scores, by income. National Health Interview Survey, 2000–06.

Income

Disentangling the effects of income on mental health from the reverse effect is difficult because mental health disorders, like other health conditions, are also likely to have a direct effect on income through labor market outcomes, and perhaps household formation. **Figure 2** uses data from the National Health Interview Survey to examine the correlation between mental health (measured by K6 score) and household income in the US. This figure illustrates the strong correlation between income and mental health. However, epidemiological evidence suggests that although the within country correlation between income and mental health is strong, there is a weaker correlation between income and mental health across countries and some evidence suggests a higher prevalence of mental disorders in higher-income countries. Evidence from quasi-natural experiments and randomized experiments suggests a causal relationship with income having a direct effect on mental health. Evans and Garthwaite (2010) examine the effects of the US Earned Income Tax Credit (EITC) expansions on maternal mental health. Comparing low-income mothers with two children, who received substantially more in benefits, with mothers with one child, they find that mothers with two children had significantly fewer days with poor mental health over the previous 30 days. In Canada, Milligan and Stabile (2011) find strong evidence that additional income through child benefits has significant positive effects on maternal mental health and children's mental health. Although these studies suggest an effect, one cannot necessarily extrapolate these findings to other populations, including populations with higher incomes.

Case (2004) finds that members of South African households that include a pensioner are less depressed than members of other households, again suggesting a causal effect of income on mental health. Fernald *et al.* (2008) find that the effect of giving loans to previously rejected applicants reduced depressive symptoms in men, despite increased perceived stress.

Economic theory has also been applied to suicide, leading to the prediction that as with mental health in general, suicide will decrease with permanent income (Hammermesh and Soss, 1974) and population-level data on suicide rates are generally consistent with that prediction. Recent work suggests that relative income may affect suicide risk. For example, Daly *et al.* (2012) finds that, holding own income constant, a 10% increase in county income was associated with a 4.5% increase in suicide hazard, suggesting that lower social status increases suicide risk.

Macroeconomic Conditions

A consistent finding from the economic literature on macroeconomic effects on health is that physical health is countercyclical, which is commonly explained by individuals investing greater time in physical-health promoting activities when the opportunity costs of time are lower. Interestingly, the relationship between macroeconomic conditions and mental health appears to be procyclical. The research literature uses unemployment rates (measured at various level of aggregation) as a measure of local macroeconomic conditions. In periods of higher unemployment levels of suicides, various measures of mental distress, and other disabling mental disorders increase, even though measures of physical health (i.e., acute health conditions and disability) improve. For example, Ruhm (2003) finds that a 1% increase in unemployment was associated with a 7.3% decline in nonpsychotic mental disorders. Economists find that Google searches for mental health-related terms increase with unemployment rates, providing further, novel evidence of the procyclicality of mental health. Less is known about how and why macroeconomic conditions affect mental health. Three nonmutually exclusive hypotheses are that poor macroeconomic conditions reduce individuals' mental health by reducing income, by increasing rates of job loss, and by increasing overall levels of psychosocial stress.

Employment

The direct and indirect effects of unemployment on mental health have been compared with large and significant direct

effects found. Although the direct effects of unemployment are much greater than indirect effects, when measuring the consequences of unemployment at a population level, Helliwell and Huang (2011) find that the indirect effects dominate because a much larger population is affected. This work suggests the nonpecuniary effects of being unemployed on mental health are 5.5 times greater than the effects due to loss of income. They also find that higher unemployment benefits (as measured by the benefits replacement rate) do not mitigate the effect of unemployment on mental health. On the other hand, Salm (2009) examines exogenous involuntary job loss, as measured by business closures, and finds little evidence of an effect of job loss on the mental health outcomes studied.

Others have looked specifically at the effect of retirement on mental health outcomes. Dave *et al.* (2008) find significant negative effects of retirement on mental health. Again, they find that income is not the dominant mechanism by which retirement affects mental health. They explore the mechanisms for this effect and find evidence for the importance of declines in social interactions and physical activity.

Can public programs alleviate life events that affect mental health? Economic studies have failed to find significant effects of declining US welfare caseloads on maternal mental health; though there is some evidence that length of maternity leave may affect the severity of depression. Kling *et al.* (2007) examined the effect of being offered housing vouchers on health. They find being randomly assigned to receive a housing voucher did lead to lower poverty rates and residence in safer neighborhoods 4–7 years postrandom assignment. Although no significant effects were found on adult physical health outcomes, large positive effects on adult mental health outcomes were found, with a 45% reduction in relative risk of serious mental illness.

The existing literature on determinants of mental health finds that psychosocial stress negatively affects mental health, and that important psychosocial stressors can emerge from individual and societal economic conditions. Experimental research, including animal studies, identifies the causal effect of psychosocial stress on mental health in controlled laboratory settings. Observational research links individual and societal economic conditions to levels of stress, and correlates stress with mental health without identifying a causal effect of stress, per se. Economists have recently bridged the gap between these two broad areas of research by applying the tools of empirical microeconomics to the study of mental health outcomes. This economic research generally suggests that individual and societal economic conditions do affect mental health, in some cases substantially. Yet, few evaluations of policies to affect economic conditions consider these effects when evaluating the costs and benefits of programs. Generally, existing literature suggests including mental health as an outcome measure in these evaluations is likely to increase the benefits of these policies.

See also: Health Status in the Developing World, Determinants of. Intergenerational Effects on Health – *In Utero* and Early Life. Macroeconomy and Health. Noncommunicable Disease: The Case of Mental Health, Macroeconomic Effect of

References

Case, A. (2004). Does money protect health status? Evidence from South African pensions, NBER Chapters. In Wise, D. A. (ed.) *Perspectives on the economics of aging*, pp. 287–312. Chicago, IL: University of Chicago Press.

Daly, M., Wilson, D. J. and Johnson, N. J. (2012). Relative status and well-being: Evidence from US suicide deaths. *Review of Economics and Statistics*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1026351 (accessed 23.08.13).

Dave, D., Rashad, I. and Spasojevic, J. (2008). The effects of retirement on physical and mental health outcomes. *Southern Economic Journal* **75**(2), 497–523.

Evans W. N. and Garthwaite C. L. (2010). Giving mom a break: The impact of higher EITC payments on maternal health. *NBER Working Paper 16296*. Cambridge, MA: National Bureau of Economic Research.

Fernald, L., Hamad, R., Karlan, D., Ozer, E. J. and Zinman, J. (2008). Small individual loans and mental health: A randomized controlled trial among South African adults. *BMC Public Health* **8**(16), 409.

Grossman, M. (1972). On the concept of health capital and the demand for health. *The Journal of Political Economy* **80**(2), 223–255.

Hammermesh, D. S. and Soss, N. M. (1974). An economic theory of suicide. *Journal of Political Economy* **82**(1), 83–98.

Helliwell J. F.and Huang H. (2011). New measures of the costs of unemployment: *Evidence from the Subjective Well-Being of 2.3 Million Americans. NBER Working Paper 16829*. Cambridge, MA: National Bureau of Economic Research.

Kessler, R. C., Berglund, P., Demler, O., et al. (2005a). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry* **62**, 593–602.

Kling, J., Liebman, J. and Katz, L. (2007). Experimental analysis of neighborhood effects. *Econometrica* **75**(1), 83–119.

Malaspina, D., Corcoran, C., Kleinhaus, K. R., et al. (2008). Acute maternal stress in pregnancy and schizophrenia in offspring: A cohort prospective study. *BMC Psychiatry* **8**, 71.

Milligan, K. and Stabile, M. (2011). Do child tax benefits affect the wellbeing of children? Evidence from Canadian child benefit expansions. *American Economic Journal: Economic Policy* **3**(3), 175–205.

Ruhm, C. (2003). Good times make you sick. *Journal of Health Economics* **22**(4), 637–658.

Salm, M. (2009). Does job loss cause ill health? *Health Economics* **18**(9), 1075–1089.

US Department of Health and Human Services (1999). *Mental Health: A Report of the Surgeon General*. Rockville, MD: US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.

WHO World Mental Health Survey Consortium (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Journal of the American Medical Association* **291**(21), 2581–2590.

Further Reading

Chatterji, P. and Markowitz, S. (2005). Does the length of maternity leave affect maternal health? *Southern Economic Journal* **72**(1), 16–41.

Kaestner, R. and Tarlov, E. (2006). Changes in the welfare caseload and the health of low-educated mothers. *Journal of Policy Analysis and Management* **25**(3), 623–641.

Kerwin, D. and DeCicca, P. (2008). Local labour market fluctuations and health: Is there a connection and for whom? *Journal of Health Economics* **27**(6), 1332–1350.

Kessler, R. C., Chiu, W. T., Demler, O. and Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National comorbidity survey replication. *Archives of General Psychiatry* **62**(6), 617–627.

Ruhm, C. (2000). Are recessions good for your health? *Quarterly Journal of Economics* **115**(2), 617–650.

World Health Organization (2008). *The global burden of disease*. Geneva: WHO Press. 2004 Update.

Mergers and Alliances in the Biopharmaceuticals Industry

H Grabowski, Duke University, Durham, NC, USA
M Kyle, Toulouse School of Economics, Toulouse, France, and Center for Economic Policy Research, Toulouse, France

Introduction

Over the past few decades, the pharmaceutical industry has been characterized by both significant consolidation of large pharma firms as well as the vertical disintegration of the R&D process. The latter is associated with considerable entry into the discovery and development process by early-stage biopharmaceutical firms. Since the early 1990s, an evolving marketplace for new technologies through licensing agreements and joint ventures has emerged, accompanied by the growth of contract research organizations that specialize in implementing clinical trials for new drug candidates.

To put some of these changes in historical perspective, it is useful to chronicle some of the key dynamic forces affecting the pharmaceutical industry. The prevailing company structure that dominated the industry from the end of World War II through the 1980s was the large, vertically integrated multinational firm with R&D laboratories, production facilities, and marketing departments. These firms generally financed their R&D investment through internally generated funds, emphasized growth through company-developed pipelines (Grabowski, 2012; Scherer, 2010). Although entry from new start-ups began in earnest in the 1970s, most drug products associated with new technologies such as recombinant technology were in the early stages of development and years away from reaching the market.

Although the 1980s was a period of rising prices and profits for the industry, integrated multinational drug firms also faced many challenging developments. These changes included rising R&D costs (DiMasi *et al.*, 1991, 2003), the expiration of patents on major commercial products, and the beginning of intensive price competition from generics. The passage of the Waxman-Hatch Act in 1984 was a key legislative change that substantially reduced entry costs for generic firms. Generics may now enter the market by demonstrating only bioequivalence, and can rely on the clinical data provided by the originator to show safety and efficacy (Grabowski, 2007). These dynamic forces intensified in the 1990s with the rise of buyer-side market power in the form of managed care organizations and pharmacy benefit managers in the USA, and increasingly stringent price controls in other major world markets. Market and political pressures have caused declining growth in sales and profits that have been particularly evident on an industry-wide basis since the mid-1990s.

There has been increasing attention over recent years to whether the pharmaceutical industry is now in an R&D productivity crisis. Several observers have pointed to a pattern of rising R&D expenditures accompanied by a declining trend in new molecular entities since the mid-1990s. The productivity crisis idea is subject to various qualifications relating to the quality of new molecular entities, the long lags that characterize the R&D process in pharmaceuticals, and a gradual shift to a new R&D paradigm based more on biology than chemistry (Cockburn, 2006). Nevertheless, the declining trend in new products from the R&D labs, along with continuing patent expirations on prior 'blockbuster' products, has created a replacement problem for many large pharma firms.

Large-Scale Mergers

The structural response to these dynamic forces has included both large horizontal mergers as well as a growing number of development-stage agreements between large pharma firms and smaller, research-based biopharmaceutical firms. The first merger wave began in 1989–90. The annual value of pharmaceutical mergers in these two years exceeded that of any prior year in the 1980s by a considerable margin (Ravenscroft and Long, 2000). This was followed by an even larger merger wave beginning in the mid-1990s and continuing into the 2000s (Danzon *et al.*, 2007; Koenig and Mezick, 2004). After a lull of several years, two large-scale merger deals were consummated in 2009 (Pfizer–Wyeth and Merck–Schering). Combinations have included not only mergers between large pharma firms, but also the acquisitions of biotech firms by pharma firms and mergers between firms of different sizes in the emerging biotech sector.

Table 1 shows how the global market shares in the pharmaceutical industry have changed between 1989 and 2009. It shows the global market shares of the top 18 ranked

Table 1 Global shares and mergers in pharmaceuticals, 1989–2009

Rank	*2009*		*1989*	
	Company	*Share (%)*	*Company*	*Share (%)*
1	Pfizer	7.6	Merck	4.0
2	Merck	5.2	BMS	3.5
3	Novartis	5.1	Glaxo	3.1
4	Sanofi-Aventis	4.7	SKB	3.0
5	GlaxoSmithKline	4.7	Ciba-Geigy	2.9
6	AstraZeneca	4.6	AHP	2.7
7	Roche	4.4	Hoechst	2.4
8	J&J	3.6	J&J	2.3
9	Lilly	2.7	Bayer	2.3
10	Abbott	2.6	Sandoz	2.1
11	Teva	2.1	Lilly	2.1
12	Bayer	2.1	Pfizer	2.0
13	Boehinger Ing	2.0	Roche	1.9
14	Amgen	2.0	Schering-Plough	1.6
15	Takeda	1.9	MMD	1.6
16	BMS	1.9	Upjohn	1.5
17	Daiichi Sankyo	1.2	Boehringer Ingel	1.5
18	Novo Nordisk	1.1	Warner Lambert	1.4

Source: Authors' analysis based on IMS Health Care Market Share Data.

firms by sales in 1989 and 2009. All of the eight top-ranked firms in 2009 engaged in large-scale mergers as well as many small-scale acquisitions over this 20-year period. Correspondingly, several of the leading firms in 1989 have been consolidated into larger entities. The firms with asterisks beside their names in 1989 have all been acquired by, or merged into, larger surviving entities.

The top ranking firm in 2009, Pfizer, completed three major mergers over this period (involving Warner Lambert, Pharmacia-Upjohn, and Wyeth). Other top-ranked firms consummating large-scale mergers over this period include GlaxoSmithKline (merging Glaxo, Burroughs Wellcome, and SmithKline-Beecham), Sanofi-Aventis (merging Hoechst, Marion Merrell Dow, Aventis, Rhone-Polenc, and Sanofi-Synthelabs), Novartis (merging Ciba-Geigy and Sandoz), and Roche (merging Syntex and Genentech).

Some of the prominent mergers in the earlier merger waves involved cross-border mergers (e.g., Pharmacia-Upjohn, AstraZeneca, and SmithKline-Beecham). Cross-border mergers exhibited positive gains in market valuation in the event studies conducted by Ravenscroft and Long (2000). International integration of pharmaceutical activities provide benefits by allowing firms to obtain faster global uptake of new drugs in the growth phase of the pharmaceutical life cycle, and coordinating firm strategies on a multinational basis. At the same time, integrating disparate corporate cultures and headquarters can lead to substantial challenges and implementation costs (Belcher and Nail, 2000). Nevertheless, mergers have been an important factor contributing to a multinational business structure and strategic approach in pharmaceuticals. Many of the firms that were regional in nature in 1989 are now part of larger enterprises with global R&D, manufacturing, and marketing capabilities.

The share of global sales of the top 18 firms increased to 59% by 2009, compared with 41% in 1989. Despite the merger activity and increased concentration, the pharmaceutical industry is still relatively unconcentrated compared with many other industry sectors. Antitrust authorities have rarely raised concerns about the potential for increased market power, only occasionally requiring the divestiture of some product lines. Changes in company rankings also occur over time as a result of both new product introductions and patent expirations. This is reflected by the rapid growth of dedicated biotech firms like Amgen and Genentech (now part of Roche). In addition, firms whose dominant business is generics, such as Teva, have grown to a sizeable presence, reflecting the increased utilization of generic drug products over these two decades as well as their entry into specialty branded products.

Figure 1 depicts the number of mergers or acquisitions and the average deal size over the period 1990–2011. This chart is based on data from Recombinant Capital. Since 1990, there has been a dramatic upward trend in the number of reported mergers in the pharmaceutical industry, consistent with the consolidation observed in Table 1. However, there is a decline in the number of deals since 2007, reflecting in part the downturn of overall global economic activity. The average deal size line shows considerable volatility with peaks around periods of large-scale horizontal merger activity. At the same time, however, there has been a downward trend in average deal size since 2004, reflecting an apparent move away from 'mega-mergers' and toward more acquisitions of development-stage firms by pharma firms and mergers between smaller biotech firms.

Alliances

The pharmaceutical industry has made extensive use of 'markets for technology' since the late 1970s. The term refers to the licensing of ideas or technology platforms, R&D alliances, or joint ventures between firms. Markets for technology allow a vertical disintegration of the product development process, with some firms specializing in early-stage work and others in the execution of clinical trials, preparation of regulatory dossiers, and marketing. These relationships, which is generally referred to as alliances, can be a substitute for mergers and

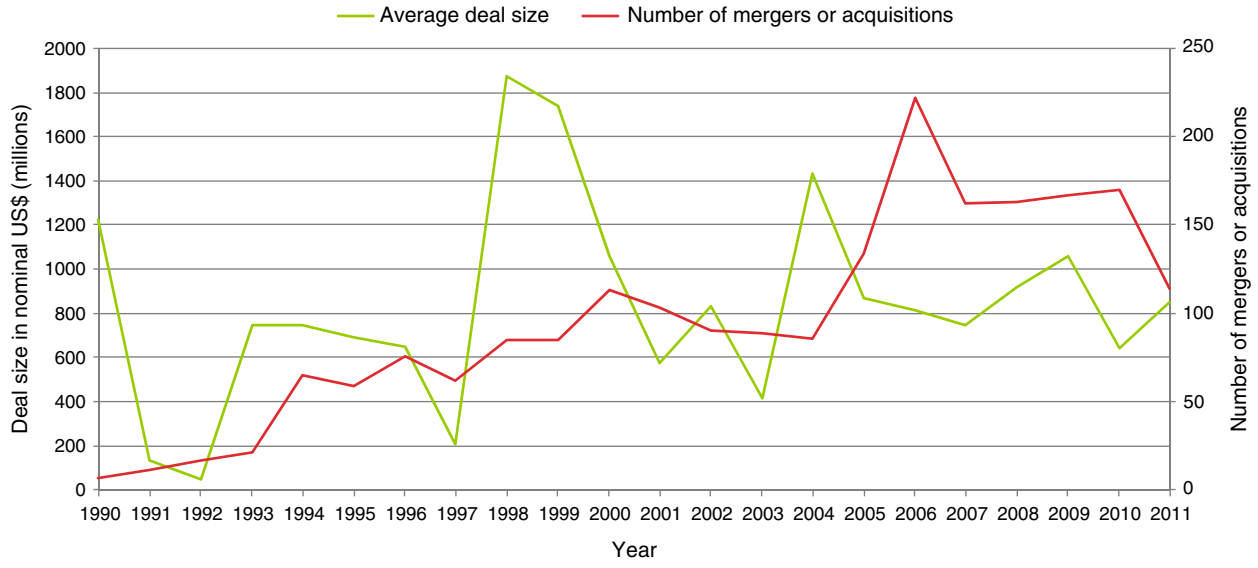


Figure 1 Merger activity and size.

acquisitions: that is, they may be a more efficient alternative to the vertical integration resulting from a merger or acquisition. However, they can also be a form of extended due-diligence during which the firms learn about whether a future merger or acquisition between partners would be desirable. A prototypical development-stage agreement would involve payments of milestones and/or royalties and some sharing of R&D expenses in the exchange for rights to develop and/or market the new products covered under the agreement. The extent of integration at the R&D stage associated with these agreements varies considerably, ranging from true joint development agreements, to transfers of development-stage products from licensors to licensees, to marketing options in exchange for development-stage funding and future payments.

Although large, horizontal-style mergers have tended to occur in waves, there has been a steady upward trend in the number and values of R&D stage alliances in the pharmaceutical industry between the mid-1990s and the onset of the global recession in 2008. During this period, the number of such alliances with market valuations of $100 million or more increased several fold in value (Recombinant Capital, 2008). With the onset of the global recession in 2007–08, however, the number of collaborative R&D deals has moderated, with a downward trend in average deal value. This downward trend reversed in 2011, perhaps signaling a return to growth in the annual number of deals and values (Cartwright, 2012).

For a subset of licensing deals, Recombinant Capital has access to the contracts themselves. Because this information is reported for only a subset of deals, it may not be representative. **Figure 2** presents a summary of how contract terms have evolved over the past two decades. There has been a pronounced increase in milestone payments relative to other deal terms, such as upfront fees and equity positions. The preference for milestone payments over upfront fees likely reflects 'lemons problem' concern; partners prefer to see results before they pay. However, the ability of R&D firms to accept these terms indicates that their financial position is more secure and they are less desperate for cash from partners.

Figure 3 shows a shift toward licensing later in the development process. Because there is a much larger set of projects available for license at early stages (the high failure rates significantly reduce the number of projects that survive until Phase II or III), most deals are in the development or preclinical stage. But by 2011, almost half of the deals were Phase I or later, and the share of Phase III or later was more than 20%. Again, this trend is consistent with both the desire to avoid financing a lemon as well as reduced financial constraints for R&D firms, who are capable of financing development through more stages than in earlier years.

The geographic coverage of licenses has been more limited recently, as is shown in **Figure 4**. This figure corresponds to the reach of licensor rights as assigned in the agreement. The 'worldwide' category includes licenses that have no geographic restriction; the remaining categories refer to the specific geographic region for which the licensor has rights. The trend toward more restricted geographic coverage is somewhat surprising, given the prevalence of multinationals with a presence in most markets and the global nature of research. It may reflect a move toward licensing arrangements whose primary focus is marketing, rather than R&D. However, it points to the continued salience of country-specific knowledge and capabilities within firms.

In the next section of the article, the motives for pharmaceutical mergers and related empirical studies on the determinants of M&A activity are considered. Section 'The Effects of Pharmaceutical Mergers and Alliances' summarizes evidence concerning the effects of mergers and alliances on different outcome measures including R&D productivity and innovation. Section 'Policy Issues' discusses current policy issues involving mergers and alliances. The final section

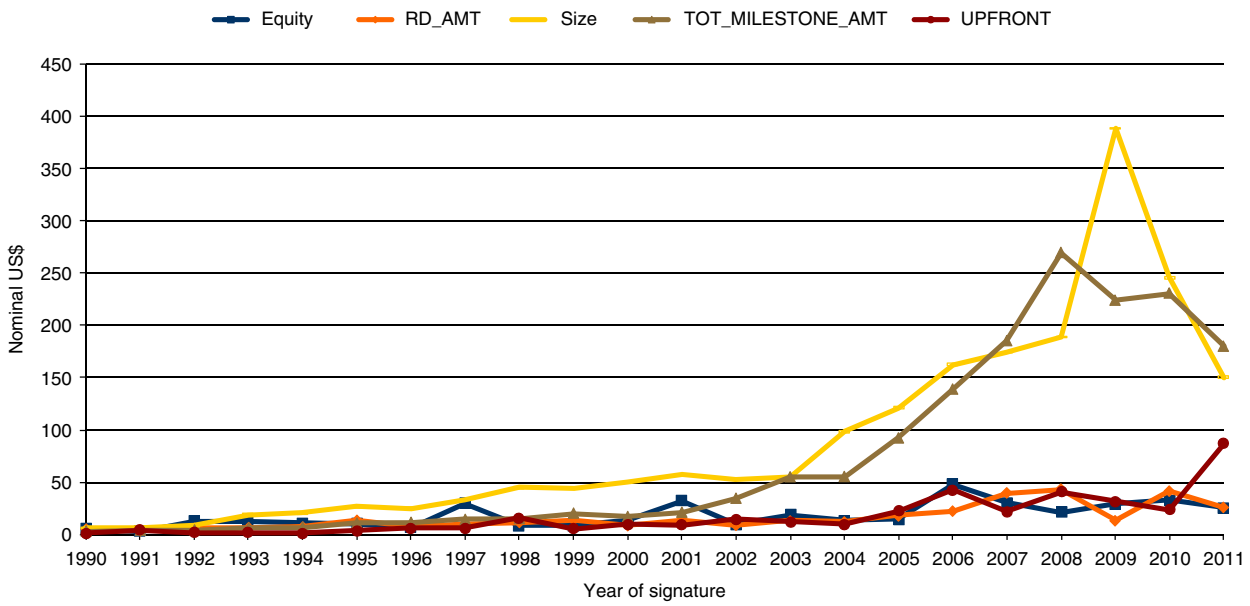


Figure 2 Average deal terms of licenses.

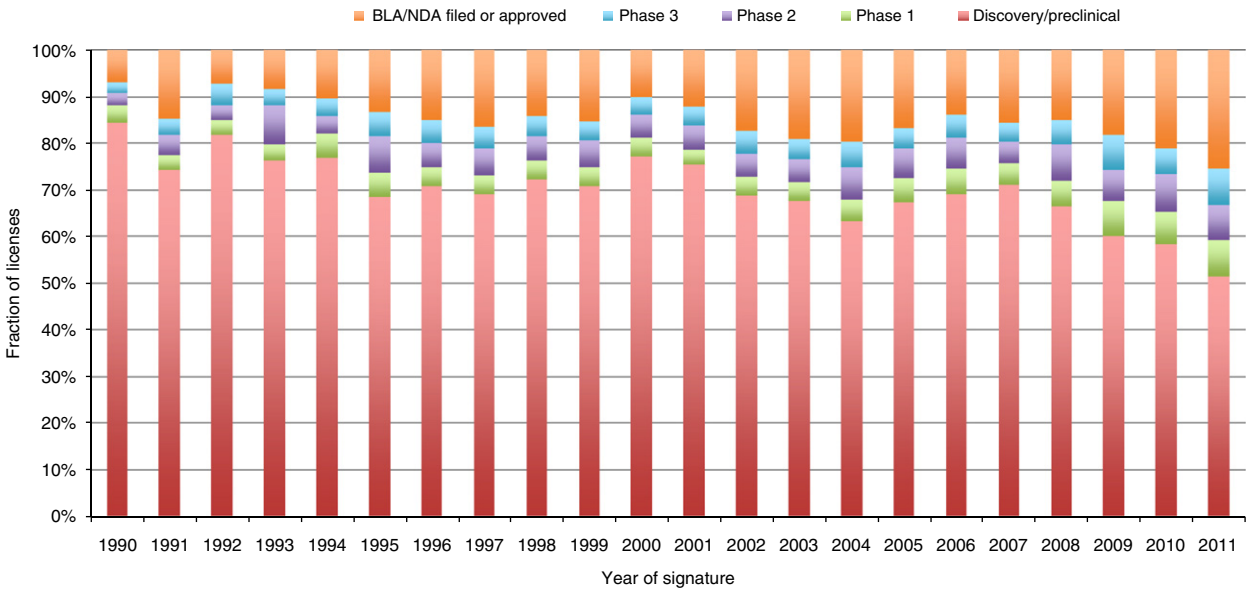


Figure 3 Licenses by stage of development.

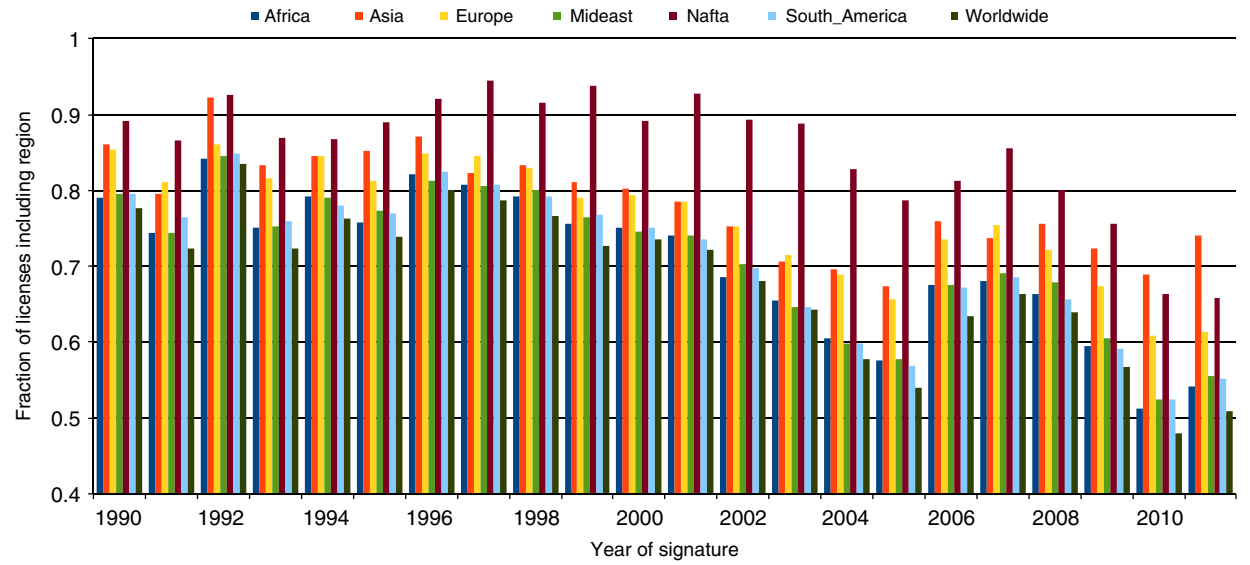


Figure 4 Regional coverage of licenses.

provides some concluding observations and interesting questions for future research.

Determinants of Mergers, Acquisitions, and Alliances

The motives for merger and acquisitions (M&As) activity can be broadly categorized into adaptive or defensive rationales versus proactive or offensive ones (Burns *et al.*, 2005). In this section, this classification is utilized to consider the economic drivers of M&As in the pharmaceutical industry. It is important to understand the rationales for mergers before attempting to evaluate studies that are focused on the effects of mergers. The role of alliances as a substitute or a complement to M&A activity is also considered.

Defensive Motives: Strategic Response to Environmental Change

The hypothesis that industry-wide shocks can precipitate merger waves appears to be a useful concept in understanding pharmaceutical mergers. The original hypothesis goes back to Gort (1969). Industry-wide shocks appear to explain merger waves in other industries like banking and telecommunications in the 1990s (Andrade *et al.*, 2001). In the case of

pharmaceuticals, the economic environment became more difficult and pipeline gaps emerged throughout the industry by the late 1980s. With stock prices under pressure, many affected drug firms were motivated to use their accumulated cash flows to acquire another firm's products and pipeline. The bidder often could pay the premium associated with these acquisitions by consolidating operations and cutting out the excessive infrastructure capacity. Various researchers have made the point that mergers and acquisitions facilitate disruptive organizational change that would otherwise meet with substantial internal inertia and resistance (Ravenscroft and Long, 2000). However, mergers are also associated with substantial integration costs that can affect the productivity of the firm in the postmerger period (Larson and Finkelstein, 1999).

Ravenscroft and Long (2000) performed one of the first analyses of pharmaceutical mergers. Their analysis covered mergers of significant value undertaken between 1985 and 1996. Using event study methodology, Ravenscraft and Long found that the large horizontal mergers and cross-border mergers created gains in overall stock market value. As in other industry studies, however, target firms captured most of the returns. Their findings are consistent with the response to the industry shocks, excessive capacity hypothesis. Their analysis of cost-cutting for large horizontal pharmaceutical mergers found a reduction in total headcount in the postmerger period ranging from 8% to 20% of the combined workforce in the premerger period. Although cost-cutting in manufacturing and marketing personnel were proportionately greater than for R&D employees, there also was a consolidation of R&D laboratories and the elimination of marginal R&D projects by several firms.

A subsequent analysis by Centerwatch of 11 large mergers (22 pharmaceutical companies) that occurred between 1989 and 1998 reported a 34% average reduction in development projects three years after the merger was consummated (Centerwatch, 2000). Neither the Centerwatch study nor Ravenscraft and Long's analysis, however, examined subsequent effects on the firms' R&D productivity or the probability of success. To the extent that these reductions in R&D activities eliminated duplicate efforts or projects with low probability of success, or facilitated more external alliances, the companies' R&D performance could have increased in the postmerger period compared with premerger one. This issue is considered further below.

Other researchers also find evidence that firms under economic stress are more likely to engage in mergers. An oft-cited firm-specific motivation for pharmaceutical M&As is to fill in gaps in a company's pipeline to maintain growth in the face of a major product's patent expirations. Patent expirations on major projects can produce rapid losses in unit sales to generic entrants and leave firms with substantial excess capacity in their marketing and sales forces. Pharmaceutical products exhibit a highly skewed distribution of revenues and returns (Grabowski *et al.*, 2002).

Two published studies have investigated this hypothesis and found that pipeline gaps and issues continue to be a key driver of merger activity. A study of 202 biotechnology and pharmaceutical mergers between 1998 and 2001 found that pharmaceutical firms that have a relatively old portfolio of marketed drugs exhibit a higher propensity to acquire another firm (Danzon *et al.*, 2007). A second study of 160 pharmaceutical mergers between 1994 and 2001 found that firms with lower scores in the strength of their R&D pipeline and fewer years of exclusivity on their marketed drugs had a greater probability of engaging in a merger (Higgins and Rodriguez, 2006).

The fact that firms in economic stress are more likely to engage in mergers creates methodological issues in evaluating pharmaceutical merger activity. In particular, one cannot simply compare merging entities to overall industry performance. Rather, it is important to construct control groups with similar firm characteristics in evaluating the effects of a merger. This issue is considered further below.

Economies of Scale and Scope

Proactive motives for mergers include increases in size to achieve critical mass and economies of scale in R&D and other firm activities. A series of papers by Cockburn and Henderson (1996, 2001) focuses on economies of scale and scope in drug R&D provides some insight on the effects of increased size on R&D productivity. They looked for the effect of scale and scope on productivity at a research program level, for 10 large firms. The advantage to these papers is that they use extremely detailed data (including program-level R&D spending) over a very long time period. They conclude that firms engaged in a broader scope of research activities are more productive than focused firms, but that scale does not matter much once the scope is controlled for. A more recent study by Danzon *et al.* (2005) finds benefits from a company's development experience as measured by the number of drugs in clinical trials, but these benefits are also subject to diminishing returns. In particular, this study finds the maximum performance measured in terms of success probabilities at different stages of the clinical development process occurs at 25 drugs in development. This is far below the number of drugs in development for the major firms listed in **Table 1**. This study is discussed further in terms of the effects of alliances on R&D productivity.

As a result of the consolidation that has occurred over the past few decades, some of the leading pharmaceutical firms now have annual R&D budgets of over several billion dollars to manage. At this size, companies may have entered a region of diminishing returns from the standpoint of managing and motivating creative individuals and coordinating their activities. It is notable that Pfizer, GlaxoSmithKline, and other firms with multibillion dollar R&D programs and several hundred R&D projects are instituting more flexible organizational structures, and delegating more decision-making authority to the heads of the various therapeutic areas (Dorey, 2001; Mathieu, 2007).

Mergers may also reflect management goals to maintain or increase firm size, even if this is not associated with economies of scale or improved long-term R&D productivity (Mueller, 1986). Although layoffs and consolidation typically occur in the aftermath of large-scale pharmaceutical M&A activities, the acquiring firm can also draw on an expanded portfolio of products and pipeline candidates to mitigate downsizing in the wake of imminent patent expirations on its major products. Managerial utility has been related to firm size in

several economic studies (Marris, 1998). This may help to explain the repeated serial use of large-scale mergers by many of the major ranked pharmaceutical firms in **Table 1**, mergers that appear to offer mainly short-run cost savings as discussed further below.

Access to New Technologies and Therapeutic Areas

Beyond economies of scale, biopharmaceutical firms may engage in mergers to gain a presence in an emerging therapeutic category that represents significant future growth opportunities. For example, the oncology class has been characterized by several new 'first-in-class' drugs in recent years (DiMasi and Grabowski, 2007). The oncology class is now the fastest growing therapeutic category among all major drug classes. The novel entities in this class have emerged primarily from the biotech sector, utilizing molecular biology techniques (e.g., new monoclonal antibody products and other targeted agents). Mergers provide a more expeditious way to enter such high opportunity fields relative to internal expansion. It can take several years or even decades to build the internal scientific capability to enter a new therapeutic area or implement a new research platform in an emerging scientific field. This appears to be an important motivation underlying both acquisitions and alliances of developing biotechnology firms by established pharmaceutical firms.

Alliances as Substitutes or Complements to Mergers

As discussed in Section 'Introduction,' larger firms are also increasingly looking to alliances and partnerships with smaller biotechnology firms as the source of new products. This suggests that scale requirements, at least in the discovery and early stages of the development process, remain modest. At these earlier stages, small research-oriented boutique firms may enjoy a number of advantages relative to their larger rivals. These include the fact that they are closer to cutting-edge technology emerging from universities and public supported basic research, are more willing to take risks on disruptive technologies, and are less bureaucratic in organizational structure (Scherer, 1999). By contrast, larger pharmaceutical and biotechnology firms may have advantages in the more advanced stages of development, where large-scale clinical trial design and regulatory coordination become important. This rationale for R&D specialization based on different comparative advantages in research versus development was advanced by Arrow (1983) in a more general model of the R&D process.

Although alliances and partnerships are an alternative to mergers as a means to acquire new technological platforms and R&D pipeline candidates, they also pose their own set of issues. Arora *et al.* (2001) find support for gains from a division of labor at alternative stages of the R&D process. There also are positive network effects associated with alliances and partnerships (Pammolli and Roccaboni, 2004; Powell *et al.*, 1996). However, partnership deals may be susceptible to a 'lemons' problem arising from agency and information problems (Akerlof, 1970; Pisano, 1997). Partnerships also raise challenging bargaining, management, and governance issues (Teece, 1998; Arora *et al.*, 2001).

Many M&As in the pharmaceutical area have occurred between firms that had first engaged in some type of alliance or partnership (Higgins and Rodriguez, 2006). This may help merging firms overcome pitfalls associated with agency problems and information asymmetries. In particular, the information gathered over time from an alliance may allow the acquiring firm to better assess the value of the acquired firm's intangible capital. It may also provide information on the resulting organization's ability to successfully integrate the strengths of the two companies. The difficulty of integrating firms with different cultures and organizational structures is an oft-cited reason for failures in mergers in the management literature (Larson and Finkelstein, 1999; Smith and Quella, 1995; Haspeslagh and Jemison, 1991).

Increasing Market Share

A traditional economic motive for mergers, of course, can be to increase market share and market power to gain competitive advantage. This has not been a major issue in the case of the large pharmaceutical mergers depicted in **Table 1**. In the USA and Europe, mergers are subject to scrutiny before their implementation by the antitrust authorities (Mueller, 1996). There are guidelines on what economic parameters can trigger challenges. In the case of pharmaceuticals, markets are defined in terms of therapeutic categories, because a drug product to alleviate pain, for example, does not compete with one that is approved for hypertension. Horizontal mergers of significant consequence, therefore, must go through a vetting process before implementation. These negotiations can result in a settlement where competitive products in the same therapeutic category are spun-off as a condition for allowing the merger.

Nevertheless, one of the distinctive areas of antitrust concerns for R&D-intensive pharmaceutical firms is in the area of innovation markets. In particular, this issue arises when two merging parties have potentially competing drug candidates in their R&D pipelines. The concern is that this merger could result in the combined firms suppressing one of the research paths in order to avoid cannibalizing the economic performance with the candidate that is carried forward. Since the 1990s, there have been several challenges of mergers in pharmaceuticals based on innovation markets. This issue is discussed in terms of policy issues considered in Section 'Summary and Concluding Comments.'

The Effects of Pharmaceutical Mergers and Alliances

Beyond the event studies discussed above, there have been a number of studies that have examined the specific effects of mergers or alliances on profits, R&D activity and other performance measures. The results are mixed in nature, and raise a number of issues and questions for further research.

Large Market Value Mergers and Acquisitions

Danzon *et al.* (2007) look directly at the effect of mergers in pharma/biotech on various measures of performance. They focus on mergers with $500 million or more of market value.

They find that these mergers are frequently a response to distress, so it is important to compare outcomes for merging firms to outcomes for other firms with similar characteristics, and they create propensity scores for this purpose. They conclude that mergers result in slower growth and a reduction in operating profit, though these effects are rather small. They also find smaller R&D growth for small merging firms. They look at performance in the first three years following a merger.

Ornaghi (2006) also looks at postmerger performance in the industry, and focuses on R&D productivity. He does not use the propensity score or economic distress index for developing a 'control group' with which to compare the performance of merging firms. He finds that in the three years following a merger, there is a decline in R&D spending as well as productivity, as measured by patents. Koenig and Mezick (2004) focus on a relatively small number of high-profile mergers consummated between 1989 and 1996 (a sample of seven large mergers). Comparing the performance of companies in the industry that undertook these mergers with a control group of firms that did not, they find that companies that merged were able to achieve more favorable postmerger productivity scores.

To summarize, although prior studies by Henderson and Cockburn (1996, 2001) and Danzon *et al.* (2005) generally find some advantage to R&D scale and scope, the studies analyzing merger effects find a weakly negative effect on R&D performance and related measures. This may be because there is no additional advantage to size at the level for most large market value mergers except perhaps for short-run cost savings reflected in the event studies discussed above. Alternatively, because many mergers are a response to distress, the counterfactual is hard to determine.

Existing studies leave open many questions for further research. First, none of them look at the long-run impact of mergers on a firm's productivity. Three years is unlikely to be enough time to pick up many changes in patenting activity, much less progression through the phases of development. Second, as Cockburn notes, patents and new chemical entities are not necessarily the best measure of output, though almost all the evidence on R&D productivity centers on these two measures. Third, the focus is on the larger mergers between public firms, and there is little attention paid to heterogeneity in outcomes. The management and finance literatures are concerned with what drives a successful merger, such as whether the R&D activities of merging firms are substitutes or complements, similarities in culture or corporate structure, the integration process, and other economic and organizational characteristics (Hitt *et al.*, 2001). These issues remain important questions for future research.

Studies of Mergers and Acquisitions Involving Development-Stage Firms and Partners

It is useful to distinguish the outcomes for M&As involving development-stage firms from those involving large-scale mergers between fully integrated pharmaceutical firms. Higgins and Rodriguez (2006) examined a sample of 160 research and development acquisitions over the period 1994–2004. Most of these deals involved an established pharmaceutical firm purchasing a development-stage company whose main assets involved new product candidates and R&D platforms. As discussed, Higgins and Rodriguez find support for the industry shocks motivation for these acquisitions of development-stage companies. In particular, they find that these acquisitions operate to effectively complement a firm's internal R&D efforts in the sense that acquirers either maintain or improve their product pipelines postacquisitions. They also find that acquirers also experience significant postannouncement, positive, abnormal returns in their market value.

Another significant finding by Higgins and Rodriguez is that firms that were engaged in an alliance before an acquisition exhibited greater success in terms of pipeline scores in the postacquisition period and abnormal market returns in the postannouncement period than did firms with no prior alliance history. This is preliminary evidence that alliances can help acquiring firms overcome agency problems (avoiding a 'lemons problem'). The information obtained over time can also improve the acquiring firm's ability to integrate the partner's assets in the postacquisition. As discussed, the difficulty of integrating firms with different cultures and organizational structures is a frequently reported reason for merger failures by business researchers.

An analysis of M&As that focuses on R&D outcomes is also undertaken (Grabowski and Kyle, 2007, 2012). In contrast to other merger studies, our analysis focuses on the effects at the R&D project level of observation. R&D outcomes are measured in terms of advancement through the various phases of drug research and market launch. It utilizes a large database of more than 4500 firms engaged in pharmaceutical R&D between 1990 and 2007. Our sample therefore includes a large proportion of development-stage companies.

Because most of the 4500 plus firms in this data set are not publicly traded in the USA, consistent time-series financial data such as R&D spending, total asset size, and other important control variables used in most other studies of mergers in this industry is lacking. Those studies focus on the performance of large firms. Our data has the advantage of including firms of varying size, at the cost of poorer information on financial data for nonpublic firms. To measure firm size, the count of active drug development projects each year is used, and four size categories are created: small (fewer than five projects underway in a year); medium (5–20 projects); large (20–50 projects); and very large (more than 50 projects).

It is found that a higher fraction of projects of firms that experienced a merger during the 1985–2006 period progress to the next phase. The differences in advancement rates are greater for the smaller merged firms at each research stage. However, the most substantial difference occurs in projects advancing from Phase III to market that originated in a firm with less than five research projects that was merged into or acquired by another company. Our general findings were confirmed in a logit regression analysis. The higher probability of market success for smaller merged firms compared with nonmerged ones is consistent with the comparative advantage of larger firms at later stages of the R&D process hypothesis (Arora *et al.*, 2001) or alternatively, with the hypothesis that large firms are better at weeding out unlikely successes earlier

in the process (Guedj and Scharfstein, 2004). Further research on this issue is warranted.

An important question for further research is the source of the observed benefits from development-stage company M&As. Our results leave open the question of whether high-performing firms are more likely to merge, or whether mergers lead to higher performance. If the latter do, mergers combine complementary skills of two firms, leading to better project selection and advancement? Are mergers necessary for the realization of these benefits, or could strategic alliances be used instead? Are firms that enter into alliances before mergers more likely to have higher probabilities of success in their R&D project as suggested by the work of Higgins and Rodriguez? These are among the open issues that are useful topics for further research.

Alliances

The most extensive published study evaluating the performance of alliances has been done by Danzon *et al.* (2005). They examine the productivity at each phase of drug development (i.e., success probabilities) for 900 firms over the period 1988–2000. They focus on experience, measured by the number of drugs a firm has in development, rather than sales in looking at economies of scale. They find that the effect of experience on productivity (advancing a drug through a phase) is positive with diminishing returns for Phases II and III, with the maximum occurring at 25 drugs in development. Products developed in an alliance tend to have a higher probability of success, at least for Phases II and III trials, and especially when the licensee is a large firm.

Arora *et al.* (2007) also examine the role of licensing and alliances in a working paper using data from 3000 R&D projects in preclinical and clinical trials in the USA in the 1980s and 1990s. After controlling for selection effects, they find licensing improves the probability of success when the licensee is a pharmaceutical firm. Their results are therefore generally consistent with the results of Danzon *et al.* (2005) on the positive benefits of alliances. Both studies are inconsistent with a 'lemons' hypothesis by Pisano (1997), at least for the typical development oriented licensing arrangement between biotech and advanced pharmaceutical firms.

Lerner and Merges (1998) use the Recombinant Capital database on licensing contracts to test the predictions of the theoretical literature on the allocation of property rights between licensor and licensee. Consistent with most theoretical predictions, they find that the R&D firm is less willing to cede control rights when it has greater financial resources. In contrast, they find little support for the theoretical prediction that the R&D firm will maintain control rights early in the development process, when its marginal contribution is higher. More recently, Lerner and Malmendier (2010) examine how contract terms are used to manage the risk that the R&D firm uses financing from the alliance to subsidize its other drug development projects. They find that in alliances where research is noncontractible, termination options that are expensive to exercise are more commonly used.

Allain *et al.* (2012) examine the relationship between market structure and the timing of pharmaceutical licensing. Alliances should result in efficiency gains if firms have comparative advantages and different stages of the R&D processes, but these efficiency gains depend on the transfer of technology occurring at the stage when the licensee has a comparative advantage. If biotechnology firms, or licensors, have private information about the quality of their drug candidates, this introduces a friction in the market for technology. Their theoretical model demonstrates that because of the potential for a lemons market, pharma firms offer prices for licenses based on the expected quality of an innovation, and these prices are too low for good innovations. Even though they are less efficient, biotech firms will elect to perform clinical trials themselves in order to prove their quality and command a higher price for a license later on. The effect of competition in this market is two-fold. Competition between potential licensees increases the bargaining power of the licensor, which increases the price the licensor expects to receive by conducting its own trials and waiting to license. But intense downstream competition between these potential licensees also erodes the profits that can be derived from the innovation and thus the price that licensees are willing to pay. In this case, the biotech firm gains less from waiting, and an increase in competition leads to more efficient licensing. Empirical analysis of data on the stage of development at which licenses were signed during 1990–2006 shows evidence of both effects of competition.

In contrast to the work on large-value mergers, the studies of alliances find positive effects on R&D performance. These studies indicate that development experience is generally associated with higher success probabilities, especially in later R&D stages. Hence, there appears to be a potentially important role for specialization across R&D stages. These findings are also consistent with the R&D productivity gains observed for development-stage firms merged into larger enterprises with more developmental experience.

These leading studies on the effects of alliances and innovation, however, also raise many issues for further research. The business alliance literature suggests a rich array of contractual terms and an evolving landscape of ventures. In this regard, Danzon *et al.* (2005) do not explicitly consider the contractual terms, the extent of integration of the R&D process, or the characteristics of the firms involved in the agreement beyond a few simple attributes relating to a firm's size and experience in performing clinical trials. Arora *et al.* (2007) adjust for product selection effects, but their analysis only considers a few characteristic variables. Both studies raise a number of issues about the underlying drivers of successful alliances for further research analysis.

Policy Issues

Innovation Markets and Antitrust Considerations

In evaluating mergers in research-intensive industries, antitrust agencies have been concerned that if two companies have potentially competing products in their R&D pipelines, a merger might increase the incentive to suppress at least one of the R&D paths. The idea that antitrust authorities should concentrate on the dynamic effects of mergers on R&D activities or innovation markets was first advanced in a paper in the

economics literature by Gilbert and Sunshine (1995). A number of merger challenges by the Federal Trade Commission (FTC) have been initiated around this innovation market concept, and the pharmaceutical industry has been a particular area of focus (Carrier, 2008).

Although the innovation market concept in antitrust enforcement has its supporters, its applications have been criticized by many economists and lawyers (Carlton, 1995; Rapp, 1995; Carrier, 2008). For pharmaceuticals, with their long and uncertain development process, critics argue that antitrust authorities should focus on drug candidates in the late-stage of development where potential competition is more easily assessed. Early-stage development activities involve relatively low costs and barriers to entry, and are also subject to high levels of uncertainty. Many firms take a portfolio approach to obtaining new product introductions at the early stages of R&D. There are also typically many parallel R&D efforts across firms searching for promising new therapeutic approaches (DiMasi and Pacquette, 2004). When a drug progresses to the final Phase III of clinical testing, however, the probability of success increases to approximately 70%, whereas costs of clinical trials also increase significantly. Antitrust concerns about potential anticompetitive effects then arguably become more relevant, particularly when there are a small number of late-stage competitors in a therapeutic class.

Carrier (2008) has analyzed nine challenges of pharmaceutical mergers brought by the FTC involving the issue of overlapping R&D activities. For example, in the first of these cases, involving the Roche–Genentech merger in 1990, Genentech had a CD4 drug candidate for HIV-AIDS in Phase I trials and Roche had a preclinical R&D program in the same class. To consummate the merger, Roche was required to offer a nonexclusive license to its CD4 drug candidate. However, none of the CD4 drugs were ever approved. Carrier (2008) finds that some of the challenges by the antitrust authorities are warranted, but others are more problematic, given the relevant characteristics of the market and an analysis of the potential costs and benefits. In the questionable cases, he argues that the FTC has attempted to protect innovation where future outcomes are uncertain and many years away from the market. By contrast, the EU has taken a less stringent approach to some of these innovation market cases (Morgan, 2001). This is clearly an evolving area of antitrust policy that warrants more research and attention by scholars.

Two former Directors of the Bureau of Economics of the FTC, William Comanor and F. M. Scherer, in a memorandum to the FTC, discuss some potential adverse consequences for alliance formation that antitrust officials should address when evaluating large-value mergers, even if there are no significant overlaps in the particular therapeutic areas of each merging firm (Comanor and Scherer, 2009). First, they point out that these large-scale mergers reduce the cohort of independent firms that are available to partner with or acquire earlier stage biopharmaceutical firms. This can lead to fewer projects originating in early-stage biopharmaceutical firms progressing to later stages of R&D process along parallel paths. Furthermore, the attendant reduction of R&D facilities and resources accompanying mergers is likely to reduce the number of internally originated projects compared with what would occur from two separate R&D-oriented firms. They are skeptical that these mergers will produce any offsetting benefits from economies of scale or scope in R&D, given the evidence from the available literature.

Although these issues and concerns did not prevent the FTC approval of large-scale mergers such as Pfizer–Wyeth or Merck–Schering Plough in 2009, they raise interesting questions for further research. In particular, even if mergers between large firms reduce overall R&D expenditures and the number of projects undertaken, it is important to understand the balance of impacts between internally generated projects and externally supported ones. Second, it is important to know the effects on early-stage versus later-stage development projects and firms. As discussed, at the present time, there appears to be a strong demand for late-stage product candidates with substantial market potential to fill pipeline gaps, so adverse effects, if any, may be primarily on deals for early-stage companies. A number of analysts have raised more general concerns about a funding gap in the technology transfer development process for start-ups with technologies that have proceeded beyond university-type basic research, but still are many years away from any commercial applications, as discussed below.

Biomedical Research Support and Technology Transfer Activities in the USA and Europe

Just as small, development-stage R&D firms now play a significant role in innovative research leading to commercialized biomedical products; in the past 25 years the university has assumed a far greater role. With the passage of the Bayh–Dole Act in 1980, and later the Stevenson–Wydler Act in the USA, universities were given increased new rights to patent federally funded research discoveries (Rai and Eisenberg, 2003). Bayh–Dole's policy goal was to increase the investment of private sector development funds for translating university research funded by federal monies into new products and processes. Before 1980, universities received fewer than 250 patents annually, compared with 3000 per year by 2002 (Association of University Technology Managers, 2003). This, in turn, has led to a large increase in university–industry licensing agreements as well as start-up companies originating from university R&D (National Academy of Sciences, 2004).

In Europe, public funding of biomedical research also has increased dramatically since the 1980s, even though total spending has remained significantly lower than in the USA. In this regard, Germany spends the most public funds on biotechnology, followed by the UK, and France (Pammolli *et al.*, 2002). In contrast to the USA (and to some extent also the UK), biomedical research in continental Europe tends to be concentrated in public research institutions and highly specialized university laboratories with little interaction with teaching, clinical practice, and industrial research (Gambardella *et al.*, 2000).

Europe has also been characterized by less mobility and interaction between university and public institution scientists on the one hand, and industry research activities and personnel on the other hand. Rather a more prevalent model is the establishment of specialized institutions, such as science and technology parks, to act as intermediaries between

publically supported biomedical research and industry (Gambardella *et al.*, 2000). To date, the US model appears more productive from the standpoint of creating new companies and fostering new technologies. However, the Bayh–Dole Act and its enhanced incentives toward licensing and commercialization have also been the subject of criticism in terms of potentially undermining the norms of open science (Dasgupta and David, 1994).

Policy Initiatives to Improve the Translation of Academic Science into New Therapies

One important issue that has emerged in recent years regarding technology transfer is a growing belief that there is a biomedical funding gap associated with early-stage preclinical R&D. These 'proof of concept' type activities (studies which provide early evidence a molecule may feasibly be developed for a particular use) are beyond the basic research questions typically investigated by university researchers. At the same time, many venture capital and private equity firms have pulled away from funding early-stage discovery companies and focused instead on companies with compounds in clinical trials. The excitement preceding and accompanying the announcement of the 'working draft' of the human genome sequence in 2000 created a surge of worldwide interest and private early-stage investment in genome-based projects. But assumptions about the time frames required to commercialize a gene-based product and the market acceptance of these new products were unrealistic, thus a period of overinvestment followed by negative returns, in effect, a bubble occurred (Klausner, 2005).

As noted in Section 'Introduction,' alliance activity has been more focused on clinical development activity in the past several years compared with early-stage research. The same appears to be the case for venture funding of start-up companies. The number of life science companies receiving first-time funding dropped from an average of over 250 companies in 2006–08 to less than 175 companies in 2007–11 (Leff, 2012). Average funding for early-stage projects have also been under pressure during this period (PWC, 2012).

The hypothesized early-stage funding gap, sometimes ominously referred to as the 'valley of death,' has been the subject of growing attention by policy makers. A number of recent initiatives to encourage public–private partnerships and other activities are focused at the precompetitive level. Their ultimate objective is to improve the low probability of success and address the R&D productivity problems of drug industry R&D discussed earlier. At the same time, there are skeptics who contend there is adequate funding by the private sector to pursue promising research leads emerging from academic science. They point to other explanations for the low probability of success, such as an increasing focus on therapeutic areas like Alzheimer's disease and oncology therapies that have low probabilities of success from a scientific perspective (Kahn, 2012; Pammolli *et al.*, 2011). Given the recent origin of the policy initiatives designed to improve the translation of academic research and increase success probabilities, it is likely to be several years before one can evaluate their effectiveness. Several of these different initiatives, many of which are at the pilot stage, are considered below.

A number of different models of nontraditional R&D collaborations have emerged to enhance biomedical innovative activity at the precompetitive level in recent years (Altshuler, *et al.*, 2010). One ambitious approach is the establishment of the National Institutes of Health (NIH) Center for Advancing Translational Sciences (NCATS) in fiscal year 2012 with a budget of more than $500 million and several new programs and initiatives. The mission of NCATS is to support research that will reduce costly and time-consuming bottlenecks in the development of new therapies (NIH, About NCATS, 2012). For example, NCATS is currently supporting efforts to identify drug targets faster and more efficiently and is collaborating with industry to develop a consortium to provide a repository and an analytical platform for target validation efforts. In addition, NCATS recently announced a collaborative pilot program with pharmaceutical firms to examine new therapeutic uses for existing molecules. In particular, eight pharmaceutical companies have contributed 58 compounds that have advanced to clinical trials, but were unsuccessful in its original therapeutic indication or not pursued for business reasons. The NIH-industry collaboration will match researchers to test these compounds for new therapeutic uses. These and several other related NCATS programs involve a multiprong approach to reengineer the precompetitive, preclinical process in order to bridge the gap between basic research and human medicines and increase the success rate of later-stage clinical compounds.

A second approach to precompetitive collaboration involves public–private consortia for process innovation or knowledge creation. Some existing examples include the Predicative Safety Testing Consortium (PSTC) established in conjunction with the FDA and the nonprofit Critical Path Institute. PSTC's 18 corporate members share information on safety testing methods and test methods developed by consortium members. The Critical Path Institute leads the collaborative process, and collects and summarizes the data. Another public–private collaborative organization, the Biomarker Consortium, addresses the development of good biomarkers for both safety and efficacy. Like the PSTC, the Biomarkers Consortium encourages scientists at competing firms to contribute their knowledge and expertise to the development of specific biomarkers. As with the PSTC, a public sector-related organization – the nonprofit foundation for the NIH – plays a key role in soliciting funding and selecting research projects (Rai *et al.*, 2008).

A third model involves so-called virtual pharma companies involving nonprofit foundations and patient advocacy organizations that provide funding and project selection oversight to bridge the gap between basic research and later-stage clinical development. Many of these groups are focused on relatively rare and neglected diseases (e.g., Cystic Fibrosis Foundation, the TB Alliance, Gates Foundation initiative on neglected diseases). These organizations have moved from sponsoring academic research proposals on an *ad hoc* basis to coordinating research agendas and activities with milestones and other contracting research approaches prevalent in downstream alliances and partnerships between pharmaceutical firms (Altshuler *et al.*, 2010).

At this point, all of the approaches to bridging the gap between basic biomedical research and therapeutic

development are somewhat experimental in character. They face significant challenges and barriers involving the alignment of incentives among participants with different expertise, resources, and objectives. Establishing a culture of mutual trust and cooperation between different stakeholders is also challenging. Because it can be difficult to delineate precompetitive from competitive activities, collaborations can give rise to free rider problems and conflicts of interest. At the same time, the potential opportunity to advance R&D productivity and facilitate innovative medicines through these precompetitive initiatives appear to warrant the continued experimentation with several forms of public–private entities. In effect, the current strategy of policy makers appears to be 'let a thousand flowers bloom' and see what approach, if any, yields positive outcomes.

Summary and Concluding Comments

As is the case in other industries, mergers in pharmaceuticals are driven by a variety of company motives and conditions. Given this is the case, it is important to take account of firm characteristics and motivations in evaluating merger performance rather than using a broad aggregate brushstroke. Research to date on pharmaceuticals suggests considerable variations in outcomes.

The empirical research on mergers is generally focused on the larger public companies. There is evidence that the large-scale mergers involving these pharmaceutical firms were driven in significant part by a series of industry-wide and firm-specific shocks. These shocks left many firms with R&D pipeline gaps associated with patent expirations (and the increased usage of generics in the USA and more stringent price controls in other countries). Although these mergers apparently have achieved cost reductions and addressed short-run pipeline problems, there is little evidence to date that they increased long-term R&D performance or outcomes. Many of the larger pharmaceutical firms listed in Table 1 continue to deal with a persistent R&D productivity problem.

By contrast, the empirical research on alliances between smaller biotech firms and larger pharmaceutical entities and on development-stage acquisitions is more encouraging in nature. There is evidence of a positive relation between a firm's experience in clinical development and the probability of successful outcomes. In particular, the 'R&D boutique' firms with a small number of research projects can apparently benefit from alliances with larger, more experienced firms, especially at the later stages of the R&D process. The work on alliances provides some support for this hypothesis, but also raises a number of issues for further research.

Our analysis of the effect of mergers on the success rates of drug development projects is generally consistent with these results from the alliance literature. In particular, using data on R&D projects from a large sample of public and private firms, it is found that a company's development experience is significantly related to the likelihood of success, especially for the large pivotal Phase III trials. Moreover, there is suggestive evidence that very small firms with only a few projects in their R&D portfolio can gain the most benefits from mergers with more experienced firms in developing new drug introductions.

Our results, and those of other studies, are subject to various qualifications and raise many questions for further research. The economics literature indicates, for example, that many acquisitions of smaller companies by larger firms are preceded in time by development-stage partnerships. This opens a fruitful line of research in terms of when alliances are a desirable alternative to mergers, and where they can be complementary in nature. More generally, there are a host of interesting research questions to be addressed relating to the various drivers of mergers and the conditions and firm characteristics that produce successful versus unsuccessful mergers. These are important issues from both a business strategy and economic efficiency standpoint.

From an antitrust policy standpoint, the larger horizontal mergers in pharmaceuticals have run into few challenges by the regulatory authorities in the USA and the European Union, given the option to spin-off competing therapeutic products to other drug firms. However, the issue of innovation markets, where firms have potentially competing development programs at various stages of the R&D process, remains a more controversial area of antitrust policy for industries like pharmaceuticals. This remains an important area for future research by law and economics scholars.

Another important area of policy review and debate concerns the apparent funding gap in transferring technology from publically supported basic biomedical research conducted largely in universities and nonprofit institutions to privately supported early-stage R&D activities primarily concentrated in start-up firms and development-stage entities. Both venture capital funding and development-stage alliances have focused more on later-stage clinical activities in recent years. A thriving market for downstream innovation activities depends on the public and private research support in upstream early-stage R&D activity. A number of initiatives and pilot projects are emerging at the NIH and elsewhere to enhance the scientific translational process from basic research to therapeutic medicines. It likely will be several years before one can assess the effectiveness of these collaborative activities.

See also: Biosimilars. Patents and Other Incentives for Pharmaceutical Innovation. Research and Development Costs and Productivity in Biopharmaceuticals. Vaccine Economics

References

Akerlof, A. (1970). The market for 'Lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* **84**, 488–500.

Allain, M.-L., Henry, E. and Kyle, M. (2012). *Inefficiency in the sale of ideas.* Toulouse School of Economics Working Paper.

Altshuler, J. S., Balogh, E., Barker, A. D., et al. (2010). Opening up to pre-competitive collaboration. *Science Translational Medicine* **2**, (52), 52cm26.

Andrade, G., Mitchell, M. and Stafford, E. (2001). New evidence and perspectives on mergers. *Journal of Economic Perspectives* **15**(2), 103–120.

Arora, A., Fosfuri, A. and Gambardella, A. (2001). *Markets for technology: The economics of innovation and corporate strategy.* Cambridge, MA: MIT Press.

Arora, A., Gambardella, A., Magazzini, L. and Pammolli, F. (2007). A breath of fresh air? Firm types, scale, scope and selection effects in drug development. Unpublished paper. Available from the authors.

Arrow, J. (1983). Innovation in large and small firms. In Ronen, J. (ed.) *Entrepreneurship*. Lexington, MA: Lexington Books.

Association of University Technology Managers (2003). *AUTM licensing survey: FY2000*. Northbrook, IL: Association of University Technology Managers.

Belcher, T. and Nail, A. (2000). Integration problems and turnaround strategies in a cross-border merger: A clinical examination of the Pharmacia–Upjohn merger. *International Review of Financial Analysis* **9**(2), 219–234.

Burns, R., Nicholson, S. and Evans, J. (2005). Mergers, acquisitions and the advantages of scale in the pharmaceutical industry. In Burns, R. (ed.) *The business of healthcare*, pp. 223–270. Cambridge: Cambridge University Press.

Carlton, W. (1995). Antitrust policy towards mergers when firms innovate: Should antitrust recognize the doctrine of innovation markets? Testimony before the Federal Trade Commission on Headings on Global and Innovation Based Competition, October 24.

Carrier, J. (2008). Two puzzles resolved: Of the Schumpeter-arrow stalemate and pharmaceutical innovation markets. *Iowa Law Review* **93**. Available at: http://papers.ssrn.com/sol3/papers.cfmabstract_id=977603 (accessed 15.04.13).

Cartwright, H. (2012). A review of deal making in 2011. *Pharma Deals Review* **2012**(1), 15–18.

Centerwatch (2000). Troubling numbers for big pharma consolidation. *In Vivo* **18**, 2–3.

Cockburn, I. (2006). Is the pharmaceutical industry in a productivity crisis? In Jaffe, A., Joshua, L. and Scott, S. (eds.) *Innovation policy and the economy*, vol. 7, pp. 1–32. Cambridge, MA: MIT Press.

Cockburn, I. and Henderson, R. (2001). Scale and scope in drug development: Unpacking the advantages of size in pharmaceutical research. *Journal of Health Economics* **20**(6), 1033–1057.

Comanor, S. and Scherer, F. M. (2009). Comments submitted to the federal trade commission on the Pfizer Wyeth and Merck-Schering plough mergers. Available at: http://www.ftc.gov/os/comments/pfizerwyeth/544915-0004.pdf (accessed 15.09.12).

Danzon, P., Epstein, A. and Nicholson, S. (2007). Mergers and acquisitions in the pharmaceutical and biotech industries. *Managerial and Decision Economics* **28**(4/5), 307–328.

Danzon, P., Nicholson, S. and Pereira, N. S. (2005). Productivity in pharmaceutical-biotechnology R&D: The role of experience and alliances. *Journal of Health Economics* **24**(2), 317–339.

Dasgupta, D. and David, P. (1994). Toward a new economics of science. *Policy Research* **23**, 487–532.

DiMasi, J. and Grabowski, H. (2007). The economics of new oncology drug development. *Journal of Clinical Oncology* **25**(2), 209–216.

DiMasi, J., Hansen, R. and Grabowski, H. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics* **22**, 141–185.

DiMasi, J., Hansen, R., Grabowski, H. and Lasagna, L. (1991). The cost of innovation in the pharmaceutical industry. *Journal of Health Economics* **10**, 107–142.

DiMasi, J. and Pacquette, C. (2004). The economics of follow-on drug research and development: Trends in entry rates and the timing of development. *PharmcoEconomics* **22**(Supplement 2), 1–14.

Dorey, E. (2001). GlaxoSmithKline present a biotech facade. *Nature Biotechnology* **19**(4), 294–295.

Gambardella, A., Orsenigo, L. and Pammolli, F. (2000). *Global competitiveness in pharmaceuticals: A European perspective*. Report Prepared for the Engerprise Directorate General of the European Commission. Available at: http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/comprep_nov2000_en.pdf (accessed 15.04.13).

Gilbert, R. J. and Sunshine, S. C. (1995). Incorporating dynamic efficiency concerns in merger analysis: The use of innovation markets. *Antitrust Law Journal* **63**, 569–602.

Gort, M. (1969). An economic disturbance theory of mergers. *Quarterly Journal of Economics* **83**, 624–642.

Grabowski, H. (2007). Competition between generic and branded drugs. In Sloan, F. A. and Heish, C.-R. (eds.) *Pharmaceutical innovation: Incentives, competition, and cost–benefit analysis in international perspective*, pp. 153–173. Cambridge, UK: Cambridge University Press.

Grabowski, H. (2012). The evolution of the pharmaceutical industry over the past 50 years. *International Journal of the Economics of Business* **18**(2), 161–176.

Grabowski, H. and Kyle, M. (2007). Mergers and alliances in pharmaceuticals: Effects on innovation and productivity. In Gugler, K. and Yurtaglu, B. B. (eds.) *The economics of corporate governance and mergers*. Cheltenham, UK: Edgar Elgar.

Grabowski, H. and Kyle, M. (2012). Mergers, acquisitions and alliances. In Danzon, P. and Nicholson, S. (eds.) *The Oxford handbook of the economics of the biopharmaceutical industry*, pp. 552–577. Oxford: Oxford University Press.

Grabowski, H., Vernon, J. and DiMasi, J. (2002). Returns on research and development for 1990s new drug introductions. *PharmacoEconomics* **20**(Supplement 3), 11–29.

Guedj, I. and D. Scharfstein. (2004). *Organizational scope and investment: Evidence from the drug development strategies and performance of biopharmaceutical firms*. NBER Working Paper 10933. Cambridge, MA.

Haspeslagh, P. and Jemison, D. (1991). *Managing acquisitions: Creating value through corporate renewal*. New York: Free Press.

Henderson, R. and Cockburn, I. (1996). Scale, scope, and spillovers: Determinants of research productivity in the pharmaceutical industry. *RAND Journal of Economics* **27**(1), 32–59.

Higgins, J. and Rodriguez, D. (2006). The outsourcing of R&D through acquisition in the pharmaceutical industry. *Journal of Financial Economics* **80**, 351–383.

Hitt, M., Ireland, R. D. and Harrison, J. (2001). Mergers and acquisitions: A value creating or value destroying strategy? In Hitt, M., Freeman, R. E. and Harrison, J. (eds.) *The Blackwell handbook of strategic management*, pp. 384–408. Malden, MA: Blackwell Business.

Kahn, D. (2012). Connecting the dots in translational research. *Nature Reviews: Drug Discovery* **11**(10), 811–812, doi:10.1038/nrd3357-c2.

Klausner, A. (2005). Mind the (biomedical funding) gap. *Nature Biotechnology* **23**, 1217–1218.

Koenig, M. and Mezick, E. (2004). Impact of mergers and acquisitions on research productivity within the pharmaceutical industry. *Scientometrics* **59**(1), 157–169.

Larson, R. and Finkelstein, S. (1999). Integrating strategic, organizational, and human resource perspectives on mergers and acquisitions: A case survey of synergy realization. *Organization Science* **10**(1), 1–26.

Leff, S. (2012). Measuring the health of the U.S. Biomedical Innovation Enterprise: A venture investor's perspective. *Presentation, Brookings Institution Conference on the State of Biomedical Innovation*. Washington DC, June 27, 2012.

Lerner, J. and Malmendier, U. (2010). Contractibility and the design of research agreements. *American Economic Review* **100**(1), 214–246.

Lerner, J. and Merges, R. P. (1998). The control of technology alliances: An empirical analysis of the biotechnology industry. *Journal of Industrial Economics* **XLVI**(2), 125–156.

Marris, R. (1998). *Managerial capitalism in retrospect*. Available at: http://www.palgraveconnect.com/pc/doifinder/10.1057/9780230376168 (accessed 16.04.13).

Mathieu, M. (ed.) (2007). *Parexel's bio/pharmaceutical R&D scoreboard 2007/2008*. Waltham, MA: Parexel International Corporation.

Morgan, J. (2001). Innovation and merger decisions in the pharmaceutical industry. *Review of Industrial Organization* **19**, 181–197.

Mueller, C. (1986). *The modern corporation profits, power growth and performance*. Sussex, UK: Wheatsheaf Books.

Mueller, C. (1996). Lessons from the United States' antitrust history. *International Journal of Industrial Organization* **14**, 415–445.

National Academy of Sciences (2004). In Merrill, A., Levin, C. and Myers, B. (eds.) *A patent system for the 21st century*. Washington, DC: National Research Council of the National Academies.

Ornaghi, C. (2006). *Mergers and innovation: The case of the pharmaceutical industry*. University of Southampton Working Paper.

Pammolli, F., Allansdottir, A., Bonaccorsi, A., et al. (2002). *Innovation and competitiveness in European biotechnology No. 7*. Office for Official Publications of the European Communities.

Pammolli, F., Magazzini, L. and Roccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nature Reviews: Drug Discovery* **10**, 428–438.

Pammolli, F. and Roccaboni, M. (2004). Technological competencies in networks of innovators. In Cantwell, J., Gambardella, A. and Grandstrand, O. (eds.) *The economics and management of technological diversification*, pp. 48–50. London and New York: Routledge.

Pisano, G. (1997). R&D Performance, collaborative arrangements, and the market for know-how: A test of the "Lemons". In *Hypothesis in biotechnology*. Mimeo. Harvard Business School, Cambridge, MA.

Powell, W., Koput, K. and Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly* **41**, 116–145.

PWC. (2012). *Dollar drought: Life sciences venture capital funding shrinks for fourth straight quarter*. August. Available at: http://www.pwc.com (accessed 15.09.12).
Rai, K. and Eisenberg, R. (2003). Bayh–Dole reform and the progress of biomedicine. *Law and Contemporary Problems* **66**, 289–314.
Rai, K., Reichman, H., Uhlir, F. and Crossman, G. (2008). Pathways across the valley of death: Novel intellectual property strategies for accelerated drug development. *Yale Journal of Health Policy, Law and Ethics* **VII**(1), 53–89.
Rapp, T. (1995). The Misapplication of the innovation market approach to merger analysis. *Antitrust Law Journal* **64**, 19–47.
Ravenscroft, J. and Long, F. (2000). Paths to creating value in pharmaceutical mergers. In Kaplan, N. (ed.) *Mergers and Productivity*, pp. 287–326. Chicago: University of Chicago Press.
Recombinant Capital (2008). *Analysts notebook trends*. Available at: http://www.recap.com (accessed 15.09.12).
Scherer, F. M. (1999). *New perspectives on economic growth and technological innovation*. Washington, DC: The Brookings Institution.
Scherer, F. M. (2010). Pharamaceutical innovation. In Hall, B. and Rosenberg, N. (eds.) *Handbook of the economics of innovation*, vol. 1, pp. 542–543. North Holland: Elsevier.
Smith, K. and Quella, J. (1995). Seizing the moment to capture value in a strategic deal. *Mergers and Acquisitions* **29**(4), 25–30.
Teece, J. (1998). Capturing value from knowledge assets: The new economy, markets for know-now and intangible assets. *California Management Review* **40**(3), 55–79.

Further Reading

Ernst & Young Online (2007). *Beyond Borders: Global Biotechnology Report 2007*. Available at: http://www.ey.com/beyondborders (accessed 15.09.12).
U.S. Food and Drug Administration, Center for Drug Evaluation and Research (Center for Drug Evaluation and Research, 1999). *From test tube to patient: Improving health through human drugs (special report)*. Washington, DC: US Government Printing Office. September.
National Institute of Health, About NCATS (Available at: http://www.ncats.nih.gov/about/about.html (accessed 15.04.13)).

Missing Data: Weighting and Imputation

PJ Rathouz, University of Wisconsin School of Medicine & Public Health, Madison, WI, USA
JS Preisser, University of North Carolina, Chapel Hill, NC, USA

Introduction and Goals

Missing Data and Their Consequences

Missing data are common occurrences in health economics and health outcomes research. The patterns of missing data can take many forms. Item missingness occurs when values are missing for selected variables (items) for a subset of subjects or cases; often the subset of cases missing one variable may not be identical to those missing another variable, thus compounding the problem. In contrast, case (record) missingness occurs when subjects selected for a study might have only minimal available data because they might be unreachable or even decline to participate. In panel (longitudinal) data, subjects might not be missing, but records for some waves might be. Dropouts are common, just as intermittent missing data arising when a subject misses a wave but participates in subsequent waves.

Missing data, even if relatively modest in scope, present a major problem when their frequency and structure call into question the validity of and conclusions from statistical analyses of a research study. The two main statistical issues are increased variability and bias. As a general rule, the loss of information resulting from the missingness increases the variability of statistical estimates, sometimes dramatically so, resulting in decreased power for hypothesis tests and wider confidence intervals. A more serious problem is bias, that is, systematic error that may result when the missing data occur disproportionately across subjects or records. Selection bias, which typically occurs when the individuals in the sample are not representative of the target population, may also be induced by nonresponse. The missing data may therefore result in reduced generalizability of study results, and may complicate statistical analyses through the need to mitigate bias.

Statistical Approaches to Missing Data: Overview

Analytic strategies for handling missing data have a long history and fall into several broad categories. The authors mention a few approaches here, and then in Sections 'Weighting' and 'Multiple Imputation' elaborate two of the most useful and popular ones for missing data.

Historically, methods were focussed on simple imputation of missing values with the aim of conducting analyses using statistical tools that required complete data. Simple imputation techniques such as unconditional mean imputation – replacing missing values with the sample mean of nonmissing values – or carrying the preceding nonmissing observation forward in a panel study were standard approaches. Conditional mean imputation improved slightly on these approaches as it replaces missing values with their sample mean on the basis of a regression model or a sample subset of the full data that is closer in some sense to the record with the missing data. In some settings, this method will yield reasonably unbiased estimators, but will often understate statistical uncertainty and inflate test statistics. Carrying the preceding nonmissing observation forward in panel studies is not recommended.

The most common strategy for handling missing data, complete-record analysis, is to simply delete the records for which data are missing, and base the analysis only on the observations without missing data. In univariate response problems, complete-record analysis ignores data on all subjects for whom some data are missing for a given analysis. In panel data, a subject may contribute data, but records for waves with missing data are ignored. This is referred to as available subject/complete-record analysis. (A subject without any missing items or waves is called a complete subject.) There is really no statistical principle guiding these decisions; rather it is mostly a matter of simplicity and convenience for software developers and analysts. Besides possible problems of bias and inefficiency, complete-record analyses suffer from the problem of different sets of records being included in different model fits, depending on the variables included in the model.

Modern methods for handling missing data frame the problem in terms of a pair of auxiliary statistical models: A model for the distribution of the missing values, and a model for the probability that a given value is missing, sometimes called the missingness mechanism. Sensitivity of inferences to these model specifications, and manipulation of them in order to obtain inferences, has been the topic of a vast missing data literature.

A minimal goal of any method is that it yields valid statistical inferences. For a method to be valid, it should produce consistent (i.e., unbiased in large samples) estimators of the parameters of interest, which are accompanied by consistent estimators of uncertainty (e.g., appropriate standard errors), thereby yielding correct test statistics and confidence intervals. Statistical efficiency (i.e., optimal use of the available data under acceptable modeling assumptions) is another important statistical desideratum.

With these three criteria in mind, one large body of approaches involves correcting bias and generating valid standard errors in complete-record analysis. Whereas other techniques are available, the most common of such approaches is the statistical weighting of complete records. In this approach, efficiency is more of a secondary concern. A key feature of weighting approaches is that a model for the distribution of the missing quantities is not needed. A model for the missingness mechanism is, however, needed in order to generate the weights.

At the opposite end of the spectrum from weighting lies full maximum likelihood estimation, which jointly estimates the model of interest and the model for the missing quantities. Maximum likelihood places the two issues of bias and

 doi:10.1016/B978-0-12-375678-7.00705-7

efficiency on equal footing. If the models are close to correctly specified, then bias will be eliminated, and the solution will be statistically efficient. However, bias and/or inefficiency can arise under misspecification. This approach is also limited because often the analysis of interest (i.e., the analysis one would do in case of no missing data) is not based on maximum likelihood.

For purposes of approximating maximum likelihood estimation, more advanced and flexible imputation techniques have been proposed. The simplest of these is conditional imputation, in which a probability model is specified and estimated for the missing quantities as a function of other nonmissing variables. A random draw from that model is then used to impute each missing value. This method improves on conditional mean imputation because it captures the variability in the data that would have been observed if they had not been missing. For small amounts of missing data, this method can work well. It is convenient and quick, not much more difficult than conditional mean imputation, and data preserving too. Its disadvantage is that the resulting measures of statistical uncertainty will be anticonservative because the approach does not account for the additional variability arising from having estimated the imputation model. Multiple imputation (MI) resolves this issue.

Goals for the Article

It is beyond the scope of this article to cover all missing data methods or to cover any one method in detail. Rather, the authors' main goal is to introduce key concepts and to provide some guidelines to help the analyst narrow down problems so that he/she may further research specific solutions. With this in mind, in the remaining sections, the authors will focus on the following three objectives. First, a discussion of missing data identifying assumptions, especially missing at random (MAR) and its variants, will be provided. Second, some leading case patterns of missing data with probably valid approaches in those settings will be discussed. Third, two model-driven approaches that are commonly used now, namely, complete-record analysis with weighting, and MI will be described in some more detail. Conditional single imputation, which is described above, is a special case of MI, so that approach will be implicitly covered as well. To focus the discussion, regression problems for univariate or for longitudinal (panel) data will be emphasized. Wherever needed, the settings of missing predictors and missing responses (e.g., attrition) will be differentiated. Also, the authors will emphasize statistical methods for MAR data. Such methods are generally easier to carry out than those for nonignorable missingness, and are more broadly applicable than those analyses that require the relatively severe assumption of missing completely at random (MCAR).

This article focuses on data that are missing by happenstance. There are many situations wherein data are missing by design, either at the item level or at the record or the unit level. Important examples include two-phase sampling designs, case-cohort designs, and outcome-dependent sampling designs for panel data. Some references for these designs are given under 'Further Reading'.

Foundational Issues

Preliminaries and Notation

To fix ideas, suppose that the analysis of interest involves the conditional distribution of response Y given predictors X, denoted as $[Y|X]$. Response Y may be univariate or, in the case of panel data or multivariate data, a vector. Predictors X are almost always row-vector valued or of higher dimension. For example, in panel data, X might be a matrix with number of rows equal to the length of Y. For clarity of exposition, an independent individual or sampling cluster is referred to as a sampling unit, and observations within that unit are considered as records. Units are considered to be stochastically independent, whereas observations within a unit may be correlated. For example, in a panel study, each participant would be a unit, and repeated waves of measurement would generate a separate record each. In a study of twins, the pair would constitute a unit, whereas each of the twins would yield his/her own record.

It is assumed that without missing data, the target of inference would be restricted to $[Y|X]$; the marginal distribution $[X]$ would be ignored. Therefore, in structural equations models wherein predictors are latent and are measured imperfectly by a set of manifest indicators, those indicators become part of Y because the measurement model is part of the analysis that takes place in the absence of missing data. Where no further specificity is needed, and recognizing that it is difficult to create a notation that is broadly applicable, Y_o and X_o will denote components of Y and X that are always observed, while those components that are potentially missing (i.e., missing for some individuals) will be denoted by Y_u and X_u.

In any analysis, it is important to take account of the pattern of missing data. Item nonresponse arises when the missing quantities Y_u or X_u constitute a set of component variables available in Y and/or X. Missing items may be similar or may vary considerably across units. For example, it may be that due to the sensitive nature of the data, many individuals are missing reports of household income. Alternatively, in a questionnaire with many items, some respondents will skip a few items or sets of items here and there, with no consistent pattern from one respondent to another. Unit nonresponse leads to record missingness, a situation where all of Y is missing, as is most of X. Depending on the sampling frame, some minimal part of X might be available. For example, for a nonresponding household in a sample survey, the investigator would still know the address and some neighborhood characteristics according to census information.

Panel data yield special types of record missingness. Most commonly, some participants drop out of a study, leading to attrition that results in monotone missingness patterns. Alternatively, some participants may not come for a given visit or respond at a given wave, but then respond in later waves, leading to wave nonresponse and intermittent missingness patterns.

For purposes of both formal theory and computational development, it is useful to introduce a variable R encoding the missingness pattern for any given unit. The specific form of R will depend on the nature of the missingness. For example,

if the only thing ever missing is a single predictor X_u, then R will be a binary indicator equal to 1 if X_u is observed, and to 0 if X_u is missing. Alternatively, in a panel study with intermittent missingness, R may be a vector of indicators with elements of 1 for observed and 0 for missing waves. R may take other forms that are necessary to fully codify the range of missingness patterns.

Some missing data analyses include auxiliary data Z. These additional variables are observed on all units that are associated with, and hence provide information on, the missing data. Auxiliary data can be quite informative in missing data problems. For example, if a key predictor is clinic-measured body mass index and it is missing on some participants, a reasonable pair of auxiliary variables can be self-reported height and weight. It is assumed that auxiliary variables are not part of the target of inference, i.e., the analyst would ignore Z in studying $[Y|X]$ if there are no missing data. For auxiliary data to be of use, it must be predictive of the missing data Y_u or X_u.

A Taxonomy of Missing Data Assumptions

Rubin (1976) has proposed a taxonomy of missing data assumptions, which has proven to be enormously successful; almost all modern missing data methods refer to it for formal identifying assumptions. Although his taxonomy is rooted in formal likelihood theory, the main ideas are sketched here, downplaying technical details.

Missing data assumptions are framed in terms of the missing data mechanism. This term does not describe the actual machinery or physical processes that lead to data being missing, but rather the stochastic dependencies of those processes on other variables at play in the analysis. Specifically, missingness mechanism refers to the probability distribution of missingness being conditional on (Y, X), viz. $[R|Y, X]$. Of course, understanding the physical processes may help the investigator in determining what is reasonable to assume where dependencies in the missing data pattern are concerned. In general, acknowledging and accounting for this probabilistic selection mechanism is central to understanding, developing, and applying modern missing data methods.

Data are considered to be MCAR when the probability of missing data is independent of both missing and observed variables. That is, R is independent of either (Y, X) or, when auxiliary data are available (Y, X, Z). Intuitively, under MCAR, complete records arise as if by a random sample from the population, because whether they are complete or not is unrelated to their realized (observed or unobserved) values. In univariate problems, complete-record analysis is a valid, albeit statistically inefficient, approach under MCAR. In panel data, under MCAR, either available subject/complete record or complete subject analysis will lead to valid but inefficient inferences.

The most common missing data assumption is MAR. Under MAR, R is independent of (Y_u, X_u) given (Y_o, X_o) or, in the case of auxiliary data, given (Y_o, X_o, Z). Under MAR, the distribution of unobserved data (Y_u, X_u) when $R=0$ could deviate from that when $R=1$, but not when controlling or stratifying on (Y_o, X_o, Z). That is, for a fixed value of (Y_o, X_o, Z), the values (Y_u, X_u) for $R=1$ (or $R=0$) are a random draw from their conditional distribution $[Y_u, X_u|Y_o, X_o, Z]$. MAR is a weaker assumption than MCAR (MCAR is a special case of MAR), but still yields identifiable models. Intuitively, the implication is that it is possible to model this conditional distribution, thus enabling generation of imputed values for (Y_u, X_u) for those subjects where they are missing. This intuition for model identifiability holds even if imputation is not formally pursued as an analysis strategy. Note that auxiliary data Z can be an important factor in meeting the MAR assumption; intuitively if Z is predictive of Y_u or X_u, it can lead to valid imputations of the missing values, imputations that would not have been available in the absence of Z.

It is useful to consider two special cases of MAR. In the first, missingness of Y_u or X_u only depends on the other covariates X_o, but not on Y_o or any auxiliary data. Because interest lies in the conditional distribution $[Y|X]$, it is possible to treat this covariate-dependent missingness essentially as MCAR. Alternatively, if missingness depends on Y_o, the authors have a true MAR situation and must account for this more carefully in subsequent analyses. Two key examples of this situation are as follows. In the first, a covariate X_u is sometimes missing, and whether or not it is missing depends on the response Y. For example, X_u may be a clinical laboratory test and the ordering of the test may be related to response Y. The second arises in panel data with attrition, where Y_o represents the responses before dropout, and Y_u represents the responses realized but not observed after dropout. This situation is termed sequential MAR if the potential missingness of the responses Y_u (after dropout) depends only on the covariates X and the components Y_o of Y being observed before dropout.

When data are not MAR, they are missing not at random (MNAR) or informative. In this case, the fact R that X_u (or Y_u) is missing is associated with its value, even after controlling for always-observed data (Y_o, X_o, Z). When missingness is nonignorable, statistical analyses can be considerably complicated because they require a model for the missingness process, and this model is based on untestable assumptions regarding the relationship of Y_u or X_u to R. As such, sensitivity analyses are often conducted to assess how inferences change over a range of parameter values for the nonignorable missingness process.

MNAR data are often termed nonignorable. The implication is that under MAR, the missingness is 'ignorable' in some sense. This is a potentially misleading statement. MAR allows the model to be identified, but the analyst cannot ignore the problem, or does so at his own peril. Indeed, whereas inferences can be safely based on the likelihood for the complete data, MAR missingness is only truly ignorable if what is missing is part of Y when there are no auxiliary data, and also if full likelihood analyses are pursued. That is, inferences based on likelihood $[Y_o|X]$ instead of on likelihood $[Y|X]$ are valid. If, however, it is a component of X that is missing, then a probability model for X_u is required, something that would not have been part of an analysis with no missing data. If nonlikelihood or conditional likelihood methods are being used, a model for the missing data mechanism is required.

Approaches: When to use What?

Before approaching any analysis, an assessment of missing data should be undertaken. One should quantify how much each variable is missing, and also document patterns of missingness. Are there a few variables missing often, or many variables wherein each is missing very occasionally? What is the relationship among the missingness of different variables? Is it the response of one or more predictors that is missing? It is then important to assess reasons for missingness during the study as this would lead to realistic assumptions regarding the mechanisms of missingness that will aid in the choice of statistical analysis.

Harrell (2001) provides rough guidelines for handling missing predictor variables X_u in univariate regression models. For 5% missingness or less, unconditional mean imputation of X_u will generally work fine. If 15% or less of observations are missing X_u, he recommends imputing it with conditional mean imputation without the response Y in the imputation model. As the missingness approaches 15%, standard errors may begin to be underestimated. If the missingness R is associated with Y (something that can be examined with the data), this approach will also begin to break down. For analyses where more than 15% of observations have at least one missing predictor and/or where missingness is strongly associated with the response, incorrect modeling assumptions could begin to introduce bias, and standard errors need to account for missing data. Weighting or MI (Section 'Multiple Imputation') is recommended.

For the missing responses Y_u in univariate models, if MAR can be assumed without auxiliary data Z, then complete-record analysis will yield valid and efficient inferences. If there are auxiliary data Z, one should examine whether they predict Y_u and/or missingness of Y_u (i.e., R). If Z is strongly predictive of Y_u but not related to missingness R, then imputation will increase statistical efficiency. If Z is related to R but not very strongly to Y_u, then weighting as a function of (Z, X) will help to avoid bias. If Z is related to both R and Y_u, then imputation is recommended to avoid bias and maximize statistical efficiency. In any case, for 5% missingness or less, complete-record analysis should work fine.

For panel data with MAR responses Y_u, for example, due to attrition, if the missingness only depends on X, but not on observed responses Y_o, available subject/complete records analysis using either maximum likelihood or semiparametric moment-based methods will yield valid and efficient inferences. If the missingness of Y_u also depends on Y_o, the best approach depends on the type of analysis being pursued. If the model is a full probability model with no auxiliary data Z, which otherwise would be fitted with maximum likelihood in the absence of missing data, then one should pursue the same approach on available subjects/complete records. In settings where a moment-based analysis is desired or where auxiliary data Z are available, then other approaches are required. Imputation is a possibility but it can be difficult to implement with both variably spaced observations and dropout at different waves in the study. Weighting is an alternative approach that relies on a model for the missingness or the dropout probability (see Section 'Weighting'). As this is a binary data model, modeling strategies for the missingness are straightforward to implement with sufficient flexibility. Covariate missingness in panel data is a more complex situation and beyond the present scope.

Weighting

Univariate Data

The general practice of weighting observed data to account for missing data comes from weighting for nonresponse in sample surveys to correct for bias, which has been discussed elsewhere in this volume. Briefly, when nonresponders differ from responders according to their distribution of measured characteristics, the responders' data are weighted so that analysis restricting to the complete data sample would resemble the analysis of the combined data of responders and nonresponders in the case where it had been observed. Use of weights assumes knowledge of some variables such as demographic characteristics, to be available on the nonresponders so that they can be placed in groups or bins with responders. The bin-specific inverse probability of being a responder is then calculated and used as the weight for observations in that bin. The same principle can be applied in a prospective study where the outcome Y at the end of the follow-up period is missing for some subjects. The general idea is to weight records inversely to their probability of being observed such that observed data with a low likelihood of being observed receive relatively high weight. When incorporating continuous as well as categorical baseline variables, instead of bins, a logistic regression model can be used to estimate the probability of being observed at follow-up for each study participant. The resulting inverse probability is then the weight used in the complete-data analysis for the outcome. The same general method of constructing weights applies to missing covariates in regression settings. When more than one covariate is missing or when missing data patterns for repeated measures are nonmonotone, implementation of weighting strategies becomes less straightforward. The next section considers the special problem of weighting for missing data due to subject dropout in panel data.

Weighted Estimating Equations for Panel Data with Dropouts

A common problem in panel studies is subjects dropping out before study completion. Whereas subject-specific random effects models estimated with maximum likelihood are a popular approach to the analysis of longitudinal data, another approach is population-averaged modeling of marginal means. Such semiparametric estimation of marginal mean regression models is especially attractive for discrete outcomes because specification of their full joint distribution is not needed. In particular, a semiparametric estimation approach using record-specific weights derived from a model for dropout addresses potential bias due to sequential MAR dropout.

To define the problem, assume that there is a set of T planned measurement times being common to all individuals, and that interest lies in the mean μ_t of response Y_t across times $t=1,\ldots,T$. However, some subjects drop out of the study before T, thereby creating a monotone pattern of missingness. Letting t' be the final observation time before dropout (for a

given subject), $Y_o=(Y_1,\ldots,Y_{t'})$, $Y_u=(Y_{t'+1},\ldots,Y_T)$, and $Y=(Y_o, Y_u)$ is obtained. To indicate missing data, $R=(R_1,\ldots,R_T)$ is also obtained, where $R_t=1$ for $t\leq t'$ and 0 otherwise. Next let X_t be the covariate vector which is used to predict μ_t, so that μ_t is a function of $X_t'\beta$. Assume that all X_t's are known, as is the case, if all covariates are baseline characteristics (time-independent) or deterministic functions of known quantities such as the planned measurement times, or are external to the measurement process.

Generalized estimating equations (GEEs) provide a common approach to marginal mean regression modeling for panel data while accounting for correlation among an individual's repeated responses. With dropout, GEE estimates β by solving the equations

$$\sum D_o V_o^{-1}(Y_o-\mu_o)=0,$$

where the sum is over all subjects in the panel. Here, $\mu_o=(\mu_1,\ldots,\mu_{t'})$, and V_o^{-1} is the inverse of the working covariance matrix of Y_o (i.e., the observed part of Y). $D_o=\partial\mu_o/\partial\beta$ and is a generalization of X_t in the normal equations of a linear regression model. GEE yields a consistent estimate of β even if the covariance structure is misspecified, provided the missing data are MCAR or when missingness depends only on covariates X_t. However, under MAR, a GEE analysis generally gives biased estimates.

In contrast, weighted GEE (WGEE) is valid under MAR even if the covariance structure is misspecified, provided the model for the probability of dropout is correctly specified. WGEE modifies GEE by solving instead the equations

$$\sum DV^{-1}W(Y-\mu)=0,$$

where μ, V^{-1}, and D now correspond to the full data vector Y, both observed and unobserved components, and W is a diagonal weight matrix with components $R_t w_t$, $t=1,\ldots,T$. Note that, because W is diagonal, it multiplies, or weights, each element $(Y_t-\mu_t)$ of $(Y-\mu)$ by $R_t w_t$. As in univariate data, w_t is equal to the inverse probability of the tth record being observed, and R_t selects the observed components of Y. In the case of a working independence covariance assumption, V is diagonal, and WGEE reduces to GEE, where W_o is diagonal with elements $w_1,\ldots,w_{t'}$. Compared to WGEE with a nondiagonal working correlation matrix, independence WGEE is rather straightforward to implement as it utilizes the same data structure as the independent GEE, with little efficiency loss when the number of subjects is large.

The benefits of WGEE relative to GEE are twofold. First, just as with the univariate analysis already described, the inverse probability weights in W serve to correct the bias due to nonresponse at the record level. Second, through the working correlation structure posited in V, multiplying by V^{-1} serves to implicitly impute components of Y_u via the observed data in Y_o. There are three main steps in implementing WGEE:

Step 1: Determine observation-specific weights w_t, $t=1,\ldots,T$, through a model for the missingness.
Step 2: Apply the weights to W in the WGEE equations and solve to estimate the $\hat{\beta}$ parameters in the marginal mean model for μ_t.
Step 3: Calculate the empirical sandwich-type estimator of the variance–covariance of $\hat{\beta}$ with the estimation of weights taken into consideration.

In Step 1, under sequential MAR, let λ_t denote the conditional probability that Y_t is observed, given Y_{t-1} is observed. This probability is typically modeled with logistic regression of R_t on X, prior responses $Y_1,\ldots,Y_{t-1}$, and possibly auxiliary data Z. Weight w_t is the inverse of the unconditional probability of being observed at wave t, estimated as the inverse of the cumulative product of conditional probabilities, $\hat{w}_t^{-1}=\hat{\lambda}_1\times\cdots\times\hat{\lambda}_t$.

With weights fixed in Step 2, β estimation is similar to that in classical GEE, alternating between regression and covariance parameter estimation. Under correctly specified models for the marginal means and the dropout process, WGEE yields a consistent asymptotically normal estimator of β. In Step 3, a sandwich estimator is generally used. This estimator is biased for the true variance of $\hat{\beta}$, because it treats the weights w_t as fixed even though they are estimated. In contrast to sampling theory, where fixed weights are used, semiparametric theory provides that this version of the weighted sandwich estimator tends to overestimate the true variance so that options for specifying fixed weights in widely available software for GEE tend to provide conservative estimates of the standard errors for $\hat{\beta}$.

In summary, the WGEE procedure corrects for bias when the GEE assumption of MCAR is in doubt. A few caveats are in order. As in GEE, choice of the working correlation structure in WGEE may affect efficiency. A particular concern with the WGEE method is that misspecification of the missingness model may cause bias in $\hat{\beta}$, which can even exceed that of GEE. Accurate choice of the working correlation can mitigate this problem. As with sample survey weights, caution should be used in applying very large weights. Finally, more complex extensions of this procedure are needed to mitigate efficiency loss.

Multiple Imputation

Overview

For relatively small amounts of missing data, single imputation is an approach that has many advantages. It is valid under MAR and can easily incorporate auxiliary data in the imputation model. It fully exploits the available data. Because data can be imputed before analyses, a single set of imputations can support multipleanalyses. Indeed, the imputation can even be done by a different data analyst other than the final data analyst. Single imputation replaces missing values with those drawn from a fitted distribution, obtained via the nonmissing values. Being drawn from a distribution, the missing values retain the same variability that would be seen if the data had not been missing.

The main disadvantage to single imputation is that it does not account for the fact that the model for imputing the data is itself estimated from the data and not known *a priori*. Specifically, once data are imputed, the final analysis produces estimates and standard errors as if the data are complete. MI solves this problem, thereby providing a complete inferential framework for conducting any type of statistical analysis when there are missing data.

MI proceeds via a three-step process. For ease of exposition, the authors sketch that process for the problem of imputing possibly multivariate missing predictors X_u; the process is similar for other missing data patterns, for example, if the missing data are Y_u.

Step 1: Posit a Bayesian model for $[X_u|Y, X_o, Z; \phi]$, governed by parameter ϕ. Specify a prior for ϕ and use the data with Bayesian inference to generate a posterior distribution for ϕ.

Step 2: Randomly draw a value $\tilde{\phi}$ from the posterior distribution of ϕ.

Step 3: For each missing X_u, randomly impute from the distribution $[X_u|Y, X_o, Z; \tilde{\phi}]$. Steps 2 and 3 are repeated M times to produce M imputed 'complete' data sets. The M data sets can then be analyzed using any valid statistical procedure. Historically, M was set at 5 or 10; with experience, methodologists have learnt that $M = 20$ often provides more reliable statistical performance with insignificant increases in computing costs or analyst's time.

MI yields several important advantages over other methods for handling missing data. The M imputations can be generated independently of the analysis using whatever variables are available. The analyst neither needs to explicitly account for the exact specification of the imputation model $[X_u|Y, X_o, Z; \phi]$ nor for the auxiliary variables Z being used in the imputation. Rather, all necessary information to make inferences is contained in the M imputed data sets. In particular, because each imputation is based on a different draw $\tilde{\phi}$ from the posterior of ϕ, the MIs capture the additional uncertainty due to the estimation of the imputation model. One implication is that the imputations can be generated as part of preliminary data processing, before generating analysis data sets. For example, in a publicly available national survey, the organization performing the survey can generate the imputed data sets and make them available for download with the original non-imputed data. A final advantage is that MI, via modern Bayesian resampling methods, is capable of flexibly handle both monotone and nonmonotone missing data patterns.

Computational Details

It is beyond the scope of this article to cover the statistical theory underlying MI. Nevertheless, it is useful to present the few central formulae prevailing in all MI routines, partly because they emphasize the simplicity and portability of the method across various analysis settings and approaches, and partly because they provide a vehicle for explicating about how MI incorporates the uncertainty due to missing data into the final analysis.

Picking up from Step 3 above, analysis is completed in two additional steps. Assume that interest lies in parameter $\hat{\theta}$ governing $[Y|X; \theta]$. For ease of exposition, the authors assume that θ is unidimensional. The extension to multidimensional $\hat{\theta}$ is straightforward, but involves both vector and matrix arithmetic, and is therefore more challenging to present.

Step 4: For each j from $1,\ldots,M$, the jth imputed data set is analyzed as if the data are complete, obtaining estimates $\hat{\theta}^{(j)}$ with corresponding standard errors $\sqrt{V^{(j)}}$.

Step 4 is repeated M times. Note that any type of analysis can be pursued to obtain estimates $\hat{\theta}^{(j)}$ and standard errors $V^{(j)}$. This could, for example, be a maximum likelihood or a generalized method of moment-based analysis. The key assumption is that the sample size is large enough for the $\hat{\theta}^{(j)}$s to be approximately normally distributed. Of practical importance, the repeated analysis in Step 4 can in almost all cases be automated, and several modern statistical packages (e.g., SAS and Stata) have written MI wrappers that can be applied to any standard analysis being available in the package.

Step 5: Estimates $\hat{\theta}^{(j)}$ and $\sqrt{V^{(j)}}$ are combined to yield final estimates $\hat{\theta}$ and $\sqrt{V}$, using the formulae shown in **Table 1**. In particular, the overall (and final) estimate of θ is the average of the M estimates. The overall variance V is comprised of two parts, the average within-imputation variance and the (scaled) between-imputation variance. The overall standard error $\sqrt{V}$ can be used in the usual way to construct normal or t-based hypothesis tests and confidence intervals. Owing to the fact that M is usually not large, Rubin (1987) has recommended about using the t-distribution with v degrees of freedom (df) for critical values, where v is given in **Table 1**. Notably, v goes to ∞ as M becomes large, or, for fixed M, as the amount of missing data become small. Specifically, as the proportion of missing data shrinks, the imputation specific estimates $\hat{\theta}^{(j)}$ will become closer and closer to one another so that the between-imputation variance will naturally shrink to zero, causing v to go to ∞.

Multiple Imputation in Practice

To implement MI in practice, the imputer needs to specify a model for $[X_u|Y, X_o, Z; \phi]$ as well as a prior for ϕ. For continuous components of X_u, the model is often based on linear regression, whereas for categorical, it is often based on logistic regression as well as its extensions for categorical and ordinal data. There is considerable benefit in having this model be fairly flexible, so if the sample size supports it, one should include nonlinear terms for continuous predictors as well as key pairwise interactions between predictors. Vague priors are generally used in an attempt to reflect a state of relative ignorance regarding ϕ.

Monotone missingness patterns are more straightforward to model because the joint distribution of all components of

Table 1 Formulae for combining multiple imputation estimates

Description	*Formula*
The overall (and final) estimate of θ[a]	$\hat{\theta} = M^{-1}\sum_{j=1}^{M}\hat{\theta}^{(j)}$
The average within-imputation variance[b]	$\overline{V} = M^{-1}\sum_{j=1}^{M}V^{(j)}$
The between-imputation variance	$B = (M-1)^{-1}\sum_{j=1}^{M}(\hat{\theta}^{(j)}-\hat{\theta})^2$
The overall, or total, variance	$V = \overline{V} + (1+M^{-1})B$
Degrees of freedom for t-distribution[c]	$v = (M-1)\left[1+\frac{V}{(1+M^{-1})B}\right]^2$

[a] M is the number of imputations and $\hat{\theta}^{(j)}$ is the estimate in the jth imputed sample.

[b] $V^{(j)}$ is the variance of $\hat{\theta}^{(j)}$ in the jth imputed sample.

[c] v can be expressed as $v = (M-1)[1+(1/\tau)]^2$, where τ is the relative increase in variance in $\hat{\theta}$ due to missing data. Owing to this simple interpretation, the quantity τ can be useful in study design and sample size calculations when it is known that some missing data will be unavoidable.

X_u can be modeled via a series of conditional models. In more complex nonmonotone missing data patterns, full conditional resampling methods are often used, wherein one component of X_u is modeled and imputed on the basis of imputed values of all the other missing components of X_u. This process is iterated with new samples of ϕ and all components of X_u in order to generate M imputed data sets.

In modern statistical packages, much of the missing data modeling and imputation is handled in procedures that automatically determine the scale of each component of X_u, specify a default model, determine whether the pattern is monotone or not, specify a prior for ϕ, and ultimately generate the imputed data sets. Default specifications can be overridden, but with modest levels of missingness, the defaults are often perfectly adequate. Those same packages often contain tools for carrying out the postanalysis summaries and inferences as described in Step 5 above. MI for clustered or longitudinal data is less systematically implemented in standard software packages. These situations are more complicated, especially if the data are imbalanced.

Finally, a comment regarding robustness. Certainly, there is a need for model flexibility in order to capture the key relationships of X_u to (Y, X_o). And, with a fairly extensive modeling machinery needed for complex missing data patterns, there are plenty of opportunities for misspecification. Nevertheless, simulation work and empirical evidence have shown that final inferences for θ are often quite robust to specifications of the missing data model. It is believed that the underlying reason for this is that most data are not missing. In most cases, less than 30% of data are missing, often even less. Hence, there is much to be gained by doing imputations even if the imputation model is only correct to first order. The degree to which incorrect models can lead to bias is in some sense bounded by the fact that at most 30% of the data (or whatever percent are missing) are the result of such incorrect model specifications. There is only so much damage that modest model misspecifications can inflict, especially when contrasted with the potential bias in carrying out naive complete-record analyses.

Conclusion and Key Literature

Missing data occurs in many if not most applied data analysis settings and can introduce bias and inefficiency if not handled properly. Quantifying the extent and structure of missing data is key to choosing appropriate methods, and this should be done using the framework of MCAR, MAR, and MNAR missingness. Two popular, widely applicable, and flexible methods are weighting and MI. These are valid under MAR missingness. Not being discussed here, NMAR missingness is especially challenging but can be approached via sensitivity analyses in the model for the missingness mechanism.

Seminal works in this area include Rubin (1976, 1987), Little and Rubin (2002), and Robins *et al.* (1995). Useful references are Schafer (1997) and Schafer and Graham (2002). Other references drill deeper on the topics discussed here or discuss related work.

References

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd edn.). New York: John Wiley.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7**, 147–177.

Further Reading

Breslow, N. E. and Cain, K. C. (1998). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11–20.

Briggs, A., Clark, T., Wolstenholme, J. and Clarke, P. (2003). Missing…presumed at random: Cost-analysis of incomplete data. *Health Economics* **12**, 377–392.

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine* **23**, 1455–1497.

Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* **55**, 244–254.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

Perin, J., Preisser, J. S. and Rathouz, P. J. (2009). Semi-parametric efficient estimation for incomplete longitudinal binary data with application to smoking trends. *Journal of the American Statistical Association, Applications and Case Studies* **104**, 1373–1384.

Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine* **21**, 3035–3054.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society, Series B* **65**, 711–723.

Schildcrout, J. S. and Heagerty, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics* **9**, 735–749.

Schildcrout, J. S. and Rathouz, P. J. (2010). Longitudinal studies of binary response data following case-control and stratified case-control sampling: Design and analysis. *Biometrics* **66**, 365–373.

Modeling Cost and Expenditure for Healthcare

WG Manning, University of Chicago, Chicago, IL, USA

Introduction

The distributions of both healthcare expenditures and utilization share a number of characteristics that make their analysis more complicated than conventional economics outcomes; see also Models for count data and Cameron and Trivedi (2013). This article will focus on the analogous issues for costs and expenditures on healthcare. The most salient of these characteristics for many health economics applications are as follows: (1) a substantial part of the population will have no costs or expenditures during the period of observation; (2) among those with any expenditure, the amount is typically dramatically skewed right; and (3) not all parts of the distribution respond the same way to covariates. Even in clinical populations based on being treated for a specific health condition, the last two characteristics are still important. It is not uncommon to find that the top 1% of the population consumes nearly a fifth to a fourth of total health resources, and the top 10% nearly half of the total. Some types of expenditures are even more skewed with the top tenth of the distribution accounting for half or more of all healthcare expenditures, such as inpatient or mental health. These upper centiles or deciles may respond differently, such as being less elastic in their response to out-of-pocket price or income. In the Health Insurance Experiment (HIE), the upper decile was composed mostly of inpatient users, who were less responsive to insurance coverage than outpatient-only users (Duan *et al.*, 1983; Manning *et al.*, 1987).

Analysts have often found that the use of the least squares estimator with such data often leads to analytical problems from highly influential outliers – catastrophic cases, or chronic care patients with many visits and substantial pharmacy costs unless the sample size is sufficiently large so that there are a substantial number of such cases. For example, see the works on risk adjustment using a large fraction of the US Medicare claims files for risk adjustment. As the sample size becomes very large, the same concern does not remain, that the results will be driven by a small number of observations with catastrophic expenditures having undue influence.

Without the luxury of enormous data files, there is often the temptation to remove these extreme cases; removing or trimming or winsorizing such cases will reduce the influence of the extreme cases but will also introduce bias because catastrophic cases do occur in health and do consume real resources. Because of these issues, analysts often find that results are not replicable across alternative samples drawn from the same population. Least squares estimation under these circumstances is inefficient and in some cases biased, especially if zeroes are prevalent. Even when the estimates are unbiased, the inferences are biased because the underlying heteroscedasticity implicit in a model with a strictly nonnegative dependent variable is ignored. The inefficiency in the estimates and bias in the inference statistics for ordinary least squares (OLS) results arise from the property that the variability in costs is often an increasing function of the mean or some other function of the covariates x.

Health economists and econometricians have adapted various methods to deal with different outcomes and issues. The largest division is between count estimators to address the integer nature of MD visits and hospital admissions, compared to estimation approaches designed for continuous positive outcomes such as expenditures; survival approaches have been used, but are less common than the major splits of counts versus continuous. A cross-cutting issue is how to address the mass of observations at zero in either type of outcome, but the specific solutions differ both by the type of data on the outcome and by the underlying research question. In the case of count data, the concerns may include the mean response or marginal and incremental effects, or the response probability that the specific values of the counts respond in a particular way – does a change in the dental insurance coverage for prophylaxis change the probability that a patient has two such visits per year?

In the case of the limited dependent variable for expenditures, the issues are often the likelihood of any use and the overall mean response or some function of these such as the marginal and incremental effect. In both cases, the skewed nature of the outcome variables may make the results sensitive to influential outliers unless suitable methods are employed.

This article will focus on the health costs and expenditure case for individuals.

With data from either a general population or a population of users, it is not uncommon for cases in the top 1% of the distribution to have values that are nine times the sample mean, have studentized residuals that are beyond the $+4\sigma$ range, and maybe in double digits for the far right tail of the distribution, especially for expenditures. If such values are coupled with deviant values for the covariates, then they may have tremendous influence on the estimates. Because cost and expenditure data are so skewed right, there are no countervailing large residuals in the left tail. The consequence is that in small and moderate-sized samples, a single case can have tremendous influence on the estimates; as the sample size N increases, this issue of influence is diminished, especially with very large data sets, such as those used in estimating risk adjusters for the US Medicare program. The one notable exception occurs when one of the rate cells is relatively rare; see comments in Mihaylova *et al.* (2011). In the HIE, one single observation accounted for approximately 17% of the mean for that insurance plan.

The issues are different for very large data sets. As the sample size N increases, this issue of influence to skewness diminishes. The major modeling issue is getting the functional form to reflect the nonlinear nature of the response, an issue that applies to data sets of all sizes.

In what follows, first, a brief introduction to alternative approaches is provided for these types of healthcare or medical-care expenditures (or continuous positive outcomes).

Then questions about the treatment of the zero mass and skewness are addressed. There are some comments on methods for assessing differential responses in different parts of the distribution. Except where noted, the discussion applies to observations for fixed-size intervals of observations, rather than unequal-sized intervals; the latter can be addressed with formal offsets in some models.

Healthcare Expenditures

In what follows, a summary is provided of a number of issues that are of econometric and statistical concern in the modeling of healthcare expenditures. Jones (2011), Mihaylova *et al.* (2011), and Mullahy (2009) provided reviews of the modeling of healthcare expenditures and continuous outcomes with a more detailed discussion related to much of what is presented here.

Addressing the Zeroes Issue

Healthcare expenditures are nonnegative, often with many zeroes for the period of observation. One of the first issues is how to address the zeroes in the econometric modeling of the mean conditional on the covariates or of marginal effects when the focus of the analysis is on a general population rather than a clinical population of healthcare users. One approach has been to break down the distribution into two or more parts using the following rule:

$$E(y_i|x_i') = \text{Prob}(y_i > 0|x_i') \cdot E(\$_i|x_i', y_i > 0) \quad i = 1,\ldots,N \qquad [1]$$

The classes of two-part estimators provide different approaches to each part, with the two terms on the right side typically modeled separately because of the conditional independence of the second part (Cragg, 1967; Duan *et al.*, 1983). This type of estimators have been extended to include multipart models as well to address additional complication, such as differential response in the right tail, largely associated with inpatient care in the HIE (Duan *et al.*, 1983; Manning *et al.*, 1987). The analytical issues for both two-part and multipart models involve choices of the estimation approach for any expenditure and for the level of expenditure that are appropriate for the data at hand.

There is an alternative approach for the case where the concern is the mean response (conditional on a set of characteristics). That alternative involves a one-part part or single-equation, nonlinear model to obtain consistent estimates of $E(y|x)$, where the underlying explanatory variables are the same as those in the two-part models in equation 1 above. There is no necessity that the functional form of the response is the same as that for the second part of eqn [1]. For two-part models, see Duan *et al.* (1983), Blough *et al.* (1999), and various papers by X. H. Zhou. For single or one-part model, see Mullahy (1998) and Buntin and Zaslavksy (2004). There appear to be two unresolved issues in the debate over one- versus two-part models. The first is how large a fraction of the observations should be zeroes to make a difference, if any, in terms of bias or efficiency. The second is to what extent is this debate about the choice of the number of parts, rather than about how complicated the covariate specification should be to fit the actual distribution across the range of predictions in one- versus two-part models? Would a one-part model with additional covariates be able to capture curvature in the data to be equivalent to a two-part model?

There is a third, less common approach that builds on bivariate normal methods for two-part models, rather than the conditioning argument behind eqn [1], that are sometimes referred to as adjusted or generalized Tobit models or Heckit models. There is mixed evidence on how well these alternative estimators behave if there are no identifying restrictions across equations, which is the most common situation in health, unlike in labor economics. There are two common misconceptions in this debate. The first is that two-part models are nested within the bivariate normal alternatives; they are not. The second is that two-part models assume no correlation between any use and level of use; there are counterexamples in the literature.

All of the models considered here may be sensitive to influential outliers. The sensitivity to extreme cases is a natural byproduct of the skewness in the data. If expenditures were to be analyzed by a standard OLS model of the form $y_i = x_i'\beta + \varepsilon_i$ where y_i is the cost or expenditure on the raw scale (dollars, Euros, or pounds) for observation i, x_i' is a row vector of observed characteristics, and β is a column vector of coefficients to be estimated. Then the effect of an individual observation i on the estimate of β can be characterized by Cook's distance or the DFITS measure. Both measures depend on how extreme the observation is in terms of both the covariate values and the residual squared. These two diagnostics can be extended to nonlinear models as well as for least squares, as well as other tests of model checking such as Pregibon's Link Test or can be extended to include more complicated nonlinearity (as in Ramsey's RESET test), or less parametrically using a modified version of the Hosmer–Lemeshow test.

Addressing Skewed Positive or Overall Expenditures or Overall Nonnegative Expenditures

There are several approaches to dealing with such data: do nothing beyond OLS, use a Box–Cox transformation of the dependent variable, use one of the generalized linear models (GLMs) appropriate for continuous outcomes, use one of the three- and four-parameter distributions, or use a flexible and robust approximations to the underlying distribution. The consequence of the all-too-common alternative of ignoring the skewed data with a least squares approach is that the results are (1) sensitive to the skewness in the dependent variable, especially if the data set is of small or moderate size (such as the Medical Expenditure Panel Survey), and some of the characteristics are rare, and (2) the inference statistics are biased given the inherent heteroscedasticity in the data has not been captured in the estimation.

Box–Cox models

One alternative is to transform the dependent variable by the natural logarithm or a power transformation to eliminate the skewness in the error; transformations may also be used to achieve a model that is linear in the parameters (as in the

Cobb–Douglas production function) or to stabilize the variance of the equation estimated (as in the square root transformation for count data or the inverse sine root transform for proportions). Some of these transformations are special cases of the Box–Cox transformation, but some are less reliant on parametric distributional assumptions. Specifically, the Box–Cox models consider transformed equations $f(y)$ if y is positive of the form

$$f(y) = (y^{\lambda} - 1)/\lambda = x\beta + \varepsilon \quad \text{if } \lambda \neq 0 \qquad [1a]$$

$$f(y) = \log(y) = x\beta + \varepsilon \quad \text{if } \lambda = 0 \qquad [1b]$$

It is often assumed that ε is either symmetric or normally distributed. The model can be estimated either by maximum likelihood estimation (MLE) or by the use of least squares for suitable values of λ. There are also two-parameter versions of the Box–Cox model that allow explicitly for the mass of observations at zeroes, but these do not always provide consistent estimates of the mean outcome, conditional on the covariates.

The advantage of the power transformation if $\lambda < 1$ for data that are skewed to the right is that it pulls in the right tail of the distribution faster than it does in the middle or the left tail. As λ decreases toward zero, the error term in the estimated equation should become more symmetric, reducing the influence of the extreme cases in the right tail of the distribution. However, too low a values of λ (such as the log, when $0 < \lambda < 1$ would be more suitable) may lead to overcorrection in the sense that it would convert a right-skewed distribution into a left-skewed one after the transformation. The log transformation is not always the optimal choice for right-skewed data.

The problem with the log and Box–Cox approaches is that one is often not interested in the transformed scale *per se* – the government does not spend log dollars or log Euros. Rather, one is actually interested in the raw scale of expenditures y and predictions or marginal effects in dollars or Euros (in $E(y|x)$ more generally). This leads to concerns about the retransformation of the results from the scale of estimation (e.g., the log or the λth power) to the scale of interest (e.g., raw or actual dollars or Euros). Because of the nonlinear nature of the log and Box–Cox transformations, the transformation cannot be simply inverted to obtain unbiased estimates of the $E(y|x)$ because $E(f(y|x)) \neq f(E(y|x))$, where $f(y)$ is the transformation. This is the retransformation problem discussed by Duan (1983), Duan *et al.* (1983), Manning (1998), Mullahy (1998), and Blough *et al.* (1999).

The difference between the two is easy to see in the case of OLS on $\ln(y)$ if the error term is log normally distributed. OLS generates unbiased estimates of $E(\ln(y_i|x_i)) = x_i'\beta$ if $E(X'\varepsilon) = 0$. However, the term $E(\exp(x_{i_i}'\beta))$ yields an estimate of the geometric mean, not the arithmetic mean of the response function. The arithmetic mean is $E(y_i|x_i) = \exp(x_i'\beta + 0.5\sigma_{\varepsilon}^2)$ if ε is i.i.d. and normally distributed, and $E(y_i|x_i) = \exp(x_i'\beta + 0.5\sigma^2(x))$ if normally distributed and heteroscedastic in x (Manning, 1998). If the error is not normally distributed, then the estimates $x_i'\hat{\beta}$ may be consistent, but one can apply Duan's (1983) multiplicative smearing factor in the homoscedastic case to provide a consistent estimate of $E(\exp(\varepsilon_i))$, or its analog in the heteroscedastic case to obtain the mean response.

The goal of most analyses is some statement about how the mean or some function of the mean of y, such as the marginal effect on the raw scale, changes with x. In general, the expectation of y depends on the variance and heteroscedasticity on the log scale and on how higher order terms depend on the covariates. If the variable of interest x_j is not discrete, then the slope of the expected value with respect to the jth covariate is given by

$$\begin{aligned} \frac{\partial E(y_i|x_i)}{\partial x_{ij}} &= \left(E(y_i|x_i)\right) \cdot \left(\beta_j + 0.5\frac{\partial \sigma_{\varepsilon}^2}{\partial x_{ij}}\right) \\ \frac{\partial E(y_i|x_i)}{\partial x_{ij}} &\neq \left(E(y_i|x_i)\right) \cdot \beta_j \\ \frac{\partial E(y_i|x_i)}{\partial x_{ij}} &\neq \left(e^{x_i'\beta}\right) \cdot \beta_j \end{aligned} \qquad [2]$$

It is the first derivative in eqn [2] that should be used in the calculation of the elasticity of the mean response or the average marginal effect, rather than either of the other two in the case of log scale heteroscedasticity if the log-scale error is normally distributed. The last one applies only if $R^2 = 100\%$ on the log scale.

In the nonnormal case or for values of λ other than zero, the derivative will depend on the power transform λ, the nature of the distribution in the absence of the additional complication of heteroscedasticity:

$$\frac{\partial E(y_i|x_i)}{\partial x_{ij}} = \beta_j \int (\lambda(x_i'\beta + \varepsilon_i) + 1)^{(1-\lambda)/\lambda} dF(\varepsilon_i) \qquad [3]$$

where $F(\varepsilon)$ is the cdf for the error term.

In the square root case where $\sqrt{y_i} = x_i'\beta + \varepsilon$ with possible heteroscedasticity, the scale of interest relationship is $E(y_i|x_i) = (x_i'\beta)^2 + \sigma_{\varepsilon}^2(x_i)$. Here the retransformation factor is additive. The retransformation factor will be multiplicative only in the case of $\lambda = 0$ and will be moot if $\lambda = 1$.

There are a number of technical issues that arise with Box–Cox models. One of these is how to deal with observations where $y = 0$. Second, the estimates of the power transform are sensitive to extreme outliers in ε. In practice, it may be difficult to tell the effect of an influential outlier from skewness in the dependent measure that is not associated with the covariates. Third, if λ is not known *a priori*, then all of the inferences should reflect that β's, λ, and σ are estimated in calculating inference statistics for eqns [2] and [3], not just the β's.

Generalized linear models

A second alternative to a least squares linear model using some $f(y)$ directly is to model the $f(E(y)$ directly and deal with the skewed expenditure data (with or without the zeroes) by addressing the property that the variance function is often an increasing function of the mean. This can be done by using some iteratively reweighted least squares alternative or the GLMs for continuous outcomes (such as gamma regression) estimated with quasi-maximum likelihood methods. In the GLM case, the analyst specifies a link function between the linear model $x'\beta$ and the mean, so that $g(E(y|x)) = x'\beta$, and a variance function $v(y|x)$ that characterizes the nature of the relationship between the mean and the variance on the raw

scale. The $v(y|x)$ function is assumed to be a function of the mean, not of individual covariates in x directly. The correct specification of the variance function results in more efficient estimators and may correspond to an underlying distribution of the outcome measure. If the distribution is from the exponential family, then the estimation can be done by quasi-maximum likelihood methods. Although one can perform an MLE using the specific distributions, the conventional GLM is more robust in the sense it does not assume the distribution beyond the first and second moments. So, for the gamma GLM, only the gamma variance (the variance function increases as the square of the mean function) is assumed and not the full gamma distribution.

If the link function is misspecified, then the estimates will provide a biased estimate of the response. Much of the work to date on healthcare expenditures has used a log link: $\ln\left(E\left(y_i|x_i\right)\right) = x_i'\delta$; I use δ has been used for the index function in the log link GLM to avoid possible confusion with β from the $\ln(y)$ model. If the $\ln(y)$ error is homoscedastic, they have the same expectation, except for the intercept where δ_0 corresponds in expectation to $\beta_0 + \ln[E(\exp(\varepsilon_i))]$. However, some papers have also used power transformations, such as the square root. Just because the log transformation is often used in transformed y models, there is no reason to assume the same for the GLM or to use $\lambda = 0$ for all cases that are skewed right. In the transformed y models, the log (or any other Box–Cox transformation) is designed to achieve symmetry in the error. In the GLM case, the goal is to find a function of a linear index (say $x'\delta$, which provides a consistent estimate of $E(y|x)$ over the range of $x_i'\delta$ and the major covariates in $x_i'\delta$. This difference between least squares on transformed y and GLMs with power or log links is very important to understand, and has been a common source of confusion between log/Box–Cox methods and GLM alternatives.

Since the late 1990s, the most commonly used distribution for GLM applied to positive healthcare cost or expenditure data has been the gamma. This is appealing working assumption for the distribution function because the standard deviation under the gamma is proportional to the mean, a property often but not always exhibited by healthcare cost data. However, this is only one of several cases where the standard deviation or the variance is a power of the mean function. If the variance is not a function of the mean, then a Gaussian assumption may be used (Mullahy, 1998; Wooldridge, 1992). If the variance is proportional to the mean, then the Poisson may be more appropriate. If the standard deviation is proportional to the cube of the mean, then the inverse Gaussian is an alternative. Other applications could employ powers of the variance as in Blough *et al.* (1999) and Basu and Rathouz (2005) approach (see below).

As long as the link function and the index function of the covariates $x_i'\delta$ are correctly specified, the GLM provides consistent estimates. The wrong mean–variance relationship or the wrong distribution function can potentially lead to substantial efficiency losses. For Box–Cox models, there may be bias if there is heteroscedasticity and efficiency loss if the underlying error is not normal. Manning and Mullahy (2001) provide a fuller discussion and simulation results that illustrate the trade-offs involved among $\log(y)$ and GLM with log links. One implication of their work is that there is no one estimation approach that is ideal or even a close second best approach for all examples. The best estimation approach depends on the application at hand and its underlying data-generating process.

Extended GLM methods

There is a hybrid of the Box–Cox type of model and the GLM. Basu and Rathouz (2005) describe an extended estimating equation (EEE) algorithm that allows one to use the data to estimate both the link function from the power family via the Box–Cox link for the mean function and the power family relationship between the mean and the variance functions.

$$\begin{aligned} E(y|x) &= \mu = g^{-1}(x\beta) \\ g(\mu_i) &= (\mu_i^{\lambda} - 1)/\lambda \\ V(y_i) &= \theta_1(\mu)^{\theta_2} \end{aligned} \qquad [4]$$

This approach is more general than the two preceding alternatives and avoids the issues that arise in the common practice of using only a discrete set of GLM alternatives for the link and mean–variance relationship. In both the Box–Cox model and the GLM, the choice of the wrong transform of y or link for $E(y|x)$ can lead to biased estimates. In the GLM, only using integer powers for the mean–variance relationship can lead to a substantial loss of efficiency. Further, the results for inference statistics will reflect the uncertainty in the estimates of λ and θ_2, thus avoiding the corresponding issue and debate in statistics over the Box–Cox transformation of y.

Other parametric approaches

Manning *et al.* (2005) propose using an exponential conditional mean regression based on the three-parameter generalized gamma distribution that could be estimated by maximum likelihood. The generalized gamma model includes the gamma, Weibull, and exponential distributions with log link, as well as models for $\ln(y)$ with normal errors. This approach also provides a robust alternative to either the GLM with log link or the OLS on $\ln(y)$ when those two alternatives do not apply, assuming that the mean is truly an exponential function of $x'\beta$. The generalized gamma is a more precise alternative than the GLM when the distribution is more skewed than is implicit in the GLM case. However, the generalized gamma is susceptible to bias in the presence of certain forms of heteroscedasticity on the log scale. Manning *et al.* (2005) propose a modification of the model that corrects for this by allowing for two index functions, one for the $\ln(\sigma)$ term and the other for the log-scale mean.

There is some preliminary interest in four-parameter distributions, such as the generalized betas of the second kind because they allow for better fit to the actual distribution of positive expenditures, as well as have several of the other alternatives as special cases if the link is log or there is a proportional response. There is related work by the distribution of income. The five-parameter distribution has not been employed to the best of the author's knowledge. Nevertheless, the generalized gamma and the four- and five-parameter versions of the generalized beta permit an explicit allowance for skewness that is not always fully captured in simpler two-parameter distributions such as the Box–Cox/log normal or the GLM with its focus on the first two moments. Jones (2011)

reports results from a parametric model in the generalized beta of the second class. Although that model allows only one of the parameters to be a function of covariates, Jones has work allowing two or more parameters being a function of covariates. This work would be more general and flexible than either the GLM (with log link) or the generalized gamma, because these are limiting cases or more restrictive variants of the generalized beta of the second kind (GB2). That statistical distribution has a richer parameterization that allows the parameters to depend on patient or other characteristics.

Differential responsiveness

In their common formulations, the previous methods do not allow for heterogeneity in the healthcare responses to covariates over the population, by service, over the distribution of expenditures, or by allowing for latent groups. Quantile methods could be used to allow for differential responses, if allowances are made for ties in the zero spenders. The multipart models, especially the four-part model used in the HIE (Duan *et al.*, 1983) allow for inpatient users to have a different response to covariates; the four-part model is also a mixture model with known or observable separation among subgroups and differs from the latent class models more often used for count data. Finite mixture models with unobserved or unknown separation can be used to approximate an arbitrary distribution and to allow for some heterogeneity in response, given the number of latent classes included.

Less-parametric approaches

Gilleskie and Mroz (2004) have suggested that one can use a series of conditional models to address the skewed nature of healthcare expenditures and the zeroes problems together. They suggest a conditional density estimator (CDE) that breaks the dependent variable (healthcare expenditures) into J different segments, modeling the probability p of being in a specific segment j as a function of the covariates x's as a polynomial function f of the covariates x, and then using subsample means of y within each of those J segments. The basic approach takes advantage of the way conditional distributions work, namely, that $E(y) = E(y|z) \cdot E(z)$. In this case, the expected value of y, given the observed characteristics of the population is

$$E(y|x) = \sum_j p_j(y \text{ in range } j) \cdot (E(y|y \text{ in range } j)) \qquad [5]$$

The overall response $E(y|x)$ for person i over the j ranges is given by

$$E(y_i|x_i) = \sum_j \left((p_j(x_i'\alpha_j)) \cdot (\mu_j) \right) \qquad [6]$$

where j is an index for segments and x's are polynomials in the underlying independent variables. Gilleskie and Mroz propose a very specific form for the probability functions $p_j(x_i'\alpha_j)$, but one can use a more general approach than they used. By breaking the dependent variables into bounded values (except for the last segment), they avoid some of the issues of robustness to skewness in y because the values of y in a specified range are not as long tailed as the whole distribution of y. By using a polynomial in the underlying covariates, they allow for a nonlinear response to the individual characteristics.

Given the complexity of the model in eqns [4] and [5], the effects of covariates are assessed using a marginal or incremental effects approach just as in the multipart models.

They find that the model performs well in a range of simulated conditions. Also, they are able to obtain well-behaved results with data drawn from the HIE. One of two remaining issues is how big the intervals should be, especially given the substantial part of the overall expenditures that are in the last (open) interval. The other is how to model means of intervals if the mean conditional on being in an interval depends on covariates.

Assessing Model Fit

The literature provides a number of tests that can be applied to most of these models to assess the quality of the fit whether the model is based on single-equation methods (with or without the zeroes), two-part or multipart models, or GLM and extended GLM. Some of the omnibus fits of test are primarily done on the scale of estimation; they include Pregibon's Link Test and Ramsey's RESET test. The Pearson correlation test between the raw-scale residual and the raw-scale prediction, and the modified Hosmer–Lemeshow test, can be performed on the raw scale, which is often the scale of interest. Buntin and Zaslavsky (2004) provide discussions of a number of these, plus others that are often used in the risk adjustment literature.

Generally, the two-part and multipart models are more difficult to assess the overall fit because there is no established analog of Pregibon's Link or Ramsey's RESET test for multipart models. For this and the CDE class of models, model assessment is limited to Pearson and modified Hosmer–Lemeshow for in-sample assessment, and the usual cross-sample validation approaches including simple split sample or k-fold methods. There are other extensions of split sample cross-validation approaches that can be used on the scale of estimation, including the type of two-parameter more parsimonious work by Copas (1997). In health services research, there is also a set of cross-validation tests that have been called Copas tests but differ in that the estimates from the first split sample are predicted to the test or validation sample on the raw scale, sometimes called the scale of interest. There the test is whether the regression of the raw scale version of the dependent variable to the test but on the raw scale is a straight line through the origin with slope 1; see Veazie *et al.* (2003) and Basu *et al.* (2006).

Quantile Approaches

With the exception of the Gilleskie and Mroz (2004) approach discussed earlier, the approaches have been largely parametric. Often these approaches have assumed that there is limited heterogeneity in response over the distribution of expenditures (inpatient vs. outpatient, or across a small number of latent classes). Another approach is to employ quantile regression methods that allow the responses to differ across the distribution of expenditures, conditional on a set of covariates based on the quantile methods reviewed in Koenker (2005).

Note that these provide a more flexible approach to modeling the response because the responses are not forced to be parallel on the scale of estimation. For example, if the inpatient and outpatient responses are different, then that can be addressed with separate models for different services or by a multipart model. But if the catastrophic inpatient cases had a different response to income and price, then the simpler models could fail to be parallel in the right tail, where so much of the total resource cost and expenditure are.

This article will not address those quantile approaches in detail.

Strengths and Weaknesses

There have been a substantial number of papers of econometric and statistical models for modeling healthcare costs and expenditures as a function of patient characteristics of interest. Many of these deal with the whole distribution, especially the substantial fraction of the cases that have zero expenditure. Many of these papers involve comparisons for alternative models with the evaluation largely limited to how well specific models do for a specific data set. One of the concerns about such studies is their generalizability to other populations or to other types of healthcare expenditures. A second concern is that the performance in a specific application may reflect overfitting of badly skewed data, because many of the papers use within-estimation sample methods to evaluate the relative performance of the alternatives.

Manning and Mullahy (2001) report simulated comparisons of several exponential conditional mean models under a range of different data-generating mechanisms. These include the use of OLS on $\ln(y)$, various types of GLMs with log links (Gaussian, Poisson, and gamma). In each case, the true response was exponential conditional mean $E(y|x) = \exp(x\beta)$. They examined a number of data-generating mechanisms that lead to varying degrees of skewness, heteroscedasticity on the log scale, and even heavy-tailed distributions for $\ln(y)$. In the absence of heteroscedasticity on the log scale, they found that both the OLS on $\ln(y)$ and GLM with log link models provided consistent estimates of $E(y|x)$. But if the true model was linear on the log scale with an additive error term that was heteroscedastic in x, the log OLS provided biased estimates of $E(y|x)$ without suitable retransformation.

Although the GLM were always consistent, the choice of variance function assumption had substantial impact on the efficiency of the estimation, especially when compared with the results based on OLS on $\ln(y)$. Further, the loss of efficiency relative to OLS on $\ln(y)$ increased as the data became more skewed or became heavy tailed on the log scale.

Because of the potential for bias from OLS on $\ln(y)$ and of efficiency losses from the GLM with log link models when the log-scale error variances are large or the log-scale error is heavy tailed, they propose an approach for determining which estimation approach is better for a specific data set.

Basu and Rathouz (2005) also simulate the behavior of various GLM versus their proposed EEE extension of the GLM. They find evidence of bias and inefficiency when the incorrect power transformations are used for either or both of the link and the variance functions. But they do not examine the traditional Box–Cox estimator.

Except for Basu and Rathouz (2005) and subsequently Basu *et al.* (2006), there is little in the literature outside the exponential conditional mean that is not specific to a particular data set or health condition. But there is no reason why other link or Box–Cox style transformations of the dependent variable could not be used or seriously considered. More flexible alternatives could be used and then the results reported as incremental or marginal effects.

With data as skewed or with as many zeroes as healthcare data have, there is always a concern about overfitting. Traditional split-sample or cross-validation tests, or the more parsimonious Copas's (1997) tests on the scale of estimation can be employed to assess overfitting in the narrower sense of the term. But there is also a concern about how well the model fits on the scale of ultimate interest, especially if payment/risk adjustment issues are involved. Some of these have extended to split-sample, out-of-sample tests, and to test/validation sample methods on the scale of interest by Veazie *et al.* (2003) and Basu *et al.* (2006). These provide more confidence in the results than ones conducted on the statistical scale-of-estimation, as Copas and others have done; see Hill and Miller (2010) for a recent example of a scale-of-interest comparison.

Conclusions

Because of the very skewed nature of healthcare costs and expenditures, analysis based on simple regressions of costs or expenditures are not robust in data sets with the number of observations encountered by most analysts. The focus in most applications is on finding more robust estimates of the mean response than simple OLS, conditional on the covariates, or the marginal and incremental effects of particular policies or treatments. The literature offers a number of alternative estimation strategies for expenditures. These include both single and multipart models using a range of options to deal with skewness in general or in the positive cases: Box–Cox transformations (especially the log) of the dependent variable, and GLMs, and a less restrictive version of the GLM-type approach (the extended GLM or EEE), a broader class of distributional assumptions (the generalized gamma and the generalized beta), a discrete and less parametric approximation, have been suggested as alternative estimators.

At this point, it does not appear that any specific approach dominates the modeling of the data with continuous outcomes beyond the nonuser subset. Instead, it appears that econometric model needs to be able to address the research question and to match the characteristics of the data if the estimate is to be relatively efficient, with little bias, and to pass the Cox test. To paraphrase Cox and Draper, all models are wrong, but some are useful. That is, some econometric methods provide better approximations than others.

Acknowledgments

This article has benefited from the support of the Harris School of Public Policy Studies at the University of Chicago

and the thoughtful comments from colleagues and from the associate editors Anirban Basu and John Mullahy.

See also: Models for Count Data

References

Basu, A., Arondekar, B. V. and Rathouz., P. J. (2006). Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of comorbidity. *Health Economics* **15**(10), 1091–1107.

Basu, A. and Rathouz, P. (2005). Using flexible link and variance models. *Biostatistics* **6**, 93–109.

Blough, D. K., Madden, C. W. and Hornbrook., M. C. (1999). Modeling risk using generalized linear models. *Journal of Health Economics* **18**, 153–171.

Buntin, M. B. and Zaslavsky, A. M. (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics* **23**, 525–542.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data* (2nd edition). Cambridge: Cambridge University Press.

Copas, J. B. (1997). Using regression models for prediction: Shrinkage and regression to the mean. *Statistical Methods in Medical Research* **6**(2), 167–183.

Cragg, J. (1967). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **35**, 829–844.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* **78**, 605–610.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics* **1**, 115–126.

Gilleskie, D. B. and Mroz, T. A. (2004). A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**, 391–418.

Hill, S. C. and Miller, G. E. (2010). Health expenditure estimation and functional form: Applications of the generalized gamma and extended estimating equations models. *Health Economics* **19**(5), 608–627.

Jones, A. M. (2011). Models for health care. In Clements, M. and Hendry, D. (eds.) *Handbook of economic forecasting*, pp. 625–634. Oxford: Oxford University Press.

Koenker, R. (2005). *Quantile regression (Econometric Society Monographs)*. Cambridge: Cambridge University Press.

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* **17**, 283–295.

Manning, W. G., Basu, A. and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**(3), 465–488.

Manning, W. G. and Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics* **20**(4), 461–494.

Manning, W. G., Newhouse, J. P., Duan, N., et al. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review* **77**(3), 251–277.

Mihaylova, B., Briggs, A., O'Hagan, A. and Thompson, S. G. (2011). Review of statistical methods for analyzing healthcare resources and costs. *Health Economics* **20**(8), 897–916.

Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* **17**, 247–281.

Mullahy, J. (2009). Econometric modeling of health care costs and expenditures: A survey of analytical issues and related policy considerations. *Medical Care* **47**(7 supplement 1), S104–S108.

Veazie, P. J., Manning, W. G. and Kane, R. L. (2003). Improving risk adjustment for Medicare capitated reimbursement using nonlinear models. *Medical Care* **41**(6), 741–752.

Wooldridge, J. M. (1992). Some alternatives to the Box–Cox regression model. *International Economic Review* **33**, 935–955.

Further Reading

Basu, A. (2005). Extended generalized linear models: Simultaneous estimation of flexible link and variance functions. *The Stata Journal* **5**(4), 501–516.

Models for Count Data

PK Trivedi, Indiana University, Bloomington, IN, USA

Introduction

Count data regression is now a well-established tool in econometrics. If the outcome variable is measured as a nonnegative count, y, $y \in N_0 = 0, 1, 2, \ldots$ and the object of interest is the marginal impact of a change in the variable x on the regression function $E[y|x]$, then a count regression is a relevant tool of analysis. Fully parametric formulations of count models accommodate this discreteness property of the distribution, with probability mass at nonnegative integer values only. By contrast, some semiparametric regression models accommodate only nonnegativity but not discreteness. For such data, a linear regression is generally not efficient, and hence the standard count model is a nonlinear regression. A fully parametric formulation is more attractive if the researcher's interest is in the full distribution of the data, whereas a semiparametric model like nonlinear least squares is often used when the focus is on the conditional mean only.

Cross-section and panel data on event counts, for example, doctor visits, hospital admissions, and many measures of health-care utilization, are very common in empirical health economics. This has contributed to the use of count data regressions. This entry covers the case where all regressors are exogenous or predetermined. The empirically important case in which some regressors are endogenous is treated in other articles. The focus here is on classical inference, but all models considered here are amenable to Bayesian analysis.

Poisson Regression

The starting point of many count data analyses is the Poisson regression, derived from the Poisson distribution, for the number of occurrences of the event, with probability mass function

$$\Pr[Y = y] = f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots, \qquad [1]$$

where μ is the intensity or rate parameter. The first two moments of this distribution, denoted $P[\mu]$, are $E[Y] = \mu$, and $V[Y] = \mu$, which is the well-known equidispersion property of the Poisson distribution. The Poisson regression results from the parameterization $\mu = \mu(x)$, where x is a K-dimensional vector of exogenous regressors. The usual specification of the conditional mean is

$$E[y|x] = \exp(x'\beta) \qquad [2]$$

Standard estimation methods are fully parametric Poisson maximum likelihood (PML), 'semiparametric' methods such as nonlinear least squares, or (weighted) moment-based estimation, based on the moment condition $E[y - \exp(x'\beta)|x] = 0$ possibly further augmented by the equidispersion restriction used to generate a weighted moment function.

According to standard maximum likelihood theory, if the Poisson model is parametrically correctly specified, the maximum likelihood estimator (MLE) $\hat{\beta}_p$ is consistent for β, with covariance matrix estimated by

$$\hat{V}[\hat{\beta}_P] = \left(\sum_{i=1}^{N} \hat{\mu}_i x_i x_i'\right)^{-1} \qquad [3]$$

where $\hat{\mu}_i = \exp(x_i'\hat{\beta}_P)$. The use of this formula can be misleading if the equidispersion assumption is incorrect.

The Poisson regression is founded in the Poisson point process for the occurrence of the event of interest. This process is a characterization of complete randomness, which excludes any form of dependence between events, either cross sectionally or over time, and any form of nonstationarity. In the Poisson regression, these assumptions are conditional on the covariates x_i, which reduces the restrictiveness of the model. Even when analysis is restricted to cross-section data with strictly exogenous regressors, the basic Poisson regression is restrictive for most empirical work. First, the mean-variance equality restriction will be violated if there is significant unobserved heterogeneity in cross-section data – in which case (conditional) variance will exceed (conditional) mean. This feature, referred to as overdispersion (relative to the Poisson), manifests itself in a variety of data features, most noteworthy being the excess zeros problem.

Overdispersion

Overdispersion results from many different sources and is consistent with a variety of different deficiencies of the Poisson regression, including serial dependence, spatial dependence, or contemporaneous dependence of events. This motivates replacing the Poisson distribution with more flexible functional forms that can accommodate overdispersion as well as its other specific limitations.Thus, many functional forms are generated as Poisson mixtures. Replace the parameter μ_i in (1) by $\mu_i\nu_i$, where ν_i represents individual-specific, independently and continuously distributed, separable unobserved heterogeneity. Next make an assumption about the distribution of ν_i Finally, derive a new functional form by integrating out ν (subscript omitted) – a mathematical operation equivalent to averaging with respect to the assumed distribution of ν. A shortcut is to start with the more flexible functional form (i.e., the mixture distribution) without going through the intermediate mathematical step.

Provided the conditional mean is correctly specified, the PML estimator (PMLE) is consistent but not efficient in the presence of overdispersion. Considering overdispersion, one can use the pseudo-ML or quasi-ML approach, again using the PMLE for point estimates but computing the (Eicker–White) robust estimate of the variance–covariance matrix using an expression of the form $\hat{V}_{Rob}[\hat{\beta}_P] = A^{-1}BA^{-1}$, where $A = \sum_{i=1}^{N} \hat{\mu}_i x_i x_i'$ and $B = \sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2 x_i x_i'$ and

 doi:10.1016/B978-0-12-375678-7.00716-1

$\hat{\mu}_i = \exp(x_i'\hat{\beta}_P)$, that is, the point estimates of β are as in pseudo-ML theory, but its sample variance is obtained robustly.

Efficient estimation of overdispersed model is possible if more specific parametric assumptions are invoked. The negative binomial (NB) regression is an example of a mixture model. It can be derived as a Poisson–Gamma mixture. Given the Poisson distribution $f(y|x,v) = \exp(-\mu v)(\mu v)^y/y!$ with the mean $E[y|x,v] = \mu(x)v$, $v > 0$, where the random variable v represents multiplicative unobserved heterogeneity, a latent variable assumed to be independent and separable has Gamma density $g(v) = v^{\alpha-1}\exp(-v)/\Gamma(\alpha)$, with $E[v] = 1$ and variance α ($\alpha > 0$). The resulting mixture distribution is the NB2 (negative binomial with quadratic variance), which has mean $E[y|x] = \mu(x)]$ and variance $V[y|x] = [1 + \alpha\mu(x)]\mu(x) > E[y|x]$, thus accommodating overdispersion. Relative to the Poisson, the overdispersed distributions have more probability mass at zero and high values of y. The same approach can be used with other mixing distributions.The NB1 variant of the NB, which has variance linear in $\mu(x)$, and the Poisson–lognormal mixture are two popular alternatives to the Poisson. The formulae for the probability mass function of NB1 and NB2 distributions are shown in Table 1.

Test of overdispersion

The null hypothesis of equidispersion can be tested by postulating an alternative overdispersed model. A formal test of the $V[y|x] = E[y|x]$ property can be based on the equation

$$V[y|x] = E[y|x] + \alpha E[y|x]^2$$

which is the variance function for the NB1 model. Equivalently, $H_0{:}\alpha = 0$ against $H_1{:}\alpha > 0$ is tested.

The test can be implemented by auxiliary regression of the generated dependent variable $((y - \hat{\mu})^2 - y)/\hat{\mu}$ on $\hat{\mu}$, without an intercept term, and performing a t test of whether the coefficient of $\hat{\mu}$ is 0. However, the rejection of the null does not automatically indicate a suitable alternative model because that outcome is consistent with failure of the null model in many different ways.

Alternatives to the Poisson

The NB regression has been found to fit well many types of data, including those with 'excess zeros.' It has an analytical closed form. Functional forms resulting from other mixing assumptions, for example, the lognormal, often do not have a closed form, although estimation using either simulation-based methods or quadrature methods for numerical integration is straightforward to implement.

The classic gamma heterogeneity assumption underlying NB2 is somewhat special. Modern approaches, however, can handle more flexible models where the latent variables are nonseparable, which means that in principle unobserved heterogeneity impacts the entire distribution of the outcome of interest. Quantile regression and finite mixtures are two examples of such nonseparable models.

The literature on new functional forms to handle overdispersion is large and still growing. Despite the availability of many functional forms, a relatively small class of models has attained much popularity in health econometrics. This includes especially the NB regression (NBR) presented above, the two-part model (TPM) or the hurdle model, and the zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB), which collectively dominate the applied literature and form the set of basic parametric count regression models. These models are all mixtures of Poisson-type models. Discussion of these models provided as follows.

Hurdle Model

The hurdle model or TPM relaxes the assumption that the zeros and the positives come from the same data generating process. The zeros are determined by the density $f_1(\cdot)$, so that $\Pr[y=0] = f_1(0)$ and $\Pr[y>0] = 1 - f_1(0)$. The positive counts come from the truncated density $f_2(y|y>0) = f_2(y)/(1 - f_2(0))$, that is multiplied by $\Pr[y>0]$ to ensure that probabilities sum to 1. Thus, suppressing regressors for notational simplicity,

$$f(y) = \begin{cases} f_1(0) & \text{if } y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y \geq 1 \end{cases} \qquad [4]$$

Table 1 Selected mixture models

	Distribution	$f(y) = Pr[Y = y]$	*Mean; Variance*
1	Poisson	$e^{-\mu}\mu^y/y!$	$\mu(x)$; $\mu(x)$
2	NB1	As in NB2 below with α^{-1} replaced by $\alpha^{-1}\mu$	$\mu(x)$; $(1+\alpha)\,\mu(x)$
3	NB2	$\dfrac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\frac{1}{\alpha}}\left(\dfrac{\mu}{\mu+\alpha^{-1}}\right)^{y}$	$\mu(x)$; $(1+\alpha\,\mu(x))\,\mu(x)$
4	Hurdle	$\begin{cases} f_1(0) & \text{if } y = 0, \\ \dfrac{1-f_1(0)}{1-f_2(0)} f_2(y) & \text{if } y \geq 1 \end{cases}$	$\Pr[y>0\|x]E_{y>0}[y\|y>0,x]$
5			$\Pr[y>0\|x]V_{y>0}[y\|y>0,x]$ $+\Pr[y=0\|x]E_{y>0}[y\|y>0\|x]$
6	Zero inflated	$\begin{cases} f_1(0)+(1-f_1(0))f_2(0) & \text{if } y = 0, \\ (1-f_1(0))f_2(y) & \text{if } y \geq 1 \end{cases}$	$(1-f_1(0))(\mu(x)+f_1(0)\mu^2(x))$
7	Finite mixture	$\sum_{j=1}^{2}\pi_j f_j(y\|\theta_j)$	$\Sigma_{i=1}^{2}\pi_i\mu_i(x)$; $\Sigma_{i=1}^{2}\pi_i[\mu_i(x)+\mu_i^2(x)]$

This specializes to the standard model only if $f_1(\cdot)=f_2(\cdot)$. This model can handle both excess zeros and too few zeros.

A hurdle model has the interpretation that it reflects a two-stage decision-making process, each part being a model of one decision. The two parts of the model are functionally independent. Therefore, ML estimation of the hurdle model can be achieved by separately maximizing the two terms in the likelihood: one corresponding to the zeros and the other to the positives. A binary outcome model is used to model the positive outcome and a truncated Poisson or NB for the second part. The first part uses the full sample, but the second part uses only the positive-count observations.

For certain types of activities, such a specification is easy to rationalize. For example, in a model that explains the number of packs of cigarettes smoked per day, the survey may include both smokers and nonsmokers. The first part of the hurdle model determines whether or not one smokes and the second part determines the intensity, i.e., the number of packs smoked, given that at least one pack is smoked.

Models for Zero-Inflated Data

The zero-inflated model was originally proposed in manufacturing quality control settings to handle data with excess zeros relative to the Poisson. Like the hurdle model, it supplements a count density $f_2(\cdot)$ with a binary process with density $f_1(\cdot)$. If the binary process takes value 0, with probability $f_1(0)$ then $y=0$. If the binary process takes value 1, with probability $f_1(1)$, then y takes count values $0,1,2,\ldots$ from the count density $f_2(\cdot)$. This lets zero counts occur in two ways: as a realization of the binary process and as a realization of the count process when the binary random variable takes value 1.

Suppressing regressors for notational simplicity, the zero-inflated model has density

$$f(y)=\begin{cases} f_1(0)+(1-f_1(0))f_2(0) & \text{if } y=0 \\ (1-f_1(0))f_2(y) & \text{if } y\geq 1 \end{cases} \quad [5]$$

As in the case of the hurdle model, the probability $f_1(0)$ may be a constant or may be parametrized through a binomial model like the logit or probit to capture dependence on observable factors. Once again, the set of variables in the $f_1(\cdot)$ density need not be the same as those in the $f_2(\cdot)$ density. However, identifying the separate roles of factors that affect $f_1(0)$ and $f_2(y)$ is generally challenging.

Finite Mixture Models

The NB model is an example of a continuous mixture model. An alternative approach uses a discrete representation of unobserved heterogeneity to generate a class of models called finite mixture models (FMM) – a particular subclass of latent class models.

An FMM specifies that the density of y is a linear combination of m different densities, not necessarily from the same parametric class, where the *jth* density is $f_j(y|\beta_j)$, $j=1, 2, \ldots, m$. Thus an m-component finite mixture is

$$f(y|\beta,\pi)=\sum_{j=1}^{m}\pi_j f_j(y|\beta_j), \quad 0\leq\pi_j\leq 1, \quad \sum_{j=1}^{m}\pi_j=1 \quad [6]$$

For example, a two-component ($m=2$) mixture of $f_1(y|x,\beta_1)$ and $f_1(y|x,\beta_2)$ may reflect the possibility that the sampled population contains two 'types' of cases, whose y outcomes are characterized by distributions with different moments. The mixing fraction π_1 is, in general, an unknown parameter which could, for additional flexibility, be parameterized in terms of observed variable(s) z.

The FMM specification is attractive for empirical work in cross-section analysis because its functional form is flexible. Mixture components may come from different parametric families, although commonly they are specified to come from the same family. The mixture components permit differences in conditional moments of the component distributions, and hence in the marginal effects. In an actual empirical setting, the latent classes often have a convenient interpretation in terms of the differences between the underlying subpopulations. However, the number of latent classes is generally unknown and has to be treated as an additional parameter. This is a nontrivial complication that has often been handled as a model selection problem that is solved using penalized likelihood criteria. Bayesian analyses using Dirichlet process mixtures potentially allow for additional flexibility.

The main features of some popular models are summarized in Table 1.

Quantile Regression for Counts

Quantile conditional regression (QCR) is a robust semiparametric methodology for continuous response data. The conditional quantile function is a more general object of interest than the traditional conditional mean because it allows us to study potentially different responses in different quantiles of the outcome variable, and thus the entire distribution. It is consistent under weak stochastic assumptions and is equivariant to monotone transformations of the outcome variable. It is attractive because it potentially allows for response heterogeneity at different conditional quantiles of the variables of interest. By extending it to count data regression, one can overcome the standard and restrictive models of unobserved heterogeneity based on strong distributional assumptions. Further, QCR permits the study of the impact of regressors on both the location and scale parameters of the model and thus supports a richer interpretation under weaker distributional assumptions. A difficulty arises because the quantiles of discrete variables are not unique as the cumulative distribution function is discontinuous with discrete jumps between flat sections. By convention, the lower boundary of the interval defines the quantile in such a case. However, recent theoretical advances have extended QCR to a special case of count regression.

The key step in the extension of QCR to counts involves replacing the discrete count outcome y with a continuous variable $z=h(y)$, where $h(\cdot)$ is a smooth continuous transformation. The standard linear QCR methods are then applied to z. The particular continuation transformation used is $z=y+u$, where $u\sim U[0,1]$ is a pseudorandom draw from the uniform distribution on $(0,1)$. This step is called 'jittering' the count. Point and interval estimates are then retransformed to the original y-scale, using functions that preserve the quantile properties.

Panel Data

Count data panel models, like their linear counterparts, use three main frameworks: population-averaged (PA) models, random-effect (RE) models, and fixed-effect (FE) models. All are widely used in health econometrics. Maximum likelihood as well as moment-based estimation is common.

Given the scalar dependent variable y_{it} with vector of regressors x_{it} where i, $i=1,\ldots,N$, denotes the individual and t ($t=1,\ldots,T$) denotes time, the case of 'short panel' (small T) is empirically in health applications.

Many complications of count data panel models stem from discreteness of y and nonlinearity of the conditional mean. Assume multiplicative individual-specific scale effect α_i applied to exponential function,

$$E[y_{it}|\alpha, x_{it}] = \alpha_i \exp(x_{it}'\beta) \quad [7]$$

As x_{it} includes an intercept, α_i may be interpreted as a deviation from 1 because $E(\alpha_i|x)=1$.

Pooled or Population-Averaged Models

Operationally speaking, pooling involves 'stacking' cross-section observations and applying cross-section methods. Pooling multiple cross-section involves a strong assumption. The observations $y_{it}|\alpha_i, x_{it}$ are treated as independent, after assuming $\alpha_i=\alpha$, which implies absence of unobserved heterogeneity. The pooled model is also called the PA model. For parametric models, it is assumed that the marginal density for a single (i, t) pair,

$$f(y_{it}|x_{it}) = f(\alpha + x_{it}'\beta, \gamma) \quad [8]$$

is correctly specified, regardless of the (unspecified) form of the joint density $f(y_{it},\ldots,y_{iT}|x_{il},\ldots,x_{iT},\beta,\gamma)$. The PA specification accommodates some assumption about conditional dependence between y_{it}, over both i and t. Serial correlation of y_{it} suggests conditional dependence. Hence, after estimating the pooled model (by ML or moment-based estimator such as nonlinear least squares), 'residuals' may be dependent; hence a panel-robust or cluster-robust (with clustering on i) estimator of the covariance matrix can then be applied to adjust standard errors for such dependence.

The pooled model for the exponential conditional mean specifies $E[y_{it}|x_{it}] = \exp(\alpha + x_{it}'\beta)$. The model can be estimated by the efficient generalized method of moments (GMM) estimator, which embeds the generalized estimating equations (GEE) estimator in the statistics literature, which is based on the conditional moment restrictions, stacked over all T observations,

$$E[y_i - g_i(\beta)|X_i] = 0 \quad [9]$$

where $g_i(\beta) = [g(x_{i1},\beta), \ldots, g(x_{iT},\beta)]'$ and $X_i = [x_{i1}, \ldots, x_{iT}]'$.

Although the foregoing analysis is for additive errors, there are multiplicative versions of moment conditions that will lead to different estimators. Because of the greater potential for having omitted factors in panel models of observational data, fixed and random-effect panel count models are more flexible and plausible alternatives to the PA model.

Random Effects Models

A REs model treats the individual-specific effect α_i as an unobserved random variable, uncorrelated with the regressors $x_{i1}, \ldots, x_{iT_i}$, with specified mixing distribution $g(\alpha_i|\gamma)$, analogous to the cross-section case. Then α_i is eliminated by averaging over its distribution – an operation which is mathematically equivalent to integrating out α_i from the conditional distribution $f(y_{it}|x_{it},\alpha_i,\beta,\gamma)$. The resulting unconditional density for the *ith* observation is $f(y_{it}, \ldots, y_{iT_i}|x_{i1}, \ldots, \beta, \gamma, \eta)$ may have an analytical closed form for some combinations of $\{f(\cdot), g(\cdot)\}$, but in general it is handled numerically. Randomness restricted only to the intercept is easier to handle numerically than when it applies to both intercept and slope parameters, which means that the RE is not separable, as in an FMM.

As in the cross-section case, the NB panel model can be derived under two assumptions: first, y_{ij} has Poisson distribution conditional on μ_i and second, μ_i are independent, identically distributed (i.i.d.)-γ distributed with mean μ and variance $\alpha\mu^2$. Then, unconditionally $y_{ij} \sim \text{NB}(\mu_i, \mu_i + \alpha\mu_i^2)$. Although this model is easy to estimate using standard software packages, it has the obvious limitation that it requires a strong distributional assumption for the random intercept and it is only useful in the special case when the regressors in the mean function $\mu_i = \exp(x_i'\beta)$ are time invariant.

A potential limitation of the foregoing RE panel models is that they may not generate sufficient flexibility in the specification of the conditional mean function. Such flexibility can be obtained using a finite mixture or latent class specification of RE.

Fixed Effects Models

Given the conditional mean specification

$$E[y_{it}|\alpha_i, x_{it}] = \alpha_i \exp(x_{it}'\beta) = \alpha_i \mu_{it} \quad [10]$$

a FEs model treats α_i as an unobserved random variable that may be correlated with the regressors x_{it}. Such dependence may be present if one or more regressors are endogenous. In this sense the FE model deals with a limited form of endogeneity. It is known that ML or moment-based estimation of both the PA Poisson model and the RE Poisson model will not identify β if the FE specification is correct. The main difficulty in handling the otherwise more attractive FE model is that,in general in nonlinear panel models, the nuisance parameters α_i cannot be easily eliminated from the model, which is the incidental parameters problem.

However, under the assumption of strict exogeneity of x_{it}, the basic result that there is no incidental parameter problem for the Poisson panel regression is now established and well understood. The conditional likelihood principle can be used to eliminate α and to condense the log-likelihood in terms of β only.

Table 2 below displays the first-order condition for FE PMLE of β, which can be compared with the pooled Poisson first-order condition to see how the FEs change the estimator. The difference is that μ_{it} in the pooled model is replaced by $\mu_{it}\bar{y}_i/\bar{\mu}_i$ in the FE PMLE, where $\bar{\mu}_i = T^{-1}\sum_t \exp(x_{it}'\beta)$ and $\bar{y}_i = T^{-1}\sum_t y_{it}$ are time averages The multiplicative factor $\bar{y}_i/\bar{\mu}_i$

Table 2 Selected moment conditions for panel count models

Model	*Moment or model specification*	*Estimating equations or moment condition*
Pooled Poisson	$E[y_{it}\|\mathbf{x}_{it}] = \exp(\mathbf{x}'_{it}\beta)$,	$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}(y_{it}-\mu_{it})=0$ where $\mu_{it}=\exp(\mathbf{x}'_{it}\beta)$
Population averaged Poisson random effect (RE)	$E[y_{it}\|\alpha_i,\mathbf{x}_{it}] = \alpha_i\exp(\mathbf{x}'_{it}\beta)$,	$\rho_{ts}=\mathrm{cor}[(y_{it}-\exp(\mathbf{x}'_{it}\beta))(y_{is}-\exp(\mathbf{x}'_{is}\beta))]$ $\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\mu_{it}\frac{\overline{y}_i+\eta/T}{\overline{\mu}_i+\eta/T}\right)=0$ $\overline{\mu}_i=\mathrm{T}^{-1}\sum_t\exp(\mathbf{x}'_{it}\beta);\ \eta=\mathrm{var}(\alpha_i)$
Poisson fixed effect (FE)	$E[y_{it}\|\alpha_i,\mathbf{x}_{it}] = \alpha_i\exp(\mathbf{x}'_{it}\beta)$	$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\mu_{it}\frac{\overline{y}_i}{\overline{\mu}_i}\right)=0$,

is simply the ML estimator of α_I; this means that the first-order condition is based on the likelihood concentrated with respect to α_i.

The result about the incidental parameter problem for the PMLE model does not extend to the FEs NB1 model (whose variance function is quadratic in the conditional mean) if the FEs parameters enter multiplicatively through the conditional mean specification. This fact is confusing for many practitioners who observe the availability of the FEs NB option in computer packages.

Moment Function Estimation

Modern literature considers and sometimes favors the use of moment-based estimators that may be potentially more robust than the MLE. The starting point here is a moment condition model which mimics the differencing transformations used to eliminate nuisance parameters in linear models, that is, moment condition models are based on quasi-differencing transformations that eliminate FEs. This step is then followed by application of one of the several available variants of the GMM estimation, such as two-step GMM or continuously updated GMM.

Two alternative formulations are

$$y_{it} = \exp(x_{it}'\beta + \alpha_i)u_{it} \qquad [11]$$

$$y_{it} = \exp(x_{it}'\beta + \alpha_i) + u_{it} \qquad [12]$$

where, in the first case $E(u_{it})=1$, the x_{it} are predetermined with respect to u_{it}, and u_{it} are serially uncorrelated and independent of α_i. A quasi-differencing transformation eliminates the FEs and generates moment conditions whose form depend on whether one starts with eqn [11] or eqn [12]. Several variants are shown in Table 2 and they can be used in GMM estimation. Certainly, these moment conditions only provide a starting point and important issues remain about the performance of alternative variants or the best variants to use. In essence these specifications lead to GMM estimation which differ in terms of the weights attached to the moment conditions.

Conditionally Correlated Random Effects

Given the difficulty of eliminating FEs for any flexible functional form, an extension of the RE model offers an attractive alternative. The standard RE panel model assumes that α_i and x_{it} are uncorrelated. Instead, suppose that they are conditionally correlated. This idea, originally developed in the context of a linear panel model, can be interpreted as a compromise between fixed and random effects, that is, if the correlation between α_i and the regressors can be controlled by adding some suitable regressors, then the remaining unobserved heterogeneity can be treated as random and uncorrelated with the regressors. Although in principle a subset of regressors may be introduced, in practice it is more parsimonious to introduce time-averaged values of time-varying regressors. This is the conditionally correlated random effects (CCRE) model. This formulation allows for correlation by assuming a relationship of the form

$$\alpha_i = \overline{x}_i'\lambda + \varepsilon_i \qquad [13]$$

where $\overline{x}$ denotes the time average of the time-varying exogenous variables and ε_i may be interpreted as unobserved heterogeneity uncorrelated with the regressors. Substituting this into the above formulation essentially introduces no additional problems. To use the standard RE framework, however, it is needed to make an assumption about the distribution of ε_t and this will usually lead to an integral that would need evaluating. Estimation and inference in the pooled Poisson or NLS model can proceed as before. This formulation can also be used when dynamics are present in the model.

Dynamic Panels

As for linear models, inclusion of lagged values is appropriate in some empirical models. An example is the use of past research and development expenditure when modeling the number of patents. When lagged exogenous variables are used, no new modeling issues arise from their presence. When lagged dependent variables are introduced, for example, $y_{it} = \exp(\gamma y_{it-1} + x_{it}'\beta + \alpha_i)$, additional complications arise. First, in short panels, initial condition y_{i0} will have a persistent influence. Second, the presence of zero outcomes or large lagged outcomes can induce instability. The literature offers a number of solutions whose suitability should be determined on a case by case basis. However, a relatively simple and appealing solution of the initial condition problem is to embed the dynamics in a CCRE model by assuming that

$$\alpha_i = \gamma_0 y_{i0} + \overline{x}_i'\lambda + \varepsilon_i \qquad [14]$$

where the individual-specific effect absorbs the initial condition. Finally, the i.i.d. component ε_i can be dealt with using the RE framework for the panel data. Estimation of such a model may entail numerical integration to deal with the i.i.d. error and/or the initial condition.

Further Reading

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* **1**, 29–53.

Cameron, A. C. and Trivedi, P. K. (2009). *Microeconometrics using stata revised edition*. College Station, TX: Stata Press.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, 2nd edn. New York: Cambridge University Press.

Deb, P. and Trivedi, P. K. (2002). The structure of demand for medical care: Latent class versus two-part models. *Journal of Health Economics* **21**, 601–625.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* **52**, 701–720.

Hausman, J. A., Hall, B. H. and Griliches, Z. (1984). Econometric models for count data with an application to the patents – R and D relationship. *Econometrica* **52**, 909–938.

Koop, G., Poirier, D. and Tobias, J. (2008). *Bayesian econometric methods*. New York: Cambridge University Press.

Machado, J. and Santos Silva, J. (2005). Quantiles for counts. *Journal of American Statistical Association* **100**, 1226–1237.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.

Trivedi, P. K. and Munkin, M. (2010). Recent developments in cross-section and panel count models. In Ullah, A. and Giles, D. (eds.) *Handbook of empirical economics and finance*, pp. 87–131. London: Francis and Taylor.

Windmeijer, F. (2008). GMM for panel count data models. *Advanced studies in theoretical and applied econometrics* **46**, 603–624.

Winkelmann, R. (2003). *Econometric analysis of count data*. Berlin: Springer-Verlag.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**, 39–54.

Models for Discrete/Ordered Outcomes and Choice Models

WH Greene, New York University, New York, NY, USA

Introduction

This paper will survey the application of discrete choice models in health economics. The application of econometrics to understanding the health system takes place at several levels:

- *Aggregate*: modeling the behavior over time of aggregates such as health care spending and demographics.
- *Market*: modeling the behavior of specific markets such as hospital services markets and market structures and labor markets such as the market for nurses.
- *Individual*: modeling the behavior of individuals making decisions such as whether to have insurance, visit a physician, or how intensively to use the health care system.

It would be overly ambitious to attempt to cover the field in a single essay. This survey will focus on the third topic. The analysis of individual behavior in health economics generally involves a particular style and tradition of econometric model building. At the individual level, behavior often takes the form of discrete choices over particular sets of alternatives. To focus ideas, we consider the examples of whether individuals purchase insurance or not, or how they report their health status or satisfaction, or which type of products to purchase, or how many times they visit the physician or engage in a particular behavior such as smoking or consuming narcotics. Each of these is a discrete choice, in most cases, a binary choice between two distinct alternatives. Understanding these individual choices helps the analyst to understand aggregate behavior and, for example, the impact of policy changes on behavior.

Choice Modeling: Theory and Econometrics

There are two fundamental building blocks that underlie the methodology of discrete choice modeling, the model of random utility and the basic econometric binary choice model for choice between two alternatives. (The theory of random utility began in the 1920s. It gained great momentum with the econometric research by McFadden, 1974.) The econometric approach to analyze discrete choices departs from the assumption that an individual's behavior reflects an underlying preference structure that is consistent with the familiar constructs of microeconomic theory. That is, the individual's preferences are continuous, complete, and transitive. The implication is that choices are made in the setting of a particular process, or calculus. The nature of the choice mechanism, if not the choice itself, follows from some familiar axioms of economic theory: choices are continuous, meaning that small variations in circumstances generally lead to small or no changes in decisions; choices are complete, meaning that decision makers are able to rank any pair of alternatives presented; and choices are transitive, meaning that, in broad terms, choices are logically consistent. The model builder moves forward from this underlying preference structure to a model of the value or utility of a specific alternative, which it is convenient to label U_{ij}, with U meaning utility and j indicating the alternative. Thus, U_{ij} is the utility, or value to individual i of making choice j. The centerpiece of the econometric model is the random utility model (RUM), which states that from the point of the model builder, U_{ij} is a random variable with a particular form that will be described. It is noted at this convenient juncture, that the idea of random utility is from the point of the observer. It does not imply that individuals make decisions randomly, for example, by using a coin toss to decide whether or not to visit a physician or to consume a narcotic. The random utility model is used as a platform for the analysis on which aspects of behavior are 'observable' (or nonrandom), whereas others are 'unobservable' (so, from the observer's standpoint, random). Thus, for example, it can be confidently said that price is a determinant of whether an individual chooses to smoke. But, there are also other intrinsic and inherently unobservable features of the individual's psyche that motivate a decision to smoke; price alone is not the full story. The observed outcome appears random because observable information is not sufficient to provide a complete explanation of the observed choice.

Random Utility

For convenience, it will help to write the RUM in a particular additive form,

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

The model of random utility is stated from the point of view of the analyst – it embodies their understanding of individual behavior. It states that the utility of the observed individual is composed of a deterministic part that is amenable to model building and responds predictably to observable stimuli, and a random part that embodies the unobservable aspects of individual preferences. It is crucial to the understanding of behavior that it is assumed that under the same circumstances, the individual will always behave the same way. The randomness of random utility describes the analyst's understanding of the differences across different people.

It should also be noted that RUM is a model – a description of behavior. A common mistake is to attack the RUM as if it was an immutable statement of the precise nature of underlying utility, which would then be characterized as hopelessly naïve. The RUM is a broad characterization of the process that lies behind observations of choices that individuals make. To make it convenient to discuss ideas, a particular form for the utility function is assumed,

$$U_{ij} = \boldsymbol{\beta}'_j \mathbf{x}_{ij} + \varepsilon_{ij}$$

For the less mathematically inclined reader, it is noted, the vector symbol $\boldsymbol{\beta}'\mathbf{x}$ is used here as a shorthand for the equivalent linear equation, $U_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots \beta_K x_{ijK} + \varepsilon_{ij}$.

 doi:10.1016/B978-0-12-375678-7.00715-X

The variables x_{ijk} represent either attributes of the choices, such as the price of a particular kind of insurance policy, or characteristics of the individual, such as age, gender, or income. The definition of a utility function in this form embodies the assumption that these features influence the choice made because they influence utility. For example, an increase in the price of an insurance plan (it is assumed) will make that plan less desirable – the choice of that plan, all else equal, will afford the individual lower utility (because the plan's greater price will divert income from other activities that would provide utility). The lower utility, working a step backward, makes it less likely that an individual would choose that plan instead of others available.

Binary Choice

The second fundamental building block of a paradigm of modeling discrete choices is a mechanism that translates unobserved utility into an observed counterpart. The departure point is the choice between two alternatives. The random utilities for two alternatives, for example, 'being a smoker,' or 'being a nonsmoker,' would be represented by

$$U_{iN} = \boldsymbol{\beta}'_N \mathbf{x}_{iN} + \varepsilon_{iN}$$
$$U_{iS} = \boldsymbol{\beta}'_S \mathbf{x}_{iS} + \varepsilon_{iS}$$

These two utility functions are implied by the RUM. The assumption that observed behavior is consistent with an underlying preference structure of utility maximization implies that the (now hypothetical) individual i will make choice S if $U_{iS} > U_{iN}$ and will make choice N if $U_{iN} > U_{iS}$. Thus, the decision actually made is termed the revealed preference – the decision to be a nonsmoker reveals that the individual prefers being a nonsmoker to being a smoker. (There is a slightly inconvenient ambiguity in how to break a tie. It is assumed that if the individual is indifferent between the two alternatives, they will choose alternative N (or, the first alternative). In more formal mathematical terms, since the random variables involved are continuous, the probability of a tie is zero and one need not worry about it in developing the theory. It might be tempting to think the individual mentally tosses a coin when faced with indifference between two choices. However, this would violate the continuity assumption made earlier, and seems unrealistic as well.)

The utility maximization assumption implies an econometric model: Specifically, it can be reasoned backwards from the observed choice to the underlying outcome. In our example,

$$\text{Choice } N \rightarrow U_{iN} \geq U_{iS} \quad \text{or} \quad \boldsymbol{\beta}'_N \mathbf{x}_{iN} + \varepsilon_{iN} \geq \boldsymbol{\beta}'_S \mathbf{x}_{iS} + \varepsilon_{iS}$$

Collecting terms, this implies that if choice N is made, then it follows that $(\boldsymbol{\beta}'_N \mathbf{x}_{iN} - \boldsymbol{\beta}'_S \mathbf{x}_{iS}) + (\varepsilon_{iN} - \varepsilon_{iS}) \geq 0$. Combining and renaming terms, it is determined that choice N will be made if $\boldsymbol{\gamma}' \mathbf{x}_i + \varepsilon_i \geq 0$. This provides the underpinning of an econometric model that will help in understanding individual choices and the differences across individuals. In particular, if two individuals who make different choices are observed, two underlying sources for the difference can be asserted. The first arises from the deterministic parts of the utility functions. The second comes from the random parts. Suppose, for example, that income is the only observable difference between individuals in the analysis, income appears in the deterministic part of the utility functions. Then, according to the model, if income is higher, it makes one of the two choices more likely and the other less so. Of course, this does not imply that it can be deduced which choice the individual makes. The presence of the random term in the model implies that the information in hand suggests the impact of changes in income will be exerted only on the probabilities. Therefore, implications of the model can only be drawn in probabilistic terms. The model is completed with a specification of the probability distribution of the random term. Traditionally, the analysis has been based on one of two frameworks. The most appropriate choice for the behavior of ε would seem to be motivated by the central limit theorem, which describes the behavior of aggregates of small influences. Assuming a normal distribution produces the probit model. Although less natural from a behavioral standpoint, because of its convenient mathematical properties, the logistic distribution has often been specified instead. This gives rise to the logit model of binary choice. (Foundational work on this specific type of model appears in the bioassay literature. The modern, social science paradigm can be traced back to work on information theory by Walker and Duncan (1967). An introduction to the use of binary choice models is contained in Greene and Hensher (2010).)

Before turning to a survey of extensions, an application that should help to focus ideas is noted. Riphahn, Wambach and Million (RWM, 2003 – Riphahn *et al.*, 2003) studied the use of the health care system by a large sample of German households. Among the interesting variables in their data set is a question about health satisfaction (HSAT), a scale variable coded 0 to 10. For purposes of the application, this variable is recoded to be HEALTHY = 0 if HSAT < 6 and HEALTHY = 1 if HSAT > 6. The average of HSAT in the sample is 6.8. (This sample is a panel. There are 27 326 household years in the data, but the sample is 7293 households.) The RWM data is used to construct a probit RUM of the binary choice of whether the individual reported feeling healthier than average or not. Note that it is not appropriate to assume that the individual reports that they feel healthier than the average person in the sample – they would obviously not know that. An ongoing theme in this area of research is an understanding of how the model should accommodate the idea that different individuals would view the term 'healthy' differently. If there were an objective construct, 'health,' different individuals at the same location on the scale might still report different answers. The estimated equation shown in the first row of **Table 1** is

$$\begin{aligned}\text{Estimated}[U_{it,\text{Health}}] = &-0.414 + 0.301\,\text{INCOME} \\ &+ 0.071\,\text{MARRIED} \\ &+ 0.069\,\text{EDUC} - 0.153\,\text{PUBLIC}\end{aligned}$$

MARRIED and PUBLIC are dummy variables for marital status and whether the individual purchased public health insurance (roughly 88% of people did). (How coefficients in a probit model are computed, what they mean, and how they are used for estimation and inference in an analysis is discussed in Greene (2011, chapter 17) and in Greene and Hensher (2010).) Based on the description so far, one would infer from these results that increases in income, and education, and if

Table 1 Estimates of discrete choice models

	Constant	*Income*	*Married*	*Education*	*Public*	*Age*	*Kids*	*Health*
1. Healthy	−0.414	0.301	0.071	0.069	−0.153			
2. Healthy[a]	−0.577	0.144	−0.118	0.107	−0.171			
3. Healthy	−0.602	0.179	−0.118	0.108	−0.168			
	$\sigma=1.084$	$\sigma=0.589$						
4(a). Addon	−2.436	0.857	0.029				0.013	
4(b). Public[b]	3.921	−0.929		−0.172		−0.000 5		−0.041
5. Health[c]	1.463	0.194	0.037	0.034		−0.019 8	0.048	

[a]Estimated correlation = 0.559.
[b]Estimated correlation = 0.522.
[c]Estimated threshold parameter between outcomes 1 and 2 = 1.872.

one is married, all act to make it more likely that an individual will report feeling healthy, whereas if they purchased public insurance, they are more likely to report feeling unhealthy. An important point to note – will be returned to it below – is the possibility that the purchase of insurance included in this equation is, itself, motivated by the individual's health satisfaction. This ambiguous directionality of the causal effect in this equation is a fundamental aspect of the model building effort. (The opportunity to discuss elements of the econometric/statistical framework related to the estimation machinery – parametric vs. nonparametric, Bayesian vs. Classical, etc. – is not used. Although they are important questions, they are only secondary to the focus on the paradigm itself. See Greene (2011) for further discussion.)

Issues in Binary Choice Modeling

The binary choice model can be used to understand individual behavior or to try to forecast it. For example, in the preceding model, it is found that INCOME seems to be an important variable – its coefficient is very large compared to the others. Translating the coefficients in the model into meaningful quantities is one of the burdens of the model builder (see Greene (2011)). There are also some fundamental issues that are relevant here and in other modeling contexts described later. Several of them are noted, in particular: endogeneity, heterogeneity (and panel data) and selectivity. It will be convenient to pivot this discussion off the small model given above in the section on Extended Choice Models.

- The model for HEALTHY includes INCOME. It might be supposed that a self-report of healthy is a revelation of an underlying, objective measure of health outcome – an individual who reports that they feel healthy really is healthier than one who does not. It may well be that one's health is a determinant of their ability to earn their income, so that if a causal effect can be considered between healthy and income in the equation, it runs in both directions. That would require some special treatment in estimation of the model. Likewise, the model seems to suggest that those who purchase public health insurance feel less healthy. It is possible that individuals who feel less healthy (because they are less healthy) are more likely to purchase the insurance. Once again, the direction of the causality is uncertain. This problem of endogeneity (of the explanatory variable) is common in the analysis of health-related outcomes.
- It is difficult to argue that the model builder should ignore the heterogeneity issue. With panel data such as available here, there are a variety of strategies that can be used. The general result is that if there are features of the individual, albeit unobservable by the analyst, that can be reasonably assumed to be constant through the time period for which the data are observed, then, with some defensible assumptions, the model can be enhanced to accommodate individual heterogeneity. The 'random effects' (RE) model is an example. In this framework, the random component of the random utility model is assumed to include two parts, the overall random term, ε_{ijt} that contains the effects that vary from period to period, and a time constant term, u_{ij} that is a fixed (albeit unobserved) characteristic of the individual. The RE model assumes that this effect can be modelled as a random variable that varies across individuals. Row 2 of Table 1 shows estimates of the same model shown earlier, but with a common random effect included in the equation. The estimate of ρ of 0.559 is a measure of the variation of the time invariant part of the random part of the utility function. To assess how important this term is, the variance of this term is computed, which is $\rho/(1-\rho)=1.267$. The model assumes that the variance of ε is 1.000, so it would appear that there is greater variation in the invariant unobserved effect than in the time variant part.
- The model contains a coefficient of 0.301 on income. However, this value is translated into a meaningful measure of causal influence of income on health satisfaction, and will be translated into the same measure for every individual in the sample. This seems naïve – the impact of income on the probability of reporting good health will surely differ from one individual to the next. Although it has already been acknowledged that the model is meant to be succinct and descriptive, this still seems like a degree of realism that would prove important to consider. How one should accommodate heterogeneity in a choice model that is an econometric issue which many authors have considered (see, e.g., Train, 2003). An extension of the RE model is a similar model that allows the marginal utilities to vary over individuals as well. The random parameters model (RPM) allows coefficients to have a distribution across individuals. Row 3 of Table 1 shows the estimates of

a model with random coefficients. It requires a bit more work to translate the estimates of the structural parameters of an RPM into a meaningful set of numerical results. It can be seen in Table 1 that in the distribution of income coefficients across the sample individuals, it appears that the standard deviation is much greater than the mean.

- Data on health outcomes are often self reported. When the sample, itself, is self selected, then the generality of the econometric model is called into question. In particular, when individuals select themselves into the sample, and the motivation for participation is connected to the health outcome being analyzed, a problem of sample selection arises. Studies of drug efficacy based on data gathered from physician visits would present an example. Jones (2007) describes some methods of analyzing models with sample selection. Some general commentary on application of the models in health economics is given in Madden (2008).

Extended Choice Models

The binary choice model based on a random utility platform with a linear index function is the workhorse of discrete choice modeling in health econometrics. There are a wide range of variations on the binary choice model that accommodate different sampling frameworks and decision situations. Madden (2008), for example, discusses the choice of sample selection versus two part models. Broadly, two part models examine consumer decisions as a pair of sequential decisions. For example, in Harris and Zhao (2007), the authors examined smoking behavior. A two part model of the intensity of smoking behavior (e.g., number of packs of cigarettes per week) might involve a decision of whether or not to be a smoker, then, for smokers, a second decision, how much to smoke. The core of the model is its allowance of the determinants of the two decisions to differ – for example, price might motivate the intensity variable, but might be tangential to the base decision whether to be a smoker or not. For another example, another interesting variable in the RWM study is ADDON, which is an indicator of whether the observed individual purchased an enhanced type of health insurance. To purchase the addon insurance, one must purchase the public insurance, so a model that purports to describe ADDON must account for the condition that the individual purchases the public insurance. Not all individuals did; approximately 88% of the individuals purchased the public insurance. The model then describes these two simultaneous decisions. Row 4 of Table 1 contains estimates of such a model. The fairly high value of the correlation coefficient is interpreted to suggest that the unobserved factors that determine purchase of the public insurance also help to explain purchase of the addon insurance.

The list of similar extensions of the basic binary choice model is extensive. Two major directions of research in discrete choice modeling are examined, ordered choice and unordered choice.

Ordered Choices

The example of modeling in health economics includes a variable HSAT, which is self-reported health satisfaction. In the original data, this variable is coded 0,1, …, 10. That is, it is a 'scale' variable coded on an 11-point scale. Many surveys, such as the British Household Panel Survey (BHPS – see Contoyannis *et al.*, 2004) and the German Socioeconomic Panel (GSOEP – see RWM (2003)) that have been used above, include attitude variables such as health satisfaction (HSAT) or subjective well being (SWB). Pudney and Shields (2000) examine the promotion process in the UK nursing market. Conventional regression methods are inappropriate for modeling such variables. A natural extension of the random utility model provides an appropriate framework for the analysis. The model supposes that the observed scale variable reflects an underlying continuous preference scale. Thus, in the RUM framework, it begins with an assumption that $U_{i,\text{health}} = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$ as usual. The observed outcome is not utility, or the continuous counterpart to 'health.' Rather, the respondent is given an opportunity to place themselves in their choice on a scale, indicated '0', '1', etc. The outcomes are thus indicators of the strength feeling on the utility scale. For the 0–10 scale, for an individual in particular, a choice of 8 rather than 7 indicates an increase in perceived health, but the difference of one unit is not meaningful. Thus, the difference between 8 and 7 of 1 is not necessarily the same as the difference between 6 and 5. The observed outcome only suggests that a response of 8 represents a greater value on the preference scale than a 7 would. This observation mechanism gives rise to an ordered choice model. The ordered choice model has proved useful for many applications in health economics (and many other fields). See Greene and Hensher (2010) for an extensive survey. See Boes and Winkelmann (2004, 2006), and Contoyannis *et al.* (2004) for applications.

Like the binary choice model, the ordered choice model has been extended in many directions. For example, Harris and Zhao (2007) examined data on tobacco use behavior in which there is reason to suspect willful misreporting. The accommodation in the ordered choice model is to 'inflate' the zero outcome. An interesting problem in understanding ordered choices is that the model assumes that all individuals interpret the scaling in the same way. This is likely to be particularly problematic when models are used to compare health outcomes across cultures and countries. In the simple example above in the section on Binary Choice, the model is likely to produce different predictions if it is used to compare two countries in which one is populated by inherently optimistic individuals whereas the other is less so. The meaning of 'middling health' might be very different in the two cases, being a very negative statement in the first case noted and a positive one in the second. King *et al.* (2004) and a succession of authors have designed models that involve anchoring vignettes. Vignettes are designed to solicit attitudes on scales that (arguably) all respondents should agree on. The scale built into the ordered choice model is calibrated to accommodate these differences in the arrangement of the outcomes on the preference scale. King *et al.*'s. study used the approach to compare survey data on political efficacy in Mexico and China.

Unordered Choices

Models of discrete choice are also extended to analyze situations in which individuals select among unordered alternatives. For example, modeling a choice among different

health insurance plans is considered. To construct an application, it supposes that plans are differentiated by features such the amount of the copayment, ceiling, and coverage of specific situations and by the price of the insurance plan. An RUM of the choice of which plan to choose would assign a utility to each plan:

$$U_{i,\text{plan}} = \cdots + \beta_D \,\text{DentalLimit}_{i,\text{Plan}} + \beta_P \,\text{Price}_{i,\text{Plan}} + \gamma \,\text{Income}_i + \varepsilon_{i,\text{plan}}$$

Maintaining the assumption of utility maximization, it is assumed that the individual makes the choice that provides the greatest utility. Because it is a RUM, as in the previous cases, the model is completed by specification of the random terms. The counterpart to the binary probit model in this setting is the multinomial logit model (MNL). Since the work of McFadden (1974), the MNL has been used by generations of researchers to model unordered choices among multiple alternatives. An example is Gertler *et al.* (1987) who analyzed a Peruvian household survey of the choice of health care provider, public clinic, public hospital, or private doctor. The focus of extensive recent research has been on formulating more realistic formulations of the choice model that accommodate heterogeneity in preference structures. A leading example is Fiebig *et al.*'s (2010) development of a generalized mixed logit model, which is an elaborate form of random parameters model.

Multinomial logit models and variants such as the mixed logit model provide information on the preference structure and on effects of interest such as how price influences the choice among the alternatives. Another useful quantity revealed through the estimated preference structure is willingness to pay. It can be recalled, the model does not predict utilities – there is an inherent scaling and translation problem. But, the model does provide information about marginal utilities. The standard measure of willingness to pay for an attribute of a choice in a multinomial choice context is

$$\text{WTP} = \frac{\text{Marginal utility of feature}}{\text{Marginal utility of income}}$$

In this example, the marginal willingness to pay measure for an increase in the limit of the dental coverage of our (hypothetical) policy would be $\text{WTP}_{\text{dental}} = \beta_D/\gamma$. This would provide an empirical estimate of the amount that an individual would value (i.e., be willing to pay) for an increase in the desirable attribute of the choice, such as an increase in the amount of dental coverage in our example.

Conclusions

The econometric models examined in this essay form the platform for a large share of the empirical analysis of individual behavior in health economics. The random utility model and its corollary, the fundamental model of binary choice, are the pillars of econometric analysis throughout the social sciences. Because the choice variables in health economics often involve Likert-like scales and choices among sets of alternatives, the ordered choice and unordered choice models are natural settings in which to examine health economics outcomes.

See also: Analysing Heterogeneity to Support Decision Making. Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap. Missing Data: Weighting and Imputation. Models for Count Data. Multiattribute Utility Instruments: Condition-Specific Versions. Sample Selection Bias in Health Econometric Models

References

Boes, S. and Winkelmann, R. (2004). *Income and happiness: New results from generalized threshold and sequential models*, vol. 1175, SOI Working Paper 0407, IZA Discussion Paper, IZA.

Boes, S. and Winkelmann, R. (2006). The effect of income on positive and negative subjective well-being. University of Zurich, Socioeconomic Institute, Discussion Paper 1175, IZA.

Contoyannis, A., Jones, A. and Rice, N. (2004). The Dynamics of Health in the British Household Panel Survey. *Journal of Applied Econometrics* **19**, 473–503.

Fiebig, D., Keane, M., Louviere, J. and Wasi, N. (2010). The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science* **29**, 393–421.

Gertler, P., Locay, L. and Sanderson, W. (1987). Are user fees regressive? The welfare implications of health care financing proposals in Peru. *Journal of Econometrics* **36**, 67–88.

Greene, W. (2011). *Econometric analysis*. Englewood Cliffs, NJ: Prentice Hall.

Greene, W. and Hensher, D. (2010). *Modeling ordered choices*. Cambridge: Cambridge University Press.

Harris, M. and Zhao, X. (2007). Modeling tobacco consumption with a zero inflated ordered probit model. *Journal of Econometrics* **141**, 1073–1099.

King, G., Murray, C., Salomon, J. and Tandon, A. (2004). Enhancing the validity and cross cultural comparability of measurement in survey research. *American Political Science Review* **98**, 191–207.

Madden, D. (2008). Sample selection versus two part models revisited: The case of female smoking and drinking. *Health Economics* **27**, 300–307.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (ed.) *Frontiers in econometrics*. New York: Academic Press.

Pudney, S. and Shields, M. (2000). Gender, race, pay and promotion in the British nursing profession; estimation of a generalized ordered probit model. *Journal of Applied Econometrics* **15**, 367–399.

Riphahn, R., Wambach, A. and Million, A. (2003). Incentive effects on the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics* **18**, 387–405.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

Walker, S. and Duncan, D. (1967). Estimation of the probability of an event as a function of several variables. *Biometrika* **54**, 167–179.

Models for Durations: A Guide to Empirical Applications in Health Economics

M Lindeboom and B van der Klaauw, VU University, Amsterdam, The Netherlands

Introduction

Often one is interested in the time spent in a specific state and the effect of variables influencing the length of stay; for example, how long does a patient stay in a hospital and what is the effect of a medical intervention. The state can also be employment, and one may be interested in the effect of (changes in) health status on the probability that a worker leaves employment. Such processes are often described in the context of a duration model. In the duration models literature, the probability of leaving a specific state is referred to as the exit probability (in discrete time), or the exit rate or hazard rate (in continuous time). High exit rates are associated with short durations in the state and low exit rates with long durations.

Regression methods may not be useful in such applications. In practice, individuals are often observed for a limited time period, and, therefore, some individuals are not observed to have left the state of interest. Dealing with such censored observations is not straightforward in a regression model. It requires specifying a censored regression model, which often makes strong distributional assumptions. Furthermore, the value of some regressors may change over time. The health status of an individual can change during the course of an employment spell. It is unclear how one can include time-varying covariates in a regression model. Finally, a regression model considers only the mean duration. In applications one may be directly interested in the effect of a variable on the exit rate, or in the evolution of the exit rate over time. Example of the former is the effect of a drug on the recovery rate of a sick patient, or the effect of a health shock on the exit rate out of employment. These effects may be different early in the spell than later on.

This article reviews the literature with attention to empirical applications of such methods and the use of standard software such as STATA. The article will not fully cover all aspects of duration analysis. The remaining of this article is organized as follows. In the next section some formal concepts are introduced and simple nonparametric methods to describe duration data are provided. Section 'Parametric and Semiparametric Models' discusses different models used in applied duration analyses. Section 'Unobserved Heterogeneity in Duration Models: The Mixed Proportional Hazard' discusses the issue of unobserved heterogeneity in duration models. Section 'Other Relevant Issues in Applied Duration Analysis' provides a brief introduction to other relevant issues in the context of duration models, such as multiple spells, competing risks, and dynamic treatment evaluation. The article also includes an appendix with some relevant STATA commands, a description of a data set on sickness absence durations, and a link to this data set.

Concepts and Nonparametric Estimates

Concepts

Let T be a nonnegative random variable representing the time spent in a specific state. This can, for example, be the length of an employment spell or the duration for which a person is sick. In practice, when there are individual data on durations, outcomes of this random variable are observed. The distribution function of this random variable is given by $F(t)=\Pr(T<t)$, which denotes the probability that the individual leaves the state within t time periods. The distribution function is uniquely characterized by the so-called hazard rate, which describes the exit rate out of the state at a point in time given that the individual is still in the state. The hazard rate is in continuous time the instantaneous exit rate at time t and is denoted by $\theta(t)$. In discrete time, $\theta(t)dt$ is the probability that an individual who did not leave before time period t leaves the state within a short time interval dt after time period t.

If the hazard rate is decreasing in t, then exiting the state becomes less likely the longer the individual is in the state. In case of sickness, one might expect to see such a decreasing pattern; individuals who have been sick for only a short period are more likely to recover than individuals who have been sick for a longer period. A decreasing pattern in the hazard rate is often referred to as persistence or state dependence. The complement of the distribution function $1-F(t)$ is referred to as the survivor function $S(t)$. The survivor function describes the fraction of individuals who are still in the state after t time periods. This is thus the cumulative of not having left the state in all short time intervals before t. With high hazard rates, generally fewer people remain in the state when one proceeds over time; with low hazard rates, more people remain in the state. This illustrates the one-to-one relation between the hazard rate and the survivor function $S(t)$ (and its complement $F(t)$).

Nonparametric Estimates of the Hazard Rate: Bringing the Concepts to the Data

The appendix includes a link to a data set on individual sickness absence spells of teachers working in primary schools in the Netherlands. The data set includes individual sickness spells (t_i), an indicator whether or not the spell is right censored ($d_i=1$ for a completed spell and $d_i=0$ for a censored spell), and observed individual characteristics X_i. Censoring implies that recovery from sickness is not observed in the data. This may be the case because individuals are still sick at the end of the observation period, either because the observation window has ended or because the respondent leaves the sample (for instance, because she/he leaves the school). For

now it is assumed that the censoring mechanism is independent of the outcome variable of interest (the issue of independent censoring will be elaborated in Section 'Other Relevant Issues in Applied Duration Analysis,' where competing risks models is discussed). To define in STATA that one is using duration data, the command stset spell-length, failure(failed) should be used. Some elements of X_i may change over time, but for now it is assumed that these are fixed as of the start of the spell. The sickness spells in the data are all observed to start during the observation window. In the literature this kind of sampling scheme is referred to as a flow sample.

With data as described above, one can obtain a nonparametric estimate of the hazard rate. Recall that the intuition of the hazard rate is that it describes exit probabilities in short time intervals, given that the individual has not left the state before the start of the interval. Therefore, as a simple and direct estimate of the hazard rate after t time periods, one can take the ratio of the number of observed exits in the next short time period as a fraction of the number of individuals who are still in the sample at the start of the interval. Or in more technical terms,

$$\hat{\theta}(t)dt = \frac{\sum_{i=1}^{N} d_i I(t \le t_i < t + dt)}{\sum_{i=1}^{N} I(t \le t_i)}$$

The denominator is referred to as the risk set, i.e., those individuals who are still in the state at time period t and who are 'at risk' of leaving the state.

Often not only the hazard rate is of interest, but also the survivor function is estimated. Given the one-to-one relation between the hazard rate and the survivor function, the latter can be estimated using the estimated hazard rates. If it is imposed that the unit of time dt is 1, then the fraction surviving after the first period $S(1)$ is the fraction of people who have not left the state in the first period: $1 - \hat{\theta}(1)$. The fraction surviving after two periods $S(2)$ equals $(1 - \hat{\theta}(1)) * (1 - \hat{\theta}(2))$, and so on. There are alternative formulations for the nonparametric estimation of the survivor function. Most statistical software packages report the popular Kaplan–Meier estimator.

In STATA sts list should be used to get estimates of the hazard rate and the survivor function. With sts graph the Kaplan–Meier estimate for the survivor function is plotted. And sts graph, hazard plots the smoothed empirical hazard rate. Figure 1 shows the smoothed hazard rate for the data on sickness absence of teachers for the first 60 days. The figure shows a high recovery rate quickly after the start of sickness absence and declining hazard rates with occasional jumps thereafter. The declining hazard rate suggests state dependence. It may be that for each individual, recovery becomes more unlikely as time proceeds. Alternatively, there could be dynamic selection: those with less serious illnesses leave the state first, so that in the end the most serious cases remain. This illustrates that the nonparametric methods are useful for summarizing the data, but that it is important to control for heterogeneity within a sample of individuals. This is discussed further in the next section. Figure 2 shows the Kaplan–Meier estimate of the survivor function. At the start all individuals are present (at $t=0$, $S(t)=1$), and as time proceeds, more and more people leave the state. Note that, as

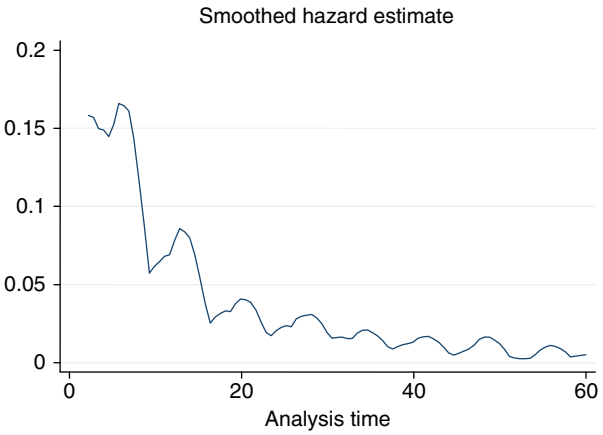


Figure 1 The smoothed hazard rate for the sickness absence example.

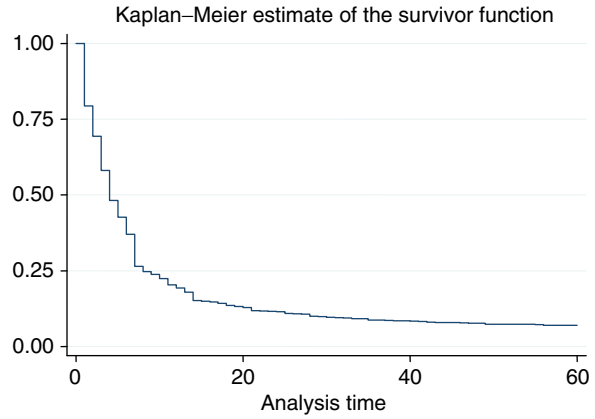


Figure 2 The survivor function for the sickness absence example.

expected, the high hazard rates in the beginning lead to a steep decline in the survivor function and the lower hazard rates later in time lead to slower decline in the survivor function.

Often, data describe different groups. For example, the data may contain sickness durations of individuals who received some treatment and of individuals who did not receive this treatment, or sickness absenteeism spells of teachers in different schools. One might be interested in whether or not hazard rates differ between groups. This can formally be tested using Logrank tests, which are nonparametric tests for the null hypothesis that the survival functions describing the durations in the different groups are identical. The underlying idea of the Logrank tests is that the order in which individuals exit the state is random in case the different groups have the same survival function. If after t time periods an exit is observed in one of the two samples, under the null hypothesis, the probability that the exit occurred in the first sample is simply the number of survivors in the first sample after t time periods as a fraction of the total number of survivors in both samples at this moment. The Logrank test is based on evaluating these probabilities for all observed exits in the data. It can also be used in case the data contain more than two groups. In STATA the comment sts test strata is used to perform Logrank tests,

where the variable strata denotes the different groups. Finally, if the data do not contain any censored observations, the ranksum test can also be used.

Parametric and Semiparametric Models

Parametric Models

The previous section mentioned that state dependence implies that the hazard rate is decreasing in the time spent in the state. There can be two reasons for observing state dependence. First, individual hazard rates are decreasing over time, which implies that for each individual, exit becomes less likely the longer the individual has already been in the state. So in case of sickness this implies that for each individual it is the case that she/he is more likely to recover in the next period if the sickness spell is still short than when the sickness spell is further progressed. The second reason for observing state dependence is dynamic selection, which means that individuals with good characteristics (i.e., with high exit rates) leave the state early. So the longer is the duration, the more the sample of survivors will move toward individuals with bad characteristics and consequently low exit rates. Also dynamic selection implies that (overall) hazard rates decrease in the elapsed duration.

In many settings, both causes for state dependence have different implications for public policy. To analyze how important both mechanisms are, often more structure is imposed on the hazard rates. A popular specification is the so-called proportional hazard (PH) specification:

$$\theta(t|X) = \lambda(t)\exp(X\beta)$$

The function $\lambda(t)$ is the baseline hazard. This describes duration dependence common to all individuals, which does not vary with individual characteristics. So a declining baseline hazard is the first explanation for observing state dependence. The role for regressors X is in the regression function $\exp(X\beta)$, which is specified such that it is nonnegative. The latter is required because hazard rates cannot be negative. The PH assumption implies that the ratio of the hazard rates of two individuals is constant over time. At any moment in time, individuals with good characteristics are more likely to leave. So when the elapsed duration is progressing, the composition of survivors moves more toward individuals with bad characteristics. The presence of heterogeneity, therefore, always causes aggregated hazard rates to decline.

The most straightforward way to estimate duration models is by using maximum likelihood. However, in the case of the PH model, this also requires parameterizing the baseline hazard. A simple parametric function is provided by the Weibull distribution, $\lambda(t) = \alpha t^{\alpha-1}$. This specification imposes a monotonic relationship for the effect of time on the hazard rate. For $\alpha > 1$ the hazard rate increases with time (positive duration dependence), for $\alpha < 1$ the hazard rate decreases over time (negative duration dependence), and for $\alpha = 1$ the hazard rate is constant over time (exponential distribution). This is illustrated in Figure 3.

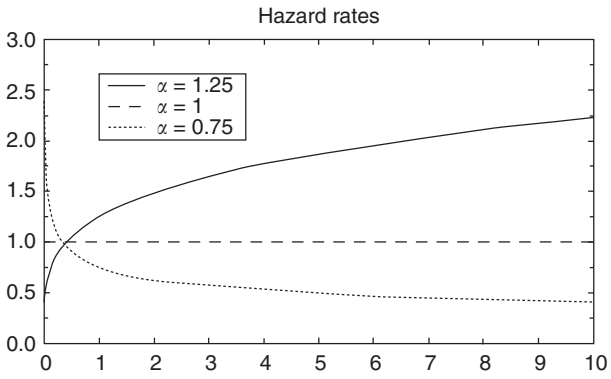


Figure 3 Hazard rates with Weibull duration dependence, for different values of α.

In STATA, PH models can be estimated using the command streg varlist, distribution (Weibull) nohr. Varlist is the set of observed characteristics included in X. The baseline hazard follows a Weibull distribution, but also other specifications can be used. For example, a log-normal specification allows for a nonmonotonic pattern of duration dependence. Finally, nohr ensures that in the output, the parameter estimates for β are reported rather than hazard ratios.

Semiparametric Models: Piecewise Constant Specification of the Baseline Hazard

Maximum likelihood estimation provides only consistent estimators if the model is specified correctly. Most choices for the baseline hazard involve specific functional forms and these may be too restrictive. For example, the Weibull distribution assumes that the hazard rate is constant, monotonically increasing or monotonically decreasing. This assumed duration dependence pattern may be violated in practice and will then lead to inconsistent estimates of the regression parameters β. Heckman and Singer (1984) provide a piecewise constant specification for the baseline hazard, which minimizes the distributional assumptions. The idea is that the baseline hazard takes different values on prespecified time intervals:

$$\lambda(t) = \exp(\lambda_k) \quad \text{if} \quad c_{k-1} < t \leq c_k$$

For $k = 1,2,\ldots,K$, with c_0, the lower bound, set to zero and c_K, the upper bound, set to infinity. The λ_ks are parameters to be estimated along with the regression parameters β. The cutpoints c_k are chosen in advance by the researcher. The piecewise constant specification requires a normalization. Either one can exclude an intercept from the regression function $\exp(X_i\beta)$, or a restriction should be imposed on the λ_ks. Most straightforward is to fix $\lambda_1 = 0$.

The key advantage of the piecewise constant specification is that by making the intervals very small, the specification can approximate any function arbitrarily close. In practice, when estimating a PH rate model with piecewise constant duration dependence, in each interval at least some exits should be observed. If there are no exits in an interval, the parameter λ_k cannot be estimated. Furthermore, without making strong extrapolation assumptions, the duration dependence pattern cannot be estimated beyond the latest observed exit in the

data. This implies that after estimating the model, it is, for example, difficult to estimate the expected duration. Consequently, it may be preferable to present estimates of the median or other quantiles of the distribution. In practice when estimating hazard rate models with piecewise constant duration dependence, it is advised to start with a very small number of broad intervals, and subsequently split intervals until no substantial improvement in the model is found. A complication is that most statistical software packages, such as STATA, do not directly allow for the option of piecewise constant duration dependence. However, in case one divides the data in discrete time intervals (such as weeks or months), the piecewise constant duration dependence translates into different intercepts for each time interval. STATA can handle time-varying regressors. The next section elaborates on the issue of time-varying regressors.

Semiparametric Models: Cox's Partial Likelihood

Even though one can try to minimize the distributional assumptions by choosing, for example, piecewise constant duration dependence, the maximum likelihood estimation requires the full specification of the hazard rates. The risk of misspecification, therefore, always remains present. An alternative estimator can be based on the rank order in which individuals exit the state. If $t_{(1)}$ describes the shortest observed duration in the sample, then the probability that individual i leaves the state at $t_{(1)}$ conditional that someone leaves at $t_{(1)}$ equals

$$\Pr\left(\text{individual } i \text{ leaves} \middle| \text{someone leaves at } t_{(1)}\right) = \frac{\theta(t_{(1)}|X_i)}{\sum_{j=1}^{N}\theta(t_{(1)}|X_j)}$$

$$= \frac{\exp(X_i\beta)}{\sum_{j=1}^{N}\exp(X_j\beta)} \qquad [1]$$

This probability is thus the ratio of hazard rates of individual i and the sum of all individuals in the risk set. Owing to the proportionality of the hazard rate, this probability does not depend on the baseline hazard $\lambda(t)$. This holds not only for the shortest duration in the spell. For all observed completed durations, the probability that a specific individual exits is the ratio of the regression function for this individual over the sum of regression functions of all survivors at that duration. Cox partial likelihood estimation only evaluates points where exits are observed in the data, and uses the probabilities provided in eqn [1] above. So without making any functional form assumptions on the baseline hazard, the parameters β can be estimated. In STATA the command stcox varlist, nohr is used for partial likelihood estimation.

Because partial likelihood estimates only β, the estimation results cannot be used for making predictions on durations, such as computing median durations (for individuals with specific characteristics). If one wants to make such predictions or is interested in the duration dependence pattern, the baseline hazard should be estimated. The Breslow method gives a procedure to retrieve the baseline hazards after partial likelihood estimation. To obtain the estimate for the baseline hazard in case of partial likelihood estimation in STATA, the options basehc (hazard) and basesurv (survivor) can be used.

Additional Complications in Estimating Duration Models

So far this article made the implicit assumptions that each spell is followed from the inflow in the state, that one observes the exact duration of the spell, that censoring is exogenous, and that individual characteristics remain constant during the spell. Next, the article briefly discusses the consequences if the data do not fit these assumptions.

Duration data may not be registered in continuous time, but describe in which time interval a spell ends. For example, the NFHS India describes for child mortality, the exact day of death if a child died within the first months after birth, the months of death if the child died between 1 month and the first birthday, and the year of death if the child died after the first birthday. In such cases the appropriate likelihood contribution should be based on the probability statements associated with such observations. Unfortunately, statistical software packages such as STATA do not handle this in their standard commands.

More problematic is the case in which spells are not followed from the inflow in the state. When, for example, looking at mortality rates of older individuals, often the data describe individuals who were at some calendar time older than a particular age. Or when looking at transitions out of employment into retirement, the data may contain a sample of older persons who are still working. These kinds of samples are referred to as stock samples, because they describe at a specific calendar time the stock of individuals who are in a state. Inference using stock samples is more complicated than using inflow samples. In stock samples individuals with bad characteristics (for leaving the state) are overrepresented. Individuals with bad characteristics experience, on average, longer time periods in the state. Therefore, they are more likely to be in the state at the calendar time of sampling and thus to be included in the stock sample. For example, when considering individuals on sickness absenteeism at a specific calendar time, this will include a relatively large share of long-term sick people, so people with more serious conditions.

In the most general set-up, a stock sample includes retrospective information (the elapsed duration e that an individual already has been in the state) and prospective information (residual duration r beyond the time of sampling). Inference can be based on the retrospective information, the prospective information, or both retrospective and prospective information. The expressions for the likelihood function become cumbersome and cannot be estimated with standard software in all but one case: the case in which one looks at the distribution of the duration r conditional on the elapsed duration e. This can be implemented in STATA by using the subcommand origin of the stset command.

So far it was assumed that censoring ($d_i=0$) is independent of the duration of interest (T). This may not always be the case. For instance in a clinical trial, terminally ill patients may be removed from the trial before the trial has ended. In this case the censoring is informative on the hazard rate. In case censoring is not exogenous, competing risks models should be used. Section 'Other Relevant Issues in Applied Duration Analysis' briefly discusses this.

Finally, until now it was implicitly assumed that the vector X includes only variables that are constant from the start of the spell, but some characteristics may change over time. For

instance, time spent in employment may depend on the health and the health status may change over time, or individuals may move to another region. The hazard rate can easily be modified to allow for time-varying characteristics. For example, one can write the PH rate as $\theta(t|X_i(t)) = \lambda(t)\exp(X_i(t)\beta)$ and the discussion above still applies. However, exogeneity of the regressors becomes an issue. The process $X(t)$ must be (weakly) exogenous, implying that values of $X(t)$ are only influenced by events that have occurred up to t and these events are actually observed. In the example of healthy lifetime, this excludes the situation where the individual knows that future health will fall (e.g., because of a chronic illness) and in anticipation reduces hours worked ($X(t)$). The article returns to this in Section 'Other Relevant Issues in Applied Duration Analysis' when discussing dynamic treatment evaluation. In case the exogeneity condition holds, STATA can be used for estimating. See the **stsplit** command in STATA to reorganize the duration data so that one can deal with time-varying covariates.

Unobserved Heterogeneity in Duration Models: The Mixed Proportional Hazard

Often the vector of observed characteristics X does not contain all variables relevant for leaving the state. For instance, in the context of individual lifetime, genetic factors may be relevant but unobserved, and the same may hold for factors like time preference or risk attitude driving health investment behavior. Ignoring unobserved heterogeneity causes the models to fail to control for dynamic selection, and thus the estimator for the baseline hazard $\lambda(t)$ will be inconsistent. Incorrectly ignoring the presence of unobserved heterogeneity not only generates spurious duration dependence, but also the estimates of β are biased toward zero. The latter holds even if the unobserved heterogeneity is orthogonal to the variables included in X. Therefore, one would like to allow for the presence of unobserved heterogeneity within the hazard rate.

Let V describe the unobserved characteristics in the hazard rate, and it is assumed that unobserved characteristics V are independent of observed characteristics X, i.e., $V \perp X$. The unobserved factors are random effects and follow at inflow ($t=0$) in the state a distribution $G(v)$. The individual exit rate is often specified using a mixed proportional hazard (MPH) specification:

$$\theta(t|X,V) = \lambda(t)\exp(X\beta)V$$

As time proceeds, individuals with good characteristics (high values of V) are more likely to leave the sample. This implies that among the survivors at time t, individuals with bad unobserved characteristics will be overrepresented. And the longer the time period t, the more the sample of survivors will move toward individuals with bad characteristics. This will lead to a discrepancy between the inflow distribution $G(V)$ of V and the distribution of V among the survivors in the sample ($G(V|X,\ T \geq t)$). As time proceeds, the observed hazard $\theta(t|X)$ declines faster than the baseline hazard $\lambda(t)$.

In the presence of unobserved heterogeneity, the observed hazard rate $\theta(t|X)$ does not factorize into t and X anymore. Therefore, it is not possible to use Cox's approach to eliminate the baseline hazard from the partial likelihood function. This implies that maximum likelihood estimation should be used. In STATA the command streg varlist, duration (Weibull) frailty(gamma) nohr can be used to optimize this loglikelihood function. It is, however, necessary to specify the unobserved heterogeneity distribution $G(v)$, for example, a gamma distribution. Furthermore, two normalizations are necessary on the duration dependence $\lambda(t)$, the regression part $\exp(X_i\beta)$ and the mean of the unobserved heterogeneity $\mathbf{E}[V]$ in the inflow.

Misspecification of $G(V)$ leads to biases in the estimators for $\lambda(t)$ and β if there is substantial censoring in the data. Furthermore, misspecification of $\lambda(t)$ usually causes significant biases in the regression parameters. It is therefore advisable to use flexible functional forms for the mixing distribution and for the baseline hazard. The previous section already introduced the piecewise constant specification as very flexible for the baseline hazard. Usually a mass-point distribution is considered to be the most flexible distribution for the unobserved heterogeneity. The idea is that V can take M different values, each acting as constants on the hazard rate:

$$\Pr(V = v_m) = p_m \qquad \text{with } p_m \geq 0 \quad \text{and} \quad \sum_{m=1}^{M} P_m = 1$$

In theory any distribution can be approximately arbitrarily close if M is chosen to be sufficiently large. However, in most applications M does not exceed 3. STATA does not have a command for a mixing distribution with discrete mass points. When optimizing the loglikelihood function, it is advised to start with a small M and then to add new points until the loglikelihood function does not improve anymore. The latter often implies either that new points have a very small probability mass or that the location of two mass points converge toward each other.

Other Relevant Issues in Applied Duration Analysis

Multiple Spells

So far it was implicitly assumed that for each individual only one spell is observed. However, it may be that the data contain multiple spells sharing the same unobserved component. For example, because within the observation period many workers experience multiple spells of sickness absenteeism, or because children born in the same family have the same unobserved component when modeling child mortality.

To formalize multiple spells, consider a cluster $c = 1,\ldots,C$ containing I_c observed spells. The hazard rate of observation $i = 1,\ldots,I_c$ in cluster c equals

$$\theta_{ic}(t|x_{ic}) = \lambda_c(t)\exp(X_{ic}\beta)$$

Each cluster is allowed to have a separate baseline hazard $\lambda_c(t)$ including duration dependence, cluster fixed effects, and possible interactions between both. To estimate the regression parameters β, stratified partial likelihood estimation can be used. The idea is similar to partial likelihood, but the risk set is

defined within clusters. The cluster-specific baseline hazards $\lambda_c(t)$ are eliminated and not estimated. This also implies that covariates that do not vary between observations in the same cluster are eliminated from X_{ic} and the corresponding covariate effects cannot be estimated. Obviously, clusters without observed exits do not contribute to the stratified partial likelihood function. So when modeling child mortality, only families in which multiple children are born and in which at least one child died contribute to the stratified partial likelihood function.

Finally, one cannot include all covariates in X_{ic}. The variables in X_{ic} should be (weakly) exogenous and may not be related to the observed exit of any individual in the cluster other than via the hazard rate. A violation of the weak exogeneity assumption may be if mothers make the decision to breastfeed a newborn child based on expected survival. Also if birth spacing might depend on the death of earlier born children, there may be a problem (in particular, if the observation period is limited and some spells are right censored). So breastfeeding and birth spacing should in those cases not be included in X_{ic}. However, if such variables are not included in X_{ic}, then there might be unobserved differences between observations in the same cluster that are not captured by the cluster specific effects, which again would violate the specification of the hazard rate above.

When baseline hazards are the same for all clusters, stratified partial likelihood estimation and partial likelihood estimation should give similar estimation results for the regression parameters β. Comparing the estimates from both procedures yields a test for similarity of the baseline hazards. Under the null hypothesis that all clusters have the same baseline hazard, both stratified partial likelihood estimation and partial likelihood estimation are consistent, but partial likelihood estimation is more efficient. Under the alternative hypothesis, only stratified partial likelihood estimation is consistent. This implies that a Hausman test can be used.

In STATA the command stcox varlist, strata(cluster) nohr provides estimation results where each cluster is allowed to have a separate baseline hazard. After estimating the coefficients these can be stored using est store sple. The same can be done for partial likelihood estimation of the regression coefficients (hereafter ple). When the results from both regressions are stored, the Hausman test can be done using the STATA command Hausman sple ple.

Competing Risks Models

It is mentioned above that duration models deal with censoring relatively easy, but it also stressed that this is only the case when censoring is exogenous. Exogenous censoring occurs, for example, when an individual is still in the state at the end of the observation period. There may, however, also be other reasons for censoring. For example, when considering sickness absenteeism of teachers, recovery may not be observed when a sick teacher quits working at a school. In such a case the censoring may be related to the process of recovery, because those individuals with more serious health conditions may decide earlier to quit working. In the case that censoring is not exogenous, a competing risks model should be used, which implies jointly modeling the process until recovery and the process until censoring.

There are also cases in which a researcher might be interested not only in the duration until leaving a specific state but also in the exit destination. For example, when modeling the age of death, the cause of death might be relevant. In these cases competing risks models are useful. Usually, data only describe the first exit, so except for the shortest duration, all other durations are latent. In the context of mortality, where individuals can die due to different causes (risks), death due to, for instance, a heart attack means that the individual did not die of cancer. Stated differently, the duration of dying of cancer is censored at the point where the individual dies of other diseases.

If the hazard rate of one cause is independent of death via other causes, then censoring due to these other diseases can be treated as exogenous and the parameters of the different hazard rates can be estimated by using successive analyses with the standard command of STATA (or other statistical software). Often it is likely that the different competing risks are not independent of each other. For example, healthy people are less likely to die both from cancer and from cardiovascular diseases. Erroneously assuming independence leads to incorrect inference.

Dynamic Treatment Evaluation

Often researchers are interested in the causal effects of (policy) interventions. A researcher might, for example, be interested in how a medical intervention affects the length of sickness. The effect of the medical intervention might depend on the moment in the sickness spell at which the treatment starts or the elapsed duration since the start of treatment. Estimating dynamic treatment effects is complicated, not only because there can be the usual endogeneity in assigning treatment to individuals, but also because there can be dynamic selection. If the treatment starts during the spell, individuals exposed to the treatment must have survived until the start of the treatment and may, therefore, have worse unobserved characteristics.

Usually one distinguishes between static and dynamic treatment evaluation. The difference is that in the static case, treatment starts at the beginning of the spell, whereas in the dynamic case treatment starts later during the spell. The key complication of static treatment evaluation compared to usual treatment evaluation is that the observation period is often limited and some spells are right censored. The methods discussed in the previous sections can be used to analyze static treatment effects, in particular when conditional on the observed individual characteristics X treatment assignment is independent of unobserved characteristics V. However, if conditional on X at the start of the spell treatment assignment is independent of V, this is not the case later during the spell. The intuition is that if treatment is successful in reducing the duration of spells, after some elapsed duration those who are still in the sample and who have been treated have, on average, worse characteristics than those who have not been treated.

Dynamic treatment evaluation is more complicated. In a dynamic setting, exclusion restrictions are often difficult to

justify. In particular, when individuals know at the start of a spell the value of an instrument describing whether or not they are likely to be treated early or later in the spell, they might already change behavior before the actual start of the treatment. For example, an individual who is on a long waiting list might search for alternative treatments, whereas an individual with a higher priority might just wait for the medical intervention. Even if the order on the waiting list is randomized, the realized order might have an effect on recovery already before the actual intervention. This implies that once an individual knows the instrument and understands the consequences, the instrument might affect the outcomes already before the actual intervention starts.

Most empirical studies focus on the *ex-post* effects of an intervention, which are the changes in hazard rates after an individual has been exposed to treatment. Empirical studies often ignore *ex-ante* effects, which describe differences in hazard rates from the beginning of the spell because at some moment in the spell, treatment may start. Exclusions restriction may be informative on the presence of the *ex-ante* effects. To identify the *ex-post* effects of treatment, a so-called *no-anticipation* assumption is required. This no-anticipation assumption imposes that individuals do not change behavior before the intervention after learning the actual moment of the intervention. So conditional on observed characteristics X and unobserved characteristics V, the intervention does not have any effect of hazard rates before the start of the intervention. This does not imply that the start of treatment is assigned randomly (conditional on observed characteristics X).

If the no-anticipation assumption holds, the *ex-post* effect of treatment δ can be specified within the MPH rate model for exits from the state

$$\theta_e(t|X,s,V_e) = \lambda_e(t)\exp(X\beta_e + \delta \cdot I(t>s) + V_e)$$

In this specification s describes the start of treatment and the indicator function $I(t>s)$ denotes if after t units of time the individual is exposed to the treatment. The key problem in the estimation is that the timing of the start of treatment s is often not independent of unobserved characteristics V. Therefore, the treatment effects model is extended by a PH rate model for the start of treatment

$$\theta_p(s|X,V_p) = \lambda_p(t)\exp(X\beta_p + V_p)$$

By allowing the unobserved terms in both hazard rates to be dependent, the model allows for endogeneity of the start of treatment. This model is often referred to as *timing-of-events* model.

To estimate the timing-of-events model, maximum likelihood estimation should be used. This requires the specification of the baseline hazards $\lambda_e(t)$ and $\lambda_p(t)$, as well as the joint distribution of the unobserved heterogeneity terms V_e and V_p. The model can allow for heterogeneous treatment effects. In particular, the treatment effect δ can be made dependent on observed characteristics X, unobserved characteristics V, the elapsed duration in the state t, the moment of starting treatment s, and thus also on the elapsed duration since the start of the treatment $t-s$.

This timing-of-events approach has in dynamic settings a number of advantages over other methods for treatment evaluation. Most important is that if treatment starts at some moment during a spell, it is difficult to define control groups for those individuals who receive treatment. In many cases, all individuals will receive treatment at some point when staying sufficiently long in the state. Therefore, individuals who are observed to be untreated are often so because they left the state relatively fast. Next, the timing-of-events model allows for selection on unobservables. Furthermore, most alternative methods require discretizing time, which has the disadvantage that results might be sensitive to the choice of the unit of time. Finally, the timing-of-events model explicitly models dynamic selection and the entry into treatment, which might be of interest in itself.

See also: Health Econometrics: Overview. Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap. Models for Count Data. Models for Discrete/Ordered Outcomes and Choice Models

Appendix: A Description of the Data and ome STATA Codes for Duration Models

STATA dataset Sickness_spells.dta: Link to the dataset
STATA dataset Sickness_spells.dta: Variable definition

Schooled	School number
Teachid	Teacher identification number
Spnr	Spell number
Splength	Length of spell in days
Year	Observation year
Rcensor	Dummy right censored = 1
Start	Year that employee entered school
Birthyr	Year of birth of employee
Gender	Gender (male = 1)
Marstat	Marital status (1 = single; 2 = married/cohabiting; 3 = divorced; 4 = widow(er)
Contract	Contract (1 = fixed; 2 = temporary; 3 = 50/50)
Hours	Hours of work
Lowgroup	Lower classes (classes 1, 2, 3, and 4)
Classize	Number of pupils in the class (97 = teacher has more than one class)
Schsize	Number of pupils in school
Teachnr	Number of teachers in the school
Public	Public school
Catholic	Catholic school
Protest	Protestant school
Special	Special school
Urban	Urbanization (1 = rural to 5 = big city)
Province	Province

Some STATA codes for duration models

```
* Open the data file
use "……\sickness_spells.dta"
*Make some transformations
Generate failed=1-rcensor
*Defining the duration data, treating the data as single record data
stset splength, failure(failed)
```

*describe the spell data

```
Stdes
```

*Kaplan–Meier estimates of the survivor function and the hazard rate

```
sts list
```

*Plotting the hazard rate (by subgroups)

```
sts graph, hazard
sts graph, hazard by(gender)
```

*Some additional tests for differences between the hazards for different groups

```
sts test gender
```

*Estimate simple parametric models and plot the hazard and survivor functions

*No unobserved heterogeneity

*Exponential model and a Weibull model

```
streg birthyr gender ……., distribution (exponential) cl(schoolid) nohr
streg birthyr gender……., distribution (weibull) cl(schoolid) nohr
stcurve, hazard
```

Reference

Heckman, J. J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**, 271–320.

Further Reading

Abbring, J. H. and van den Berg, G. J. (2003). The non-parametric identification of treatment effects in duration models. *Econometrica* **71**, 1491–1517.

Abbring, J. H. and van den Berg G. J. (2005). Social experiments and instrumental variables with duration outcomes. Discussion Papers 05–047/3. Amsterdam: Tinbergen Institute.

Abbring, J. H. and Heckman, J. J. (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects and dynamic discrete choice, and general equilibrium policy evaluation. In Heckman, J. J. and Leamer, E. (eds.) *Handbook of econometrics*, vol. 6B. Amsterdam: Elsevier Science.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Ham, J. C. and Rea, S. A. (1986). Unemployment insurance and male unemployment duration in Canada. *Journal of Labor Economics* **5**, 325–353.

Han, A. K. and Hausman, J. A. (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**, 1–28.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge: Cambridge University Press.

Ridder, G. (1984). The distribution of single-spell duration data. In Neumann, G. R. and Westergard-Nielsen, N. C. (eds.) *Studies in labor market dynamics*. Berlin: Springer.

Ridder, G. (1987). *The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence*. Mimeo.

Ridder, G. and Tunali, I. (1999). Stratified partial likelihood estimation. *Journal of Econometrics* **92**, 193–232.

Salant, S. (1977). Search theory and duration data: A theory of sorts. *Quarterly Journal of Economics* **91**, 39–57.

Van den Berg, G. J. (2001). Duration models: Specification, identification, and multiple duration. In Heckman, J. J. and Leamer, E. E. (eds.) *Handbook of econometrics*, vol. 5. Amsterdam: North Holland.

Van der Klaauw, B. and Wang, L. (2011). Child mortality in rural India. *Journal of Population Economics* **24**, 601–628.

Monopsony in Health Labor Markets

JD Matsudaira, Cornell University, Ithaca, NY, USA

In recent years there has been a surge in interest in models of imperfect competition in the labor market, and monopsony in particular (Boal and Ransom, 1997; Bhaskar *et al.*, 2002; Manning, 2003; Ashenfelter *et al.*, 2010; Manning, 2011). The term 'monopsony' was introduced by Joan Robinson in her 1933 book *The Economics of Imperfect Competition*. Taken literally, monopsony means a situation with only one buyer in a labor market and textbook discussions of Robinson's theory have accordingly been largely confined to 'company town' examples such as a coal mine in a rural town that is the only local employer. But Robinson's discussion makes it clear that monopsony power allowing employers to set wages may exist whenever frictions in the labor market give rise to an upward sloping labor supply (LS) curve at the individual firm level. (In this light, recent allusions to a 'new monopsony' framework represent a new shift of attention amongst sources of monopsony enumerated in Robinson's original discussion rather than new conceptual insights. Of course theoretical developments in the interim, particularly in the area of search theory, have helped make information related frictions more salient as a potential source of monopsony.) Such frictions might arise because "there may be a certain number of workers in the immediate neighborhood and to attract those from further afield it may be necessary to pay a wage equal to what they can earn near home plus their fares to and fro; or there may be workers attached to the firm by preference or custom and to attract others it may be necessary to pay a higher wage. Or ignorance may prevent workers from moving from one firm to another in response to differences in the wages offered by the different firms" (Robinson, 1933, p. 296).

Writing in 1946, Lloyd Reynolds predicted "The view that labor-market imperfections result in a forward-rising supply curve of labor to the firm... first elaborated by Mrs. Robinson... seems well on the way to being generally accepted as a substitute for the horizontal supply curve of earlier" (Reynolds, 1946, p. 390). But despite this early enthusiasm, monopsony models have rarely been invoked to characterize important segments of the US labor market, as the 'company town' metaphor led many economists to dismiss their broader relevance. One important exception is the labor market for registered nurses (RNs), where Yett (1970) suggested that monopsony was the most likely explanation for the consistent nurse 'shortages' reported by hospitals at least since World War II. To Yett, and many economists since, oligopsony seemed a natural model for the nurse labor market since many hospitals operate in counties with few other hospital competitors for workers, and there are reasons to believe that the geographic and occupational mobility of nurses is low. Indeed, the lion's share of empirical papers directly investigating the predictions of monopsony models cited in three recent reviews of the literature involve the nursing labor market (Boal and Ransom, 1997; Manning, 2003, 2011). (The only other occupation frequently featured in the monopsony literature is teachers (Landon and Baird, 1971; Boal, 2009; Falch, 2010).

In this brief research synthesis, it is attempted to illustrate why health labor markets so often evoke monopsony models to economists and review the empirical evidence on their relevance. Nearly all the relevant empirical literature concerns nurses, so it will largely be confined to the discussion of the same. It is started by providing a sketch of different models that imply an upward sloping LS curve, and thus some market power, for individual employers. It will be seen that these models provide an alternative, and arguably more persuasive explanation for several empirical facts that neoclassical models have struggled to rationalize. For example Manning (2003) argues that vacancies, wage dispersion across firms for similar workers, and employer provision of general skills training are all suggestive of monopsony. The second Section Suggestive Evidence therefore presents some facts about the nurse labor market with an eye toward assessing such a *prima facie* case for the plausibility of monopsony. Since many of these facts can be explained in a neoclassical framework, their existence does not constitute a 'severe test' (Mayo, 1996) of whether the labor market is monopsonistic. The third section reviews how economists have attempted to test for monopsony and assess the overall body of evidence. It is ended by discussing some areas that might be fruitful avenues for future research.

Why might we care whether monoposony is a better model of labor markets in health care? One reason is simply that monopsony models may provide a more accurate explanation for why many labor market phenomena, such as wage dispersion, vacancies, or large employers paying higher wages are observed (Manning, 2011, 2003). But the extent of monopsony power in the market also has important consequences for public policy. The implication of most monopsony models is that wages will be set lower in equilibrium than would be the case under perfect competition. If so, then labor will be inefficiently allocated across nursing and nonnursing sectors with too few nurses, obviously a concern in the context of the nursing shortages that have continued well past Yett's study and into the present. Some researchers in this area have suggested that the government should therefore be more active in monitoring nurse compensation, or perhaps promote unionization or a mandated wage to provide countervailing power to nurses in wage setting. Another public policy area where the extent of monopsony is important is the large public investment in nurse education. If employers are able to capture a substantial portion of the returns to nurses' human capital investments due to their ability to set wages below marginal products, then government subsidies may be less attractive relative to relying on private employers to pay for (part of) nurse training.

Models and Predictions

Since by now there are many excellent survey treatments of the various models implying monopsony power, here only a short

sketch of some of these models and their empirical predictions is provided (Boal and Ransom, 1997; Bhaskar *et al.*, 2002; Manning, 2003, 2011). In a perfectly competitive labor market workers are fully informed about the wage offerings of alternate employers, other jobs that are perfect substitutes for their own are readily available, and job mobility is costless. Under these assumptions, the LS curve facing a firm is infinitely elastic because a firm reducing their wage even slightly will see all its workers leave to work for other firms. At the same time, employers have no reason to offer anything above the market wage, since they can get all the workers they need at that wage. The monopsony models discussed in this section are all concerned with explaining why these assumptions might not be satisfied.

Employer Concentration or Collusion

The most straightforward case of monopsony arises when a single firm constitutes the only buyer of labor in a particular market. The classic work of Bunting (1962) showed that high employer concentrations were rare for the US labor market as a whole, leading many economists to dismiss the relevance of this type of model. In health care, however, this isolated firm model may apply more frequently as many hospitals operate in counties with few other competitors for nurse labor. (Of course, even in rural areas significant alternate employment opportunities may exist for nurses in doctors' offices, schools, etc.)

With a single employer for an occupation, the LS curve to the firm is the supply curve for the whole market. The market supply curve is typically assumed upward sloping, reflecting the idea that higher wages should induce more workers of that occupation to enter the labor market of the area, or workers already in the area to switch to the occupation in question. As described in Section Nurse Labor Supply, economists have presumed this elasticity to be low for nurses, making it more likely they might be vulnerable to being 'exploited' – in a sense defined below – by a monopsonistic employer.

Assuming a firm produces its output using only labor (L) its profit maximization problem can be written as

$$\max_{L} \pi(L) = pF(L) - w(L)L$$

where $\pi(\cdot)$ is firm profit, $F(\cdot)$ is the firm production function or revenue normalizing output price to unity, and $w(L)$ is the wage required to attract L workers (i.e., the LS function). The first-order conditions of this problem imply the following wage-setting rule

$$\frac{MRP - w}{w} = \frac{\partial w}{\partial L}\frac{L}{w} = \varepsilon^{-1} \qquad [1]$$

where MRP is the firm's marginal revenue product, t is the elasticity of LS with respect to the wage. Unless the LS curve is perfectly elastic, eqn [1] suggests that the monopsonist will pay workers less than their marginal revenue product, and employment will be lower than in the perfectly competitive case. Since Pigou (1924), economists have sometimes called the ratio on the left of eqn [2], the 'rate of exploitation', referred to below as E. Note that a nearly identical result can be derived from a model with multiple firms in the same market that collude to maximize joint profits (Boal and Ransom, 1997, p. 91).

Robinson's classic graphical analysis of equilibrium in a monopsonistic labor market is depicted in Figure 1. With an upward sloping LS curve, the marginal cost of hiring labor (MCL) lies above the LS curve so long as the employer must pay workers the same wage (i.e., cannot perfectly wage-discriminate). Thus, the employer hires workers up until the marginal cost of doing so equals the value of the marginal revenue product at L^*. As described above, there will be a gap between wages (w^*) and workers' marginal revenue product (MRP^*). But note also that at the wage prevailing in equilibrium, labor demand exceeds LS. In other words, firms face a shortage of labor, or 'vacancies' (Archibald, 1954), equal to $L' - L^*$. This is still an equilibrium because in fact firms cannot hire additional workers at the wage w^* as workers demand higher wages to increase their supply – in other words, firms are '(labor) supply constrained'. In this model, minimum wage policies that raise wages above w^* and as high as MRP^* can increase both employment and wages.

With a few large employers in the market, the situation might best be thought of in terms of oligopsony where there may not be collusion but employers do consider other firms' actions in their hiring decisions. The Cournot model provides a tidy analytical result: if employers choose labor to maximize profits taking other firms' employment levels as given, the first order conditions of the model suggest an employer-specific E_i equal to

$$E_i = \frac{MRP_i - w}{w} = \frac{L_i}{L}\varepsilon^{-1}$$

and thus an average (employment-weighted) market level E of

$$E = \sum \frac{L_i}{L}E_i = \varepsilon^{-1}\sum \left(\frac{L_i}{L}\right)^2 = \varepsilon^{-1}H \qquad [2]$$

where H is the Herfindahl index of employment concentration. The latter condition suggests a relationship between the level of exploitation (and thus wages) at the market level and the concentration of employment so long as the inverse LS elasticity is nonzero (i.e., so long as the labor market is not perfectly competitive). Note, however, that correlations across markets between H and wages are only indicative of

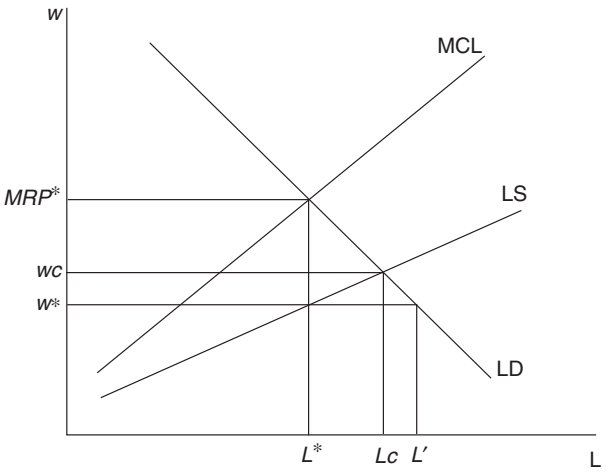


Figure 1 Equilibrium in a monopsonistic labor market.

monopsony power if total labor demand (i.e., the sum of firms' *MRP)* and LS are held constant.

What testable predictions arise from models of employer concentration or collusion? The most frequently explored prediction in the empirical literature is the idea that across markets, higher concentration should lead to greater 'exploitation' and lower wages. In part, such a prediction is generated by the assumption that collusion between employers is more likely to be enforceable with relatively few actors. Although intuitive, it's important to keep in mind that this assumption has little empirical evidence behind it – little is known about the prevalence of collusive agreements among employers across market structures. Short of collusion, models of oligopsony do generally predict a relationship between concentration and wage levels like eqn [2] above. It is important to remember that this prediction discriminates between competition and monopsony only if market level labor demand and supply are held constant. In general, markets that are more concentrated (e.g., rural areas) may differ along both of these lines leading to indeterminate biases – unaccounted for differences in demand would tend to amplify any negative correlation between concentration and wages, whereas differences in supply would tend to lead to a positive bias. (These presumptions about the direction of correlation are based on the assumption that areas with more concentrated employment probably have lower unobserved demand and supply shocks.)

Worker or Firm Heterogeneity

Another source of employer market power is that alternate firms may not be viewed as perfect substitutes in the eyes of job seekers due to differences in worker preferences or differentiation amongst firms. Bhaskar *et al.* (2002) present an intuitive model that illustrates the idea, and a similar idea is developed by Staiger *et al.* (2010) in their analysis of the market for RNs, discussed below. (Also see Bhaskar and To (1999) for a more detailed exposition of the model.) In their model, workers are uniformly distributed along a 1 km road with two firms (firm 0 and firm 1) located at either end. The cost of transportation to and from the work is t per kilometer, so a worker who lives x kilometers from firm 0 incurs a cost tx if he works for firm 0 and $t(1-x)$ if he works for firm 1. Thus, transportation costs differentiate the desirability of employment at the two firms for a given worker's location. If both firms paid identical wages, then all else equal workers would simply choose to work for the nearest employer. If firm 0 increased its wage by a small amount and firm 1 did not, it would attract more workers from firm 1 but clearly not all of them because a small wage increase would not be enough to compensate workers who live close to firm 1 for their cost of travel. Instead, LS would vary continuously and positively with the wage. The greater is t, the transportation cost parameter, the greater a wage increase will be necessary to increase LS for each firm and conversely with $t=0$ the supply curve will be perfectly elastic.

It should be clear here that transportation costs are simply a metaphor for some nonwage aspect of a job that affects the relative utility workers derive from employment at a particular employer. In nursing markets it is easy to imagine many such dimensions along which jobs might differ, including not only geography but also workload (patients per day), control over their work hours, the quality of facilities, etc. Perhaps less information is available regarding how much these other job aspects affect LS decisions – in other words, what is the magnitude of t for these aspects?

Equilibrium Search

Another strand of the literature on monopsony invokes search frictions as the cause of firm-level upward sloping LS curves. Search models provide an alternative, dynamic way of viewing firm-level LS. A firm's level of employment L_t can be written in terms of its previous level of employment, the separation rate of employees from the firm $s(wt)$, and the number new recruits $R(wt)$:

$$L_t = [1 - s(w_t)]L_{t-1} + R(w_t)$$

In this context, wages influence the size of the firm through the flows of workers to and from the firm. In steady state, the dynamic LS equation can be written

$$L(w) = \frac{R(w)}{s(w)}$$

or in elasticity terms,

$$\varepsilon_{Lw} = \varepsilon_{Rw} - \varepsilon_{sw} \qquad [3]$$

In other words, the LS elasticity to the firm can be estimated as the difference between the elasticity of new recruits and the elasticity of the quit rate with respect to the firm's wage (Card and Krueger, 1995). Manning (2003) suggests the simplification of assuming the elasticities on the right-hand side of eqn [3] are equal in magnitude, so the elasticity of LS is just twice the elasticity of separations with respect to the wage. In this light, a firm has market power whenever the elasticities of recruits or separations are less than infinite.

The model of Burdett and Mortensen (1998) is commonly invoked in this literature, and is striking in that it shows that a lack of perfect information (i.e., a finite job offer arrival rate) can generate monopsony power for firms even with identical workers and infinitely many small firms. (See Manning 2003) for a simplified presentation of this model and discussion.) One key result of the model is that the ratio of the job arrival rate to the separation rate indexes the degree of market power firms possess in a market. Manning (2003, p. 44) shows that this parameter is monotonically (negatively) related to the fraction of new hires (recruits) who come from nonemployment, and so the latter is positively related to the extent of monopsony in the market. The intuition is that the higher the fraction of new recruits coming from nonemployment, the less is the direct competition among employers for workers as fewer workers are leaving one firm for another.

Suggestive Evidence

The models of monopsony outlined in the Section Equilibrium Search all suggest a variety of symptoms that might

suggest their presence in the labor market. Before turning to a review of the literature formally testing the implications of such models, it is perhaps useful to review some features of the nurse labor market that hint at either the motivations or the predictions of these models. This evidence is suggestive, and meant only to encourage the reader to view the nursing market through a monopsonistic lens, as it has led many economists to do already. The literature that more formally tests the monopsony hypothesis is reviewed in the following section.

A preliminary note is that since nearly all of the empirical studies of monopsony in health labor markets use data from the US, international comparison is not provided. Although the basic forces at play in other developed countries are likely to be the same as those identified in the theoretical discussion in the Section Equilibrium Search, there are reasons to believe the degree of monopsony power for employers in other countries may differ from that found for firms in the US. For example, Acemoglu and Pischke (1999) argue that lower labor mobility in Germany gives firms there more monopsony power than in the US and helps explain the higher prevalence of employer sponsored training in Germany. It seems likely that differences in labor markets in the US and other developed countries like this exist in health care as well, so the discussion in this section should be viewed in this light.

Vacancies

More than any other feature, the presence of nurse-shortages evidenced by high vacancy rates has motivated the *prima facie* case that nurse labor markets are monopsonistic. Since the 1950s, hospitals have reported RN vacancy rates ranging as high as 10–20% (Yett, 1970; Buerhaus *et al.*, 2009). Amongst economists, early explanations for such shortages suggested that they reflected a state of dynamic disequilibrium whereby increases in the demand for nurses were constantly outpacing increases in supply, and wages were adjusting only slowly. Thus, at the (lagging) wage prevailing in the market, demand for nurses frequently exceeded supply (Blank and Stigler, 1957; Arrow and Capron, 1959). Invoking the work of Archibald (1954) who wrote "we will find oligopsony in the labour market whenever there are few employers of a given type of labour in an area and the cost of mobility is positive," Yett (1970) argued instead that hospitals were oligopsonists and so vacancies could exist even in equilibrium whatever dynamic (out of equilibrium) shortages may or may not be occurring in addition.

Although comprehensive national data are not available, more recent statistics suggest that relatively high vacancy rates persist for several types of nurses across a range of employment settings. A 2010 survey of 572 community hospital CEOs by the American Hospital Survey found RN, licensed practical nurse (LPN), and nurse aide vacancy rates of 4%, 4%, and 5%, respectively, which is low by historical standards due to the recession causing more nurses to join the labor force (Association, 2010). (One should be cautious about comparing reports of vacancy rates across different information sources, as the definitions used appear to vary significantly.) Among nursing homes, a 2007 survey found vacancy rates of 16.3%, 11.1%, and 9.5% for RNs, licensed practitioner nurses, and certified nurse assistants, respectively. (This is based on information from 3828 responding nursing homes from a 2007 survey of all 15 558 nursing homes in the US conducted by the American Health Care Association (Table 1, http://www.ahcancal.org/research_data/staffing/Documents/Vacancy_Turnover_Survey2007.pdf%20accessed%20May%2030,%202011).)

Concentration and Collusion

Another aspect of the nurse labor market that has evoked monopsony models is the relatively high employment concentration ratios. Yett (1970) suggested that the emergence of nursing shortages coincided with consolidation of RN employment in hospitals around World War II, and it remains true that the plurality of RNs work for hospitals. Many hospitals, in turn, operate in labor markets with few other employers: Yett (1970, p. 378) cites the figure that 60% of hospitals are in a health service area with fewer than six hospitals. More recent statistics suggest approximately 60% of US counties are served by only one hospital and approximately 25% of all hospitals are the only hospital in their county. (Author's tabulations of 2004 Area Resource File data. Of course hospitals are not the only employer of nurses but, especially in less populous counties, they account for the lion's share of employment.) That said, there are other employers of nurses such as nursing homes, doctors offices, schools, etc. and their presence in even the most rural markets may provide enough competition to prevent nurse exploitation in the sense of eqn [1].

Arguably the most direct evidence on monopsony would be the discovery of explicit agreements amongst employers to lower wages. Although economists tend to be skeptical of the sustainability of such arrangements among any substantial number of employers (For an early example, see Rosen (1970). He was also skeptical that hospitals, as nonprofit organizations, would engage in such rent appropriating behavior – a concern echoed by Pauly (1969). A classic reference is Stigler (1964).), there is scattered evidence that such arrangements have existed amongst hospitals. For example, in a survey of metropolitan hospital associations by Yett (1970), 14 of the 15 respondents reported having established successful 'wage-standardization' programs and the 15th asked for information about establishing one. Devine (1969) provides similar evidence on hospital collusion over wages in Los Angeles in the 1960s.

More recently, in 2006 nursing groups filed class action lawsuits against hospital chains in several large cities across the US (Greenhouse, 2006). These lawsuits alleged that the hospitals shared information about the wages of competitors for the purpose of keeping RN wages low (Miles, 2007), in violation of antitrust laws. An interesting feature of all of this evidence is that the collusion apparently took place in large metropolitan areas with many employers. In this light, the notion that the opportunity to collude is limited to those areas with high employer concentrations seems dubious. Indeed, the opportunity to share information on wages with competitors through consulting companies that conduct compensation surveys may enable a substantial degree of 'arm's-length' or tacit collusion.

Nurse Labor Supply

A common part of *a priori* arguments for monopsony in the nurse labor market is the notion that the market-level elasticity of LS for nurses is low due to high mobility costs. With few other hospital employers in an area, any job switching, particularly for RNs, would likely entail either relocating to a different area or switching to a different occupation. Yett argued both costs were substantial since (1) many nurses were married and their location decisions were likely constrained by their husbands' careers and (2) "few other occupations for which nurse training provides any advantage pay competitive salaries" (Yett, 1970, p. 381), so it would be costly for a nurse to switch to another occupation. To support this claim, Yett cited the results of a survey of nurses that reported low mobility rates overall, and suggested only 4–8% of nurses reported changing jobs because they were dissatisfied with their pay. More recently, Shields (2004) reviews a range of studies of nurse LS conducted over the past four decades, and reports the overall conclusion that nurse (market level) LS is indeed quite unresponsive to wages.

An implication of monopsony first highlighted by Robinson (1933, p. 302) is that employers may be able to discriminate between different types of workers. In particular, if there are two groups of workers with different LS elasticities to the firm, the profit-maximizing employer will offer a lower wage to the group with less elastic supply. Robinson provided a theoretical example where a wage differential emerged between male and female workers assumed equally productive, but men were organized into a trade union and thus had perfectly elastic supply at the union negotiated wage rate. In the nurse labor market, this phenomenon is a potential explanation for the wage premium paid to temporary contract, or registry, nurses.

In general, however, it is not clear that nurses stand out from other workers in the sense of having markedly different LS elasticity. For example, motivated by the intuition of the Burdett and Mortensen (1998) search model outlined above, Hirsch and Schumacher (2005) calculate the fraction of new RN hires from nonemployment using Current Population Survey data for the years 1994 to 2002, and find that 41.8% of recruits come from nonemployment on average each year. Surprisingly, however, they find that the ratio is actually higher (51.2%) for a nonnursing control group (women with a college degree). Although this does not imply that the RN market is not monopsonistic, it does cast some doubt on whether the nursing market is distinctively so.

Other Features

The features of the nurse labor market described in Section Nurse Labor Supply help to explain why health labor markets are singled out disproportionately in discussions of monopsony. There are other empirical 'puzzles' that have been documented more broadly in the economy, however, that some have argued are more naturally explained by monopsony models than by a model with a perfectly competitive labor market (Bhaskar *et al.*, 2002; Manning, 2003). Without going into much detail, it is noted in passing that many of these 'puzzles' are characteristic of the nurse labor market as well.

One hallmark of the competitive model is the 'law of one wage' (Bhaskar *et al.*, 2002, p. 156), or the prediction that similarly productive workers should all receive the same wage at jobs that are similar in their nonwage attributes. In nursing labor markets, wage differentials across employers are quite common. For example, in 2005 among 363 nursing homes in Los Angeles county, the average hourly wage level of RNs was US$28.2 (2005 dollars). But the firm at the 10th percentile of the firm-average wage distribution paid US$23.4 and the firm at the 90th percentile paid US$32.9, or 40% more. Such differentials are more pronounced for more skilled nurses, but are present for nurse aides as well. Among the same set of nursing homes, nurse aides made US$9.7 on average but the firm at the 90th percentile paid US$10.9 on average, or 25% more than the firm at the 10th percentile (US$8.7). (These data are taken from Matsudaira (2010), and are described in more detail there. Note that a limitation is that they represent average wages paid to a particular occupation by a firm. It is likely that this understates the degree of wage dispersion, though part of the differentials will reflect differences in composition among workers across firms, e.g. in experience. Also see Machin and Manning (2004) for an interesting and more detailed study of wage dispersion among nursing home workers in the UK.)

Some economists have argued that wage dispersion might reflect differences in worker productivity or compensating differentials for nonwage aspects of the jobs at different firms. But another piece of evidence argues against such explanations: turnover is negatively related to wage levels. Using data on the same set of Los Angeles nursing homes from 1981 to 2004, a regression of turnover rates for nurse aides on their average wage levels and a set of facility and year fixed effects suggests that raising hourly wages by $1 reduces turnover by approximately 7.4 percentage points (approximately 10%). If wage differences reflect unobserved productivity differences, there should be no reason for low wage workers to leave their jobs at higher frequency since they would not expect to be able to get the higher wage jobs. Similarly, differences driven by compensating differentials are inconsistent with the turnover result as switching from a low to a high wage job would not yield an expected utility gain.

Another interesting feature of labor markets that has challenged competitive theory is the fact that many employers seem to provide and pay for general human capital training to their workers. With perfect labor mobility, one might be skeptical that firms have an incentive to do this since they would be unlikely to recoup any investment in their workers' human capital that is not firm-specific Becker (1993). Despite this, it appears common for hospitals to pay for general skills training for their nurses, consistent with the view that imperfections in the labor market allow firms to recoup part of their investment in their employees general skills (Acemoglu, 1997). For example, May *et al.* (2006) document that many hospitals in their survey offer in-house nurse training programs, or subsidize their staff's training at nearby nursing schools. Benson (2011) provides a test of monopsony of sorts based on this reasoning, showing that hospitals with higher concentration ratios in their metropolitan area are more likely to subsidize training to RNs.

Empirical Studies of Monopsony

What direct evidence is there about whether nurse labor markets are actually monopsonistic? Previous studies on this question can be grouped into two broad categories: (1) an early literature that attempts to determine whether there is a link between employment concentration and the level of nurse wages, and (2) a more recent and smaller set of papers attempting to directly estimate the facility level elasticity of LS for nurse employers. Each of these literatures are discussed in turn.

Concentration and Wages

The first serious empirical test of the hypothesis that nursing markets are monopsonistic was Hurd (1973). Using data from the 1960 Census, Hurd estimates the relationship between hospital employment concentration and median nurse earnings across the 100 largest Standard Metropolitan Service Areas (SMSAs). The regression model includes control variables for the cost of living, the percentage of nurse employment accounted for by hospitals, the percentage of hospital employment accounted for by federal hospitals and the percentage accounted for by state and local hospitals, and the percentage of nurses receiving earnings who worked for less than 50 weeks. Hurd finds that holding these other factors constant, there was a significant negative relationship between concentration, measured by the share of employment of the eight largest hospitals, and median nurse earnings. In auxiliary analyses using wage data from the Bureau of Labor Statistics for a subset of cities he finds a similar relationship, and interpreted his findings as supportive of the monopsony hypothesis.

Subsequent studies by Link and Landon (1975), Feldman and Scheffler (1982), and Robinson (1988) (Robinson (1988) tests the prediction that the employment of nurses will be relatively low in areas with higher hospital employment concentrations.) confirm Hurd's findings, using hospital-level data and slightly different measures of firm concentration (e.g., a Herfindahl index based on hospitals' share of total beds in a city). As emphasized above, however, a correlation between concentration and wages is only evidence of monopsony if other factors related to LS and demand are held constant across markets.

More recent studies, such as Adamache and Sloan (1982) and Hirsch and Schumacher (1995), suggest that the relationship between concentration and wages may not be robust to better controls for such factors, such as population density and the wages of alternative occupations. The best study in this literature is probably Hirsch and Schumacher (1995), who pursue a two-step strategy for testing the prediction that hospital concentration depresses wages. Using data from the 1985 to 1993 Current Population Survey Outgoing Rotation Group (CPS-ORG) files, they first identify a control group of workers for each type of three nurse occupations (RNs, LPNs, and nurse aides) based on similarities in educational requirements. Then, for each occupation they estimate the relative wage gap between nurses and their respective control group separately for each of 252 geographic areas – including 202 metropolitan areas and 50 nonmetropolitan area state groups – controlling for worker characteristics and time effects. In a second-step regression, they then test whether hospital concentration and market size are correlated with the nursing wage differential and find no evidence that supports such a claim for any of the nurse occupations considered. (In a follow-up study, Hirsch and Schumacher (2005) use a similar design to see whether an alternative measure of monopsony power – the fraction of RN recruits from nonemployment – is correlated with wages across geographic areas. They find again no evidence for monopsony power for hospitals.)

The research design in Hirsch and Schumacher (1995) is clearly better than earlier studies in that measuring nursing wage differentials relative to a control group more effectively controls for market specific differences in demand and supply conditions that apparently confound earlier estimates. That said, the design relies heavily on the premise that nonnursing occupations are not subject to monopsony, or more accurately that variations in monopsony power in nonnursing labor markets is uncorrelated with variations in hospital concentration and market size. If monopsony is a more pervasive feature of labor markets for other occupations, then the approach may understate the wage-setting market power of hospitals.

Labor Supply to the Firm

As discussed in Section Models and Predictions, the key distinguishing feature of monopsony models relative to perfect competition is an upward sloping LS curve to individual firms. Despite its importance, only a handful of studies have attempted to credibly estimate this elasticity parameter. The reasons for this are readily apparent: the observed changes in wage and employment levels across firms are in general the result of changes to both supply and demand. So long as the supply curve is shifting, the observed equilibria cannot reliably be used to identify its slope. This is a well-known problem in economics dating back at least to Haavelmo (1943), but the solution of finding a 'demand shifter' to use as an instrumental variable for the observed quantities, tricky in the best of settings, is particularly difficult in this case. Since firm-level LS is of interest, any instrument must act differentially only on the labor demand of the specific firm in question so market wide phenomena are off the table.

Sullivan (1989) was the first to make a serious attempt at estimating the firm-level elasticity of LS to individual hospitals using data from 1979 to 1985. He derived LS equations based on different assumptions about the nature of competition among firms in an oligopsonistic setting, and controls for hospital and region specific fixed-effects to estimate the (inverse) elasticity of supply. For example, assuming Nash equilibrium in employment levels his estimating equation is

$$w_{rit} = \alpha i + \delta_r t + \theta n_{rit} + \gamma on_{rit} + \varepsilon_{rit}$$

where w_{rit} represents log wages for hospital i in region r at time t; n_{rit} is the log number of nurses employed; and on_{rit} the log of the sum of nurses employed at other hospitals. θ represents the inverse elasticity of supply. The simultaneity problem is addressed by using the number of caseloads and average length of stay as instrumental variables for the number of

nurses at hospitals, the notion being that these variables affect output demand and thus the derived demand for nurses but not necessarily nurse supply for a given hospital. Sullivan finds that the inverse elasticity of supply is approximately 0.79 (with standard error of 0.13) over a 1-year period and 0.26 (0.07) over a 3-year period and asserts this represents a significant amount of market power for hospitals. In a static model these elasticities can be used to compute the 'markup' of marginal product over wages using eqn [1], in this case implying wages are between 43% (for 1-year changes) and 21% (for 3-year changes) below marginal product. In a dynamic setting, however, this 'rate of exploitation' is a weighted average of short and long run elasticities where the weights are a function of a firm's discount rate. Assuming a long run elasticity of zero, Boal and Ransom (1997, p. 105) suggest Sullivan's estimates imply that wages might be set between 87% and 96% of marginal product. Using this logic, they characterize Sullivan's results as being suggestive of only slight market power for hospitals but of course such a conclusion rests on the accuracy of its assumptions.

The validity of this instrumental variable has been questioned. Staiger *et al*. (2010) point out that Sullivan's sample brackets a period when Medicare's Prospective Payment System is introduced, and suggests that much of the variation in hospital days over the period was therefore endogenous as the transition presumably may have led to independent (downward) pressure on nurse wages. Manning (2003) suggests, alternatively, that caseloads might be related to population shocks, and thus might fail the exclusion restriction.

An instrumental variable strategy is also employed by Staiger *et al*. (2010) who use legislated wage changes in Veteran's Affairs (VA) hospitals to identify the firm-level LS elasticity of RNs. Similar to the analysis in Sullivan (1989), Staiger and his coauthors adopt an explicit model of oligopsony (based on Salop (1979)) where hospitals compete most intensively with hospitals in close proximity, leading to an estimating equation where employment depends on a hospital's own offered wage but also the average wage at nearby hospitals. They demonstrate that VA wage changes affect the wage levels of RNs at nearby hospitals (up to 30 kms away), suggesting that hospitals do have the ability to set wages. Using gaps between the newly legislated wage and wages at the time of the legislation as instruments for wage changes, the estimated LS elasticity over a 2-year period ranges from approximately 0 to 0.2 with standard errors approximately 0.13 (or an inverse elasticity ranging from approximately 5 to infinity). Even using the upper bound of the 95% confidence interval from Staiger *et al*.'s estimates implies that the inverse elasticity of LS is at least 2, far from the 0 assumed by the theory of perfectly competitive labor markets.

Matsudaira (2010) attempts to estimate the degree of monopsony power for nursing home employers using a different strategy. In 2000, the state of California adopted minimum staffing regulations for nursing homes requiring them to employ a minimum number of nursing hours for each patient in residence. Depending on the gap between a home's initial staffing level and the legislated threshold, this law created more or less pressure to hire additional nurses to comply. Thus, Matsudaira uses this measure of the staffing gap as an instrument for subsequent changes in nurse employment.

Despite finding that homes initially out of compliance with the staffing law did hire significantly more nurse aides than those already in compliance, there were no differences in wage changes of aides across these groups of firms suggesting a highly elastic firm-level LS curve (i.e., an inverse elasticity close to 0). Since homes complied with the law almost exclusively by hiring nurse aides, the LS elasticity estimates for more skilled nurses (RNs and LPNs) were not estimated.

The estimates in Matsudaira (2010) differ markedly from those in Sullivan (1989) and especially Staiger *et al*. (2010). This may reflect differences in the supply elasticities of different kinds of nurses – nurse aides do not have near as much occupation-specific human capital, and so may have a broader set of alternative employers and thus more elastic LS. Or, if factors such as ignorance about alternative wage offers are the primary source of labor market frictions and this ignorance affects all occupations similarly, then the results may well be in conflict. A third possibility raised by Manning (2011) is that none of the studies are accurately measuring the firm-level LS elasticity, and that the models of firm-level LS used in the literature reviewed here are overly simplistic. This point is returned to below.

Discussion

Overall, the evidence above presents a very mixed case on the empirical relevance of monopsony models for understanding the nurse labor market. On the one hand, as in other markets there seems to be strong *prima facie* arguments that the market is monopsonistic ranging from 'smoking gun' evidence of wage-fixing, to reports of vacancies, to wage dispersion and provision of training. On the other hand, formal tests of the implications of monopsony theory have yielded varied results. The best studies on the relationship between employment concentration and wage levels suggest there is no relationship, and direct estimates of firm-level LS elasticity have produced some estimates consistent with extremely inelastic LS, and some estimates consistent with perfectly elastic LS.

It would be suggested that part of the reason for these ambivalent results is the reliance on overly simplistic theoretical models to guide empirical work. As noted above, although some models predict that collusion is more likely in more concentrated industries, there seem to be many cases where employers have colluded to keep wages low even in large, fairly unconcentrated markets. (Relatedly, Levenstein and Suslow (2006) report that although most cartels that have been studied in the product market have few members (with a median number of companies approximately 6 to 9), about one-third have more than 10 members with some having hundreds of members. They report that with cartels of many companies, industry associations often play a key role in coordination.) More work on the prevalence of collusive agreements on wages and competition for employees, and the effects of industry associations and wage-information sharing through compensation surveys would be useful, particularly given the recent legal actions taken by RNs alleging wage-fixing by hospital chains in many large MS as in the US. It may well be that traditional measures of concentration are a poor proxy for the prevalence of collusion among employers to

keep wage levels low in the nursing labor market. (Another interesting direction for future research is exploring the ways in which regulations restrict competition in the labor market. In a recent study, Kleiner and Park (2010) show evidence that state licensing rules restricting the work that can be done by dental hygienists have important effects on the earnings of hygienists and dentists. Licensing is obviously a pervasive feature of the health labor market, as are noncompete clauses among physicians. Exploring the consequences of such regulations for health care workers would be an interesting addition to the literature.)

Manning (2011) makes the point that if the firm-level LS function is not one-to-one with respect to wages, then the elasticities estimated in the literature may be incorrect. For example, if LS to the firm depends both on wages and recruitment expenditures (e.g., on advertising vacancies), then faced with a mandatory increase in employment it may be optimal for the firm to respond by increasing recruitment expenditures rather than by increasing wages. If such a model applied, the results of Matsudaira (2010) might overstate the LS elasticity to the firm, and conversely a design estimating the impact of a legislated wage increase on employment like Staiger *et al.* (2010) might understate the degree of elasticity. Other models with heterogeneity in worker quality or nonwage aspects of jobs can have similar implications. For instance, Matsudaira (2010) cautions that firms may respond to the mandate to hire more nurses by reducing their hiring standards with respect to worker quality at a given wage. If so, then the LS curve to the firm for nurses of a given quality may well be upward sloping even given his result that wages remain constant in firms that hire more nurse aides. Currie *et al.* (2005) test the predictions of a monopsony model in the context of examining the effects of hospital mergers on RN wages, and suggest that nurse effort (proxied in their context by nurse–patient staffing ratios) may be an important nonwage dimension that hospitals use to affect LS. They suggest that rather than suppressing wages, hospital mergers may lead to increased nurse effort for a given wage, a result consistent with the predictions of monopsony in their model. (The result is also consistent with a contracting model that they develop.)

The empirical literature on the importance of monopsony in the nurse labor market has yet to provide a conclusive answer. As suggested above, however, this is in part because our theoretical understanding of frictions in the labor market has evolved. Unfortunately it is difficult to formulate tests that would allow one to definitively reject monopsony or perfect competition under all theoretical formulations, and the tests that suggest themselves are hard to capture in the 'real world.' To wit, the thought experiment "what would happen if one employer was randomly forced in a market to increase its wage holding the wages of all competitors constant?" is easy to posit but near impossible to observe in the wild. Moreover, formulating direct tests of more general models of monopsony (Manning, 2006) present a challenge since many other determinants of LS are hard to observe at the firm level. In general, there are little data on worker quality, nonwage attributes of jobs, recruitment expenditures, and the other margins along which employers might adjust in response to (firm-specific) labor demand shocks. However, there may be relatively greater opportunities to advance a research agenda along these lines in the nurse labor market due to a long history of health management studies focused on understanding the determinants of nurse turnover and job satisfaction.

Acknowledgment

The author thanks Matthew Freedman for helpful comments on an earlier draft, and the National Science Foundation for financial support under grant SES-0850606.

References

Acemoglu, D. (1997). Training and innovation in an imperfect labour market. *The Review of Economic Studies* **64**, 445–464.

Acemoglu, D. and Pischke, J. S. (1999). Beyond becker: Training in imperfect labour markets. *The Economic Journal* **109**, F112–F142.

Adamache, K. W. and Sloan, F. A. (1982). Unions and hospitals: Some unresolved issues. *Journal of Health Economics* **1**, 81–108.

American Hospital Association (2010). The state of America's hospitals – Taking the pulse. Available at: http://www.aha.org/research/policy/2010.shtml (accessed 26.07.13).

Archibald, G. C. (1954). The factor gap and the level of wages. *The Economic Record* **30**, 187–199.

Arrow, K. and Capron, W. (1959). Dynamic shortages and price rises: The engineer-scientist case. *The Quarterly Journal of Economics* **73**, 292–308.

Ashenfelter, O. C., Farber, H. and Ransom, M. R. (2010). Labor Market Monopsony. *Journal of Labor Economics* **28**, 203–210.

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis with special reference to education*, 3rd ed. Chicago, IL: University of Chicago Press.

Benson, A. (2011). *Firm-sponsored general education and mobility frictions: evidence from hospital sponsorship of nursing schools and faculty.* Cambridge, MA: Massachusetts Institute for Technology. unpublished Mimeo.

Bhaskar, V., Manning, A. and To, T. (2002). Oligopsony and monopsonistic competition in labor markets. *Journal of Economic Perspectives* **16**, 155–174.

Bhaskar, V. and To, T. (1999). Minimum wages for Ronald McDonald monopsonies: A theory of monopsonistic competition. *Economic Journal* **109**, 190–203.

Blank, D. M. and Stigler, G. J. (1957). *The demand and supply of scientific personnel.* Cambridge, MA: National Bureau of Economic Research.

Boal, W. M. (2009). The effect of minimum salaries on employment of teachers: A test of the monopsony model. *Southern Economic Journal* **75**, 611–638.

Boal, W. M. and Ransom, M. R. (1997). Monopsony in the labor market. *Journal of Economic Literature* **35**, 86–112.

Buerhaus, P. I., Auerbach, D. I. and Staiger, D. O. (2009). The recent surge in nurse employment: Causes and implications. *Health Affairs* **28**, w657–w668.

Bunting, R. L. (1962). *Employer concentration in local labor markets.* Chapel Hill, NC: University of North Carolina Press.

Burdett, K. and Mortensen, D. T. (1998). Wage differentials, employer size and unemployment. *International Economic Review* **39**, 257–273.

Card, D. and Krueger, A. B. (1995). *Myth and measurement: The new economics of the minimum wage.* Princeton, NJ: Princeton University Press.

Currie, J., Farsi, M. and Macleod, W. B. (2005). Cut to the bone? Hospital takeovers and nurse employment contracts. *Industrial and Labor Relations Review* **58**, 471–493.

Devine, E. (1969). Manpower shortages in local government employment. *The American Economic Review* **59**, 538–545.

Falch, T. (2010). The elasticity of labor supply at the establishment level. *Journal of Labor Economics* **28**, 237–266.

Feldman, R. and Scheffler, R. (1982). The union impact on hospital wages and fringe benefits. *Industrial and Labor Relations Review* **35**, 196–206.

Greenhouse, S. (2006). Suit claims hospitals fixed nurses 'pay. *New York Times, NY, NY.* June 21.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**(1), 1–12.

Hirsch, B. T. and Schumacher, E. J. (1995). Monopsony power and relative wages in the labor market for nurses. *Journal of Health Economics* **14**, 443–476.

Hirsch, B. T. and Schumacher, E. J. (2005). Classic or new monopsony? Searching for evidence in nursing labor markets. *Journal of Health Economics* **24**, 969–989.

Hurd, R. W. (1973). Equilibrium vacancies in a labor market dominated by non-profit firms: the 'shortage' of nurses. *The Review of Economics and Statistics* **55**, 234–240.

Kleiner, M. M., Park, K. W. (2010). Battles among licensed occupations: Analyzing government regulations on labor market outcomes for dentists and hygienists. *National Bureau of Economics Research Working Paper*. Cambridge, MA: National Bureau of Economic Research.

Landon, J. H. and Baird, R. N. (1971). Monopsony in the market for public school teachers. *American Economic Review* **61**(5), 966–971.

Levenstein, M. and Suslow, V. (2006). What determines cartel success? *Journal of Economic Literature* **44**, 43–95.

Link, C. R. and Landon, J. H. (1975). Monopsony and union power in the market for nurses. *Southern Economic Journal* **41**, 649–659.

Machin, S. and Manning, A. (2004). A test of competitive labor market theory: The wage structure among care assistants in the South of England. *Industrial and Labor Relations Review* **57**, 371–385.

Manning, A. (2003). *Monopsony in motion*. Princeton, NJ: Princeton University Press.

Manning, A. (2006). A generalised model of monopsony. *The Economic Journal* **116**, 84–100.

Manning, A. (2011). Imperfect competition in the labor market. *Handbook of Labor Economics* **4B**, 973–1041.

Matsudaira, J. D. (2010). Monopsony in the low-wage labor market? Evidence from minimum nurse staffing regulations. (in press).

May, J. H., Bazzoli, G. J. and Gerland, A. M. (2006). Hospitals' responses to nurse staffing shortages. *Health Affairs* **25**, W316–W323.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. London: University of Chicago Press.

Miles, J. (2007). The nursing shortage, wage-information sharing among competing hospitals, and the anti-trust laws: The nurse wages antitrust litigation. *Houston Journal of Health Law and Policy* **7**, 305–378.

Pauly, M. V. (1969). Discussion. *American Economic Review* **59**, 565–567.

Pigou, A. C. (1924). *The economics of welfare*, 2nd ed. London: MacMillan.

Reynolds, L. G. (1946). The supply of labor to the firm. *The Quarterly Journal of Economics* **60**, 390–411.

Robinson, J. (1933). *The economics of imperfect competition*. London: MacMillan.

Robinson, J. C. (1988). Market structure, employment, and skill mix in the hospital industry. *Southern Economic Journal* **55**, 315–325.

Rosen, S. (1970). Comment on the chronic 'shortage' of nurses: A public policy dilemma. In: Herbert, E. K. and Helen, H. J. (eds.) *Empirical Studies in Health Economics: Proceedings of the Second Conference on the Economics of Health*, pp. 390–397. Baltimore: Johns Hopkins Press.

Salop, S. C. (1979). Monopolistic competition with outside goods. *Bell Journal of Economics* **10**, 141–156.

Shields, M. (2004). Addressing nurse shortages: What can policy makers learn from the econometric evidence on nurse labour supply? *The Economic Journal* **114**, F464–F498.

Staiger, D., Spetz, J. and Phibbs, C. (2010). Is there monopsony in the labor market? Evidence from a natural experiment. *Journal of Labor Economics* **28**, 211–236.

Stigler, G. (1964). A theory of oligopoly. *The Journal of Political Economy* **72**, 44–61.

Sullivan, D. (1989). Monopsony power in the market for nurses. *Journal of Law and Economics* **32**, S135–S178.

Yett, D. E. (1970). The chronic 'shortage' of nurses: A public policy dilemma. In: Herbert, E. K. and Helen, H. J. (eds.) *Empirical Studies in Health Economics: Proceedings of the Second Conference on the Economics of Health*, pp. 357–389. Baltimore: Johns Hopkins Press.

Moral Hazard

T Rice, University of California, Los Angeles, Los Angeles, CA, USA

Introduction

The term 'moral hazard' is surely one of the most controversial in the field of health economics. Although it would seem that the connotation must be pejorative – immorality is certainly implied if one is prey to the hazard (Dembe and Boden, 2000) – it is commonly used to describe a much more benign situation in which a person with health insurance will use more services. Indeed, as Pauly (1968) pointed out long ago, "the response of seeking more medical care with insurance than in its absence is a result not of moral perfidy, but of rational economic behavior" (p. 535).

In reality, though, health economists have not viewed the concept agnostically. Moral hazard has been linked inextricably to another concept: the welfare loss from excessive health insurance. This, too, is value-laden, as it implies that social welfare would be higher if people did not have so much insurance. Some estimates put the cost of this welfare loss as high as 30% of all spending on health care in the US (Manning *et al.*, 1987; Feldman and Dowd, 1991). Evidence on the existence of moral hazard has led to increasing patient cost-sharing for health care services (Manning *et al.*, 1987; Newhouse, 1993).

How one views the concept and evidence has a profound impact on the public policies espoused – and even carried out. Because raising cost sharing requirements can be shown to reduce welfare loss under the traditional theory, this has been a policy advocated by many health economists. Others, particularly those from outside of the US, have been less sanguine about such policies as they can lead to less use of necessary care, as well as more inequity. Many countries instead rely more on quelling unnecessary utilization by providing incentives to providers rather than demanders of care.

This article is organized as follows. It begins with an explication of the traditional economic theory of moral hazard. Next, it provides some challenges to this theory. After that, empirical evidence is provided, first, from the RAND Health Insurance Experiment (HIE) and a critique of it, as well as some more recent studies. It then raises a topic of some currency: the advisability of evidence-based cost sharing based on the value that services convey. Finally, alternative ways of controlling the use of unnecessary care are presented that focus on the supply rather than demand side of the health care market. The article concludes with a call for less value-laden terminology.

Traditional Economic Theory

Before addressing moral hazard, it is useful to consider the traditional concept of consumer demand more broadly. If some key assumptions – for example, consumers are rational and well-informed – are deemed to be true (or are ignored), then what people demand (that is, what they are willing to pay for goods at different prices) is a barometer of social welfare. This is because in asserting these demands, they 'reveal themselves' to prefer one set of goods over another. It is a short leap to conclude that for society as a whole, whatever people choose will make society best off.

Not everyone, of course, agrees that demand curves can be used in such a way. American economists Ellis and McGuire (1993) take a much less value-laden approach, asserting that, "[W]e are skeptical that the observed demand can be interpreted as reflecting 'socially efficient' consumption, [so] we interpret the demand curve in a more limited way, as an empirical relationship between the degree of cost sharing and quantity of use demanded by the patient" (p. 142). Nevertheless, not only is that the first interpretation by far the most common one, but it underlies the entire notion of welfare loss discussed below.

To understand that theory it is useful to begin with the concept of 'consumer surplus.' This is defined as "[t]he difference between what a consumer pays for a good or service and the maximum they would pay rather than go without it" (Culyer, 2010). The former is set by the marketplace, the latter by the consumer's own preferences. To illustrate, suppose a pound of apples costs US$2 and a consumer is willing to buy 4 lb at that price. This fourth pound, however, is probably of less value to him or her than are the previous pounds (unless a pie is being baked requiring that much). This is because of another economics concept, 'diminishing marginal utility.' In fact the consumer might be willing to pay US$5 for the first pound, US$4 for the second, and US$3 for the third. Fortuitously, they do not have to, as the market price is only US$2. As a result, in this example they have generated US$6 worth of consumer surplus: for each pound of apples, the difference between how much they are willing to pay and how much they actually have to pay. The term, incidentally, was first used in the mid-nineteenth century by a French engineer named Jules Dupuit as a way of calculating the value of railroad bridges (Ng, 1979) (A history of Dupuit's contribution – and notably, the lack of contribution by John Marshall, who popularized the concept to the English-speaking world, can be found in Houghton (1958)).

Public policymakers are not very interested in the individual consumer as they are in the aggregation of all consumers. By summing up the consumer surplus, we can derive the value to society of a particular commodity or investment over and above its costs. This is useful to know in and of itself, but also can help policymakers choose among alternative projects in which to invest.

Pauly (1968) focused on the concept of moral hazard in critiquing a famous article by Kenneth Arrow (1963). Although Arrow raised the issue, he nevertheless argued, "The welfare case for insurance policies of all sorts is overwhelming. It follows that the government should under-take insurance in

 doi:10.1016/B978-0-12-375678-7.00916-0

those cases where this market, for whatever reason, has failed to emerge'' (p. 961).

Pauly showed that this is not necessarily the case because it fails to take into account moral hazard, which can chip away at consumer surplus. In essence, with full insurance, people would demand more services, even ones that had only marginal value. Because these services would cost (perhaps) as much to produce as others, society would suffer a welfare loss from this excessive amount of health insurance coverage. The welfare loss would equal the difference between how much it cost to produce the services and how much people were willing to pay for them. Suppose that a medical service cost $10, and a person would be willing to pay that much for up to three doctor visits per year. If, however, they had full insurance and had to pay nothing, they might demand six visits. Suppose for the fourth visit they would be willing to pay US$7, the fifth US$4, and the sixth US$1 (each still cost US$10 to produce). The sum of the welfare loss would be US$3 + 6 + 9 = US$18.

Because people use more services when they have full insurance, it costs more to provide medical care than it would otherwise. Pauly's point with regard to Arrow's comment is critical. Arrow said that government should provide insurance if it is not available. Pauly shows that this is not necessarily true: people will have to pay (in taxes) for the insurance program, but much of the spending will go toward services that they would not have chosen to purchase in lieu of insurance – services, he argued, are of less value by definition. Stated more bluntly, the individual, and therefore society as a whole, could very well be better off with no insurance than with government-provided insurance, due to the concept of moral hazard. Or as Robert Evans (1984) states disparagingly of this line of reasoning, ''The welfare burden is minimized when here is no insurance at all'' (p. 49).

The word 'could' in the previous paragraph is there advisedly. Although Pauly argues that there is a welfare loss to health insurance, there is also a gain: people obtain utility from being protected against large medical expenses. The issue, then, is determining which is larger: the welfare gain from this security, or the welfare loss described above. Feldman and Dowd (1991) took both elements into account, and concluded that the loss was far greater than the gain.

The policy implication that is generally taken away from this analysis is that consumers should share in the cost of services, or, put more graphically, 'have some skin in the game.' Patient cost sharing will reduce service usage; it is assumed that the services that are forgone will be those that bring the lowest utility (a concept returned in the section The RAND Health Insurance Experiment). Although the RAND HIE has not been discussed yet, its authors touted the societal savings that they argue were generated by the uptick in cost-sharing requirements in the US that followed publication of the study results. The study cost US$285 million in 2010 dollars; they argue that this cost was made up in only a week from savings that resulted from the lower costs associated with the increased cost sharing (Manning *et al.*, 1987).

Before going on, it needs to be pointed out that the discussion in this article focuses on 'ex post moral hazard.' This is the phenomenon that occurs when the out-of-pocket price of medical care is reduced through the possession of insurance, such that the quantity of services demanded subsequently increases. There is another type of moral hazard, known as *ex ante*. According to Culyer, this ''refers to the effect that being insured has on behavior, generally increasing the probability of the event insured against occurring'' (p. 331). For example, if you are insured you may be less likely to engage in preventive behaviors – or may take up skydiving – because of the financial protection afforded by insurance. Because *ex ante* moral hazard has received much less consideration in the health care literature, it is not discussed further here. It is more salient in other types of insurance, such as for fires. By possessing such insurance, business and homeowners may take less care in taking care of electrical wiring, installing fireproofing, etc.

Challenges to the Traditional Theory

Although it is probably fair to say that most health economists are largely comfortable with the traditional theory moral hazard, there have been both direct objections as well as indirect ones. The former concern the issue of whether there is substantial welfare loss from health insurance, and the latter relate to the notion that substantial patient cost sharing is an advisable policy.

One objection raised by the present author (Rice, 1992; Rice and Unruh, 2009) relates to the notion that one can derive accurate estimates of social welfare from traditional methods. The way in which welfare losses are calculated assumes that individuals can accurately predict (at least on average) the benefits they will derive from using a medical service. They then compare this to the cost that they have to pay, and make a decision about whether such a service is worth purchasing. If they cannot predict these benefits accurately, then the method of ascribing welfare loss to excessive health insurance is invalid.

Why is this the case? Recall from above that welfare loss is defined as the difference between how much it costs to produce the services and how much people were willing to pay for them. How much people are willing to pay is defined by the demand curve, which shows, at all hypothetical prices, how many of an item a consumer will purchase. The traditional theory assumes that what people are willing to pay is an accurate measure of how much something is worth to them, or, which can call 'utility' or 'welfare.' It assumes that people know the benefit they will derive from a service – before purchasing it – and therefore can compare it to the cost to make a purchase decision that is in their best interest.

Consider the following. A person has a number of health ailments that include an ear infection and throat pain. Treating each will involve visiting a physician, so the out-of-pocket costs are the same. Researchers, however, probably unbeknownst to the person, have found that medical care has been shown to be highly effective in treating the ear infection, but rarely effective in treating throat pain (Lohr *et al.*, 1986). It is logical to assume that a person would get more utility from the ear treatment and therefore would be willing to pay more for it – perhaps even the full price even in the absence of having insurance (This assumption is, admittedly, somewhat controversial. It may be that consumers are not interested in

the conclusions of medical researchers but instead trust their own judgments in these matters. Here a different view is taken – that consumers would generally prefer to pay more for services that are judged to be more effective by medical research). In contrast, they would perhaps be willing to pay only the cost sharing amount – which is far less than the total cost of the service – to be treated for the cough. If the person actually behaved this way, then the welfare loss calculations would appear to be valid.

In reality, however, consumers are often unaware of which service will be more useful to them. If so, then examining what services they demand when they have to pay the full price, versus what they demand when they are insured, does not provide an indication of the utility or welfare derived. In the parlance of economists, the author is positing here that the demand curve does not necessarily reflect utility or welfare when consumers do not have good information about the benefits and costs of alternative services. Empirical evidence will be examined on this issue below. To give a preview, there is some evidence to suggest that when facing cost-sharing requirements, patients cut back on service usage somewhat indiscriminately, equally reducing use of services that are deemed by experts to be most and least useful. Moreover, there is growing evidence that cost sharing result in forgoing needed services.

A second objection to the welfare loss theory has been propounded by John Nyman (1999, 2002, 2007). The traditional model of welfare loss assumes that the only benefit of insurance is that (a risk-averse) people will receive utility from the financial protection afforded by insurance. Nyman, however, asserts that there is a yet more important aspect of insurance to purchasers: it allows them to be able to afford very expensive medical procedures that, in lieu of having insurance, they would not be able to obtain. If this is the reason why people use more services when they are insured, then, he argues, there is a welfare gain rather than a welfare loss to the additional utilization that occurs when a person is insured.

Nyman provides a hypothetical example. Assume that a mastectomy costs US$20 000 and breast reconstructive surgery, another US$20 000; the total cost of care for this episode of illness is therefore US$40 000. Further assume that an uninsured woman who has breast cancer can afford the US$20 000 surgery, but does not have the resources to pay for the reconstruction. Compare that to a second hypothetical situation, where the woman is insured and therefore can afford the mastectomy and the reconstruction. Under the conventional theory there would be a welfare loss associated with the reconstruction, because the woman only demanded it when having insurance made it cheaper.

According to Nyman, it is not the reduction in price brought about by insurance, but rather the increase in effective income that generates the demand for reconstructive surgery. In effect, having insurance has increased the woman's income by making a heretofore unaffordable service, affordable. The woman, in turn, chooses to spend this new wealth on the reconstruction. When the insurance company wrote her a US$40 000 check, and the woman chose to use the money toward not only the mastectomy but also the reconstruction instead of spending it on something else, then the purchasing behavior is evidence of a welfare gain (Nyman 2007).

A third objection to the welfare loss theory is also ethical in nature but more general. As noted, the major policy implication of the welfare loss is that cost sharing (compared to free care) will increase social welfare. Two concerns rise from this. The first relates to the distribution of income; cost sharing is highly regressive, falling most heavily on those with low incomes. Moreover, the poor tend to be sicker and, if they avoid care due to its costs, are more likely to suffer the consequences of unchecked illness. This is well summarized by Evans *et al.* (1993), who wrote:

> [P]eople pay taxes in rough proportion to their incomes, and use health care in rough proportion to their health status or need for care. The relationships are not exact, but in general sicker people use more health care, and richer people pay more taxes. It follows that when health care is paid for from taxes, people with higher incomes pay a larger share of the total cost; when it is paid for by the users, sick people pay a larger share.... Whether one is a gainer or loser, then, depends upon where one is located in the distribution of both income... and health.... In general, a shift to more user fee financing redistributes net income... from lower to higher income people, and from sicker to healthier people. The wealthy and healthy gain, the poor and sick lose (p. 4).

There is a final objection to relying on patient cost sharing, as implied by the welfare loss theory. If patient cost sharing defines efficiency by reducing welfare loss, this implies that the US has the most efficient health care system in the world (or is second to Switzerland, which also has substantial cost sharing). Although this is not the place to review the evidence, the assertion that the US health care system is among the most efficient in the world is hard to justify given the far higher costs, but mediocre at best process and outcome indicators that are available from international comparative research (Rice and Unruh, 2009, ch. 10).

Evidence

The RAND Health Insurance Experiment

The RAND HIE was the most important empirical study done on the demand for medical care. It also provided evidence of the impact of cost sharing not only on use of services but also on patient health status. Researchers have used the results on moral hazard to estimate the welfare loss from excess health insurance.

Conducted between 1974 and 1982, approximately 5800 individuals in six sites (in a total of four US states) were randomized into groups that faced different cost sharing requirements. Although the actual experimental design was somewhat more complicated, the main intervention tested concerned cost sharing. Participants were assigned to pay 0%, 25%, 50%, or 95% of their medical care expenses. There were also maximums associated with how much they would have to pay each year.

The study's findings with regard to use of services and costs showed that cost sharing indeed had a substantial impact. Those who received free care spent, on average, US$750 annually, compared to US$617 for those paying 25% of costs,

US$573 for responsible for 50%, and US$540 for those paying 95% (in 1984 dollars) (Manning *et al.*, 1987). Most income groups behaved similarly, as did those who were healthy versus sick. The reductions were similar for children and adults for outpatient services. However, cost sharing did not deter inpatient utilization for children. Of note is that nearly all of the impact of cost sharing was on seeking care in the first place. Once a person entered the medical care system for an episode of illness, it did not have a marked effect on usage.

With one or two exceptions, these results were not surprising. What was surprising was that for the most part, those who paid more and used less did not experience a reduction in health status. This was measured in numerous ways, including self-assessed health status, physical functioning, role functioning, health perceptions, and mental health (Brook *et al.*, 1983). There were few exceptions to this for the sample as a whole (mainly slightly higher blood pressure and lower corrected vision). Those already at elevated risk of dying were also adversely affected, mainly due to the impact of cost sharing on blood pressure (Brook *et al.*, 1983). If there was one group that did benefit from free care, it was those with low incomes. Their risk of dying was lower and they experienced fewer serious symptoms (Shapiro *et al.*, 1986). Free care also led poorer individuals to obtain more medical examinations (Lohr *et al.*, 1986).

Interestingly, another finding by the RAND researchers was that those facing higher copayments were rather indiscriminant in their reduction of services (Lohr *et al.*, 1986). They categorized services into four groups: highly effective, quite effective, less effective, and rarely effective. The authors concluded that "cost sharing was generally just as likely to lower use when care is thought to be highly effective as when it is thought to be only rarely effective" (p. S32) and that "cost sharing did not lead to rates of care seeking that were more 'appropriate' from a clinical perspective. That is, cost sharing did not seem to have a selective effect in prompting people to forego care only or mainly in circumstances when such care probably would be of relatively little value" (p. S36). This was essentially the finding of another aspect of the experiment, which looked at the effect of coinsurance on the appropriateness of hospitalization (Siu *et al.*, 1986).

Although generally viewed as the seminal study in the area of demand, there are a number of caveats that need to be kept in mind:

- The study's results are from 30 years ago. Much has changed since then, including a dramatic drop in hospital usage in the US. Moreover, there has been a shift from fee-for-service to managed care. Managed care implies that not only the patient and their cost sharing requirements, but the health plan itself, is involved in determining which services are used.
- The study did not examine the impact of uninsurance. Everyone who participated in the study was assigned an insurance plan, so such comparisons were not possible.
- Seniors were also excluded so the results do not apply directly to them.

Some more controversial concerns have also been raised. The first concerns the internal validity of the experiment. Nyman (2007) points out that those individuals who were assigned to cost sharing were much more likely to drop out of the experiment than those assigned to free care, presumably because they did not like the prospect of facing higher expenditures.(The experiment was designed so that no one could be made worse off financially by participating, but this might not have been clear to participants who noted that they were paying 50% or 95% of their medical costs.) Just half a percent of those who were assigned to free care dropped out compared to seven percent of others. He contends that this could not only bias the results on service usage but also the health status results. If those who dropped out had stayed in, their health would have been more likely to have been adversely affected because, facing higher coinsurance, they would likely have forgone needed medical care. These criticisms were not taken lightly by the researchers who conducted the experiment, who contended that it was not in people's best interest to leave the experiment and that other factors provide more likely explanations for the differential drop-out rate. They also contend that those dropping out would have to have had remarkably different hospitalization rates than those who stayed in the study (Newhouse *et al.*, 2008). At the time of writing there does not appear to be a consensus in the literature on these issues.

A second criticism relates to external validity. Although cost sharing may reduce patients' demand, it is possible that the impact on overall utilization will be less. Consider that the experiment included, at most, 2% of the people in a geographic area – and those with considerable cost sharing, perhaps half that. This means that the experiment would have had almost no impact on the behavior of physicians and hospitals. In reality, though, if cost sharing were increased dramatically, suppliers would likely respond to reduced demand by trying to generate some more, to compensate. This implies that one cannot take the results from individuals and apply them to the population as a whole (Rice and Unruh, 2009). This criticism was raised at the outset of the experiment (Hester and Leveson, 1974), although the researchers conducting the experiment contended that the study was not "designed to replicate what would happen if various health insurance proposals were enacted into law" and that they "deliberately selected sites that vary considerably with respect to the amount of stress on the delivery system" (Newhouse, 1974, pp. 236–237).

More Recent Evidence

What is striking about the most recent evidence from the US is that it does tend to show that higher cost sharing reduces health status. It is important to note, however, that unlike the HIE, these studies are not based on true experimental designs. A few such studies are noted below:

- Trivedi *et al.* (2008), focusing on the appropriate use of mammograms, examined Medicare beneficiaries aged 65–69 years in managed care plans, a group excluded from the HIE sample. In the period from 2001 to 2004, many more plans required modest copayments (US$10 or a coinsurance of 10%). The authors found that not only cost sharing reduced screening rates by 8.3 percentage points compared to those with full insurance coverage, but that

the effect "was magnified among women residing in areas of lower income or educational levels. Screening rates decreased by 5.5% points in plans that instituted cost sharing and increased by 3.4% points in matched control plans that retained full coverage" (p. 375).

- Much research has been conducted on the appropriate use of prescription drugs. Looking again at Medicare beneficiaries, Rice and Matsuoka (2004) report on five studies where it was possible to directly assess the impact of cost sharing on mortality, and 15 others where health status effects could be inferred by examining the appropriate use of medications. In two of the five studies examining mortality, cost sharing led to higher incidence of death; in three there were no effects. Of the other 15 studies, 12 found evidence that cost sharing led to less usage of appropriate medications, and three found no effects.
- Studies have also been conducted on younger populations. For example, a study of more than half a million employees from 30 employers found that doubling of copayments reduces use of nonsteroidal antiinflammatory drugs (NSAIDS) by 45% and antihistamines by 44%. The authors conclude that, "significant increases in copayments raise concern about adverse health consequences because of the large price effects, especially among diabetic patients" (Goldman *et al.*, 2004, p. 2344). Indeed, among the diabetics, a doubling of copayments reduced their use of medications by 23%.
- In a similar vein, a study from a single large employer found that enrollees in a high-deductible plan reduced substantially their filling of prescriptions for blood pressure and cholesterol medication (Greene *et al.*, 2008). Furthermore, a study of examining an increase in copayments from US$2 to US$7 in the Veterans Administration found large reductions in drug adherence for cholesterol medication, even among those at high coronary risk (Doshi *et al.*, 2009).

In sum, these studies further impugn the notion that charging patients more increase social welfare. The author concludes this article by considering two alternatives: tailoring cost-sharing to medical effectiveness, and focusing on the supply rather than demand side of the health care marketplace.

Tailoring Cost-Sharing to Medical Effectiveness

Although the traditional economic model calls for patient cost sharing as a way of reducing moral hazard and the concomitant reduction in societal welfare, it has been suggested that there are a variety of reasons – both conceptual and empirical – that cast doubt on this interpretation. In the last section the author will address larger policy alternatives that rely on controlling service use by relying on suppliers. Here, examination is made of one other demand-side policy that is now receiving much attention: tailoring cost-sharing to the medical effectiveness of services.

This is commonly called value-based insurance design (VBID). According to Chernew and colleagues (2007), under VDIB, "cost sharing is still put to use, but a clinically sensitive approach is explicitly adopted to mitigate the adverse health consequences of high out-of-pocket spending" (Chernew *et al.*, 2007, p. W196). Fendrick *et al.* (2010) write, "[t]he basic premise … is to align out-of-pocket spending with the value of medical services" (p. 2017). Cost sharing requirements can, in theory, be the same for everyone, or tailored to the individual. The latter, although probably more effective in matching reduced cost-sharing to medical need, is much more administratively cumbersome as well as difficult for the individual patient to understand. As Robinson (2010) notes, "[t]ailoring benefit design to differences among patients will depend on the development of reliable diagnostic tests that can identify ex ante which products will be effective for which patients" (p. 2012). He suggests that VBID move away from low-cost preventive services and chronic medications to surgery, specialty drugs, implantable medical devices, and imaging services, which "constitute the new frontier for insurance design and require that value principles…" (p. 2015).

Thus far, VBID programs have focused more on reducing cost sharing for high-value and preventive services rather than raising them for low-value services. Although this does encourage more appropriate utilization, it may not be cost-saving and therefore could be unsustainable (Fendrick *et al.*, 2010). To illustrate, one US company, Pitney Bowes, eliminated cost sharing requirements for cholesterol drugs and for a blood clot inhibitor; the policy did indeed improve drug adherence (Choudhry *et al.*, 2010). On a larger scale, a large insurer, Blue Cross Blue Shield of North Carolina, showed similar results when it eliminated copayments on generic drugs and reduced them on selected brand-name drugs (Maciejewski *et al.*, 2010).

Supply-Side Policies

An alternative to focusing on moral hazard on the patient is to instead focus on the suppliers of care. This has several potential advantages, including a greater potential to control costs, the ability of experts to target which services should be encouraged, and less distributional impact than demand-side policies, which focus on ability to pay.

Just as there can be cost sharing on the demand side of the market, Ellis and McGuire (1993) argue that there is an analogous concept, supply-side cost sharing, "which seeks to alter the incentives of health care providers to provide certain services" (p. 135). Examples they list include the use of a fixed payment per hospital stay (e.g., diagnosis-related groups or DRGs) and even more broadly, the use of Health Maintenance Organizations (HMOs) rather than fee-for-service medicine. Both DRGs and HMOs focus on influencing the behavior of the provider rather than the patient.

Supply-side policies have the potential of improving social welfare by focusing not on supposed overinsurance, but rather on encouraging the use of appropriate services and discouraging inappropriate ones – that is, removing some of the waste in the medical care system, which has been estimated to be up to 30% of services used (Leape 1989; Schuster *et al.*, 1998). Much effort is being expended on this, through comparative effective research and the dissemination of practice guidelines.

Although demand-side policies are certainly in vogue now, it is still true that most policies worldwide – including in the US – focus on suppliers. Common supply-side policies such as the movement away from fee-for-service payment of physicians, incentivizing providers to provide high-quality care and avoid wasteful procedures, utilization management, and global budgets, are aimed directly at providers, not patients. It is not that they eschew patient input but the focus is clearly elsewhere. If, as is argued here, these strategies are designed to reduce waste, then it can be argued that nearly all nations act as though the waste in medical care is more through the provision of unnecessary services, and not much through excess demand stemming from overinsurance.

Conclusion

At the beginning, the author noted Pauly's comment that it is rational economic behavior rather than 'moral perfidy' that drives people to seek more care when they are insured. Despite this, there is undoubtedly a lingering effect to term 'moral hazard' that influences economists' views on the subject, and perhaps therefore, their policy prescription. In his Dictionary of Health Economics, Culyer (2010) argues, "before rushing to the conclusion that moral hazard must be controlled through coinsurance, copayments and other forms of rationing, it needs to be borne in mind that there may be reasons for wanting individuals to consume more care. Even more fundamentally, there may be reasons for entirely eschewing the idea that the demand curve reveals anything worth knowing about the value placed on health care. In that case, even if the behavioral account given of moral hazard may still stand, the ethical accusation of 'waste' fails entirely" (pp. 331–332).

Thus, one thing that might be helpful is coming up with a more neutral term for the concept. The phenomenon one wants to capture in the health insurance context is similar to what Ellis and McGuire (1993) were attempting to do when redefining the demand curve to reflect nothing about social efficiency but rather simply "an empirical relationship between the degree of cost sharing and quantity of use demanded by the patient" (p. 142). The concept to be captured here is the additional use of service as a result of possession of insurance. As such, one value-neutral term could be 'insurance-driven utilization.'

Indeed, analyzing the history of moral hazard in the context of the workers compensation field, Dembe and Boden (2000) also call for the use of less value-laden terms, "[u]nless economics intend to pass judgment on the moral conduct of system participants…" (p. 273). They further state that, "Attention is focused on the costs of increasing benefits and not on the adequacy of those benefits. Insurance is characterized as leading to more time lost from work, not as providing a valuable buffer against the economics stresses resulting from workplace injuries and illnesses. Recipients of workers' compensation benefits are characterized as engaging in malingering or fraudulent behavior and are thus classified as undeserving of those benefits. They are not characterized as hard-working individuals who have suffered an injury and who may nevertheless receive inadequate benefits from their insurance carrier" (p. 274).

The same argument can be made in the field of health economics. Charactering additional utilization that comes about from possessing health insurance in nonvalue-laden terms can widen the scope of policy options beyond simply charging people more, and provide a more positive view of the benefits that people derived from health insurance.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance in Developed Countries, History of. Private Insurance System Concerns. Social Health Insurance – Theory and Evidence

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5), 940–973.

Brook, R. H., Ware, Jr., J. E., Rogers, W. H., et al. (1983). Does Free Care Improve Adults' Health? *New England Journal of Medicine* **309**(23), 1426–1434.

Chernew, M. E., Rosen, A. B. and Fendrick, A. M. (2007). Value-based insurance design. *Health Affairs* **26**(2), W195–W203.

Choudhry, N. K., Fisher, M. A., Avorn, J., et al. (2010). At Pitney Bowes, value-based insurance design cut copayments and increased drug adherence. *Health Affairs* **29**(11), 1995–2001.

Culyer, A. J. (2010). *Dictionary of health economics*. Cheltenham, UK: Elgar.

Dembe, A. E. and Boden, L. I. (2000). Moral hazard: A question of morality? *New Solutions* **10**(3), 257–279.

Doshi, J. A., Zhu, J., Lee, B. L., Kimmel, S. E. and Volpp, K. E. (2009). Impact of a prescription copayment increase on lipid-lowering medication adherence in veterans. *Circulation* **119**, 390–397.

Ellis, R. P. and McGuire, T. G. (1993). Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives* **7**(4), 135–151.

Evans, R. G., Barer, M. L. and Stoddard, G. L. (1993). The truth about user fees. *Policy Options* **14**, 4–9.

Feldman, R. and Dowd, B. (1991). A new estimate of the welfare loss of excess health insurance. *American Economic Review* **81**(1), 297–301.

Fendrick, A. M., Smith, D. G. and Chernew, M. E. (2010). Applying value-based insurance design to low-value health services. *Health Affairs* **29**(11), 2017–2021.

Goldman, D. G., Joyce, G. F., Escarce, J. J., et al. (2004). Pharmacy benefits and the use of drugs by the chronically ill. *Journal of the American Medical Association* **291**(19), 2344–2350.

Greene, J., Hibbard, J., Murray, F., Teutsch, S. M. and Berger, M. L. (2008). The Impact of consumer-directed health plans on prescription drug use. *Health Affairs* **27**(4), 1111–1131.

Hester, J. and Leveson, I. (1974). The health insurance study: A critical appraisal. *Inquiry* **11**(1), 53–60.

Houghton, R. W. (1958). A note on the early history of consumer's surplus. *Ecnomica* **25**(97), 49–57.

Leape, L. (1989). Unnecessary surgery. *Health Services Research* **24**(3), 351–407.

Lohr, K. N., Brook, R. H., Kamberg, C. J., et al. (1986). Effect of cost sharing on use of medically effective and less effective care. *Medical Care* **24**(supplement), S31–S38.

Maciejewski, M. L., Farley, J. F., Parker, J. and Wansink, D. (2010). Copayment reductions generate greater medication adherence in targeted patients. *Health Affairs* **29**(11), 2002–2008.

Manning, W. G., Newhouse, J. P., Duan, N., et al. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review* **77**(3), 251–277.

Newhouse, J. P. (1974). The health insurance study: Response to Hester and Leveson. *Inquiry* **11**(3), 236–241.

Newhouse, J. P. (1993). *Free for all? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Newhouse, J. P., Brook, R. H., Duan, N., et al. (2008). Attrition in the RAND Health Insurance Experiment: A response to Nyman. *Journal of Health Politics, Policy, and Law* **32**(5), 295–308.

Ng, Y.-K. (1979). *Welfare economics*. London: Macmillan and Co.

Nyman, J. A. (1999). The value of health insurance: The access motive. *Journal of Health Economics* **18**, 141–152.

Nyman, J. A. (2002). *The theory of the demand for health insurance*. Stanford, CA: Stanford University Press.

Nyman, J. A. (2007). American health policy: Cracks in the foundation. *Journal of Politics, Policy, and Law* **32**(5), 759–783.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**(4), 531–537.

Rice, T. (1992). An alternative framework for evaluating welfare losses in the health care market. *Journal of Health Economics* **11**(1), 88–92.

Rice, T. and Matsuoka, K. Y. (2004). The impact of cost sharing on appropriate utilization and health status: A review of the literature on seniors. *Medical Care Research and Review* **61**(4), 415–452.

Rice, T. and Unruh, L. (2009). *The economics of health reconsidered*. Chicago, IL: Health Administration Press.

Robinson, J. C. (2010). Applying value-based insurance design to high-cost health services. *Health Affairs* **29**(11), 2009–2015.

Schuster, M. A., McGlynn, E. A. and Brook, R. H. (1998). How good is the quality of health care in the United States? *Milbank Quarterly* **76**(4), 517–563.

Shapiro, M. F., Ware, Jr., J. E. and Sherbourne, C. D. (1986). Effects of cost sharing on seeking care for serious and minor symptoms. *Annals of Internal Medicine* **104**(2), 246–251.

Siu, A. L., Sonnenberg, F. A., Manning, W. G., et al. (1986). Inappropriate use of hospitals in a randomized trial of health insurance plans. *New England Journal of Medicine* **315**(20), 1259–1266.

Trivedi, A. N., Rakowski, W. and Ayanian, J. Z. (2008). Effect of cost sharing on screening mammography in medicare health plans. *New England Journal of Medicine* **358**(4), 375–383.

Multiattribute Utility Instruments and Their Use

J Richardson, J McKie, and E Bariola, Monash University, Clayton, VIC, Australia

Introduction

A multiattribute utility (MAU) instrument consists of two parts: (1) a health questionnaire, and (2) a scoring formula which converts answers into an overall score. Each set of answers to the health questionnaire defines a 'health state.' The overall score reflects the strength of people's preferences for the state, and, consequently, it is a measure of the utility of the state as understood in economics.

Box 1 illustrates this. The EQ-5D MAU instrument consists of five single 'items', i.e. questions and response levels (see Box 2 on terminology). Each relates to a separate dimension of health (mobility, self care, usual activities, pain, and depression), which collectively constitute the 'descriptive system' or classification. The instrument combines these using the formula shown below the questionnaire. An individual answering level 1 for each item (1, 1, 1, 1, and 1) would obtain a utility score of 1.00; a person answering (3, 3, 3, 3, and 3) – the 'all worst' health state – would obtain a utility score of −0.594. As health states change (because of a health program), answers change, and the MAU instrument predicts a change in a person's utility. Someone answering (1, 1, 2, 2, and 3) before health care and (1, 1, 1, 2, and 2) afterwards would score 0.225 before and 0.725 after care, an improvement of 0.5.

Utility scores calculated this way may be used for economic evaluation, and, in particular, cost–utility analyses (CUA), which compare health program costs with the number of quality-adjusted life-years (QALYs) obtained. QALYs are calculated by multiplying an index of utility by years of life. The index must be measured on a scale on which 1.00 is 'best health' (as defined by the instrument) and 0.00 is the 'utility' of death. Consequently, in best health, the number of QALYs equals life years times 1.00 and therefore equals the number

Box 1 EQ-5D descriptive system

EQ-5D descriptive system

1. Mobility (MOB)
   - MOB 1: No problems walking about
   - MOB 2: Some problems walking about
   - MOB 3: Confined to bed
2. Self-Care (CARE)
   - CARE 1: No problems with self-care
   - CARE 2: Some problems washing or dressing
   - CARE 3: Unable to wash or dress self
3. Usual Activities (ACT)
   - ACT 1: No problem with performing usual activities (e.g., work, study, housework, family, or leisure activities)
   - ACT 2: Some problems with performing usual activities
   - ACT 3: Unable to perform usual activities
4. Pain/discomfort (PAIN)
   - PAIN 1: No pain or discomfort
   - PAIN 2: Moderate pain or discomfort
   - PAIN 3: Extreme pain or discomfort
5. Anxiety/Depression (DEP)
   - DEP 1: Not anxious or depressed
   - DEP 2: Moderately anxious or depressed
   - DEP 3: Extremely anxious or depressed

Combinations of answers ('Health states') $= 3 \times 3 \times 3 \times 3 \times 3 = 243$

EQ-5D Scoring formula

$$\begin{aligned}\text{Utility} = 1 - [&(0.069\ \text{MOB2} + 0.314\ \text{MOB3}) + (0.104\ \text{CARE2} + 0.214\ \text{CARE3}) \\ &+ (0.036\ \text{ACT2} + .094\ \text{ACT3}) + (0.123\ \text{PAIN2} + 0.386\ \text{PAIN3}) \\ &+ (0.071\ \text{DEP2} + 0.236\ \text{DEP3}) + (0.081\ \text{ANY(A)} + 0.269\ \text{ANY(B)})]\end{aligned}$$

where

[MOB2,... PAIN3] = 1 (or 0.00) if the respondent did (did not) tick the corresponding response level of the item ANY(A) = 1 if any level ≠ 1; ANY(B) = 1 if any level = 3

Note: The derivation of the formula and parameters (0.69, 0.314, etc) are explained in the text.

Box 2 MAU instrument-related terminology

Algorithm	(or formula) The rule for converting answers to a questionnaire into a number. It is constructed by 'scaling' a 'model'
Attribute	A characteristic or property, which an instrument seeks to describe, for example, vitality, depression, and mobility
Construct	An attribute, which is constructed or conceptualized as part of a theoretical explanation
Content	The scope and detail of the instrument's descriptive system: the behaviors, outcomes, or states, which determine an instrument's score
Descriptive system	(or descriptive 'classification'; or descriptive 'instrument'). The collection of items and dimensions, which describe the health state
Dimension	A collection of attributes with a common theme (a 'super construct'), for example, physical, mental, or social health. It usually consists of more than 1 item
Element	A single idea or attribute embodied in an item or dimension, for example, contentment or exhilaration but not contentment and exhilaration
Instrument	A questionnaire with an associated method for attaching a numerical value to the answers
Item	A linguistic statement generally consisting of a stem (e.g., 'in the last 7 days I was: …) plus a number of ordered response levels (e.g., 'always happy' … 'never happy')
Model	A conceptual or mathematical framework, which defines how values will be combined (e.g., simple or weighted averaging of the level of the item responses)
Reliability	See **Box 5**
Scaling	(or calibrating) The process of creating the algorithm for attaching numbers to health state descriptions. It requires a scaling instrument (e.g., TTO or SG) plus a model for combining the numbers produced by the scaling instrument
Sensitivity	The extent to which the instrument content allows the detection of changes in a health state
Validity	See **Box 5**

of life years. With death, life years times utility equals zero. In the example above, a health program which moves an individual from health state (1, 1, 2, 2, and 3) to (1, 1, 1, 2, and 2) for 10 years, would result in $0.5 \times 10 = 5$ QALYs, which would, in turn, be compared with the cost of the care, to obtain a cost per QALY.

Even when utility scores are not used, MAU instruments are useful for describing changes in health states over time and for comparing the health states of different individuals. In principle, the instrument can also be used to estimate the QALY-based burden of disease. However, this type of analysis has been dominated by the use of disability-adjusted life years, which combine the quality and length of life in a related but different way.

In principle, an MAU instrument can be generic, i.e., applicable to a wide range of health states, or it may be condition-specific and apply to only a specific disease. This article is about generic MAU instruments, and the term 'MAU instrument' is used here to refer to generic instruments.

Construction: The construction of an MAU instrument entails three steps. First, 'items' must be selected to create the questionnaire ('descriptive system' or 'classification'). Second, individuals are interviewed to obtain numerical data from which their utility – strength of preference – can be calculated for different health states. Third, a 'model' is used to attach values (utility scores) to all of the possible health states described by the instrument. The third step is necessary because the number of health states described by an MAU instrument is, generally, too large for the utility of each health state to be evaluated individually. Modeling is therefore used to extrapolate from measurements that are made to all possible health state values.

These steps have been approached differently by different research teams, and the scope and detail of the resulting descriptive systems varies considerably. The numerical data used for predicting utility have been obtained using different 'scaling' techniques including the time trade-off (TTO), standard gamble (SG), and the rating scale (RS). Models have employed different econometric techniques, sophisticated averaging, and a combination of these to derive a general formula for predicting utility scores from the numerical data (see section Instrument Use and Acceptance).

Box 3 Six multiattribute utility instruments and country of origin

QWB	Quality of Well-being Index	USA
15D	15 dimension instrument	Finland
EQ-5D	Originally EuroQol (RS and TTO versions)	Europe/UK
HUI	Health Utilities Index, 3 versions, HUI 1–3	Canada
SF-6D	Short form 6D (SF-6D (12) and SF-6D (36)	UK/USA
AQoL-8D	Assessment of Quality of Life (8D)	Australia

MAU instruments are flexible and easy to administer. However, they have their limitations. Their usefulness for evaluation is constrained by the content and sensitivity of the instrument's descriptive system and by the validity of the utility scores produced by the algorithm.

In the following section, six MAU instruments (see **Box 3**) are reviewed. Their chronology, characteristics, and construction are described and compared in the section History, Description and Construction of MAU Instruments. Section Instrument Use and Acceptance summarizes their use and recognition by health authorities. Different instruments produce different scores, as discussed in the section Comparison of Instruments. The reasons for this include differences in the theoretical traditions adopted in constructing the descriptive systems and scoring formula (section Theory and Evaluation) and differences in instrument content (section Construct and Content Validity). The implication of these differences for the validity of utility scores and therefore for policy is discussed in section Criterion Validity. Challenges to the field are outlined in the concluding section Conclusions. Additional

readings are suggested, which contain references supporting the present text.

History, Description and Construction of MAU Instruments

Chronology

Figure 1 summarizes the historical development of the six MAU instruments. Most writers in the area commence with a reference to the famous 1948 World Health Organization (WHO) definition of health as a 'state of complete mental and physical well-being and not merely as the absence of disease and infirmity.' This legitimized the concept of 'health' as a single construct. However, it did not provide a basis for measurement.

In the USA, the 'blueprint' for measurement was published in 1970 by Fanshel and Bush. This provided the theoretical basis for the earliest instruments, the health status index (1973), the Quality of Well-Being (QWB) (1976), and the Short Form 36 (SF-36) (1977). The latter was also the empirical basis for two later UK versions of the SF-6D developed by Brazier, one directly derived from the SF-36 (2002) and one from its reduced form, the SF-12 (2004).

The first UK instrument, the Rosser Index, was initially intended for hospital patients (1972) but was subsequently generalized to a generic 29 health state classification instrument, the 'Rosser-Kind Index' (1978). This was displaced by the EuroQol, which was created by a European consortium (The EuroQol Group) formed in 1987. The instrument was subsequently renamed as 'EQ-5D' and adopted for general use following creation of a scoring algorithm at the University of York in 1995. Earlier, Sintonen had created the 12D instrument in Finland, and the revised 15D was published immediately before publication of the EuroQol in 1989.

Three Canadian health utility instruments (HUI) were initiated by Torrance in 1982 for the evaluation of neonatal intensive care. These were modified for use in childhood cancer (HUI 2) in 1996 and further developed and scaled by Feeny for the adult population in the HUI 3 in 2002. The Assessment of Quality of Life (AQoL) instruments were developed in Australia by Richardson and Hawthorne. The AQoL-4D was published in 1997 and subsequently modified as the AQoL-6D in 2004. Additional dimensions were added to increase sensitivity for vision (AQoL-7D) in 2005 and for mental health (AQoL-8D) in 2009.

Description of Instruments

Tables 1–3 compare the six MAU instruments. Two broad conceptual approaches to description have been used (Table 1). Following the WHO typology, health problems result in impairment, disability and handicap; that is, body malfunction, limitations of body performance, and problems affecting life in a social context, respectively. Three MAU instruments (EQ-5D, SF-6D, and AQoL) have based their descriptions primarily on the last concept (i.e. health problems affecting life). By contrast, two MAU instruments (15D and HUI) have adopted a 'within-the-skin' approach (impairment/disability), although 15D was modified to include one handicap dimension. The QWB spans all concepts.

The resulting instruments have between 5 and 15 dimensions, with one item per dimension in the HUI 3, 15D, EQ-5D, and SF-6D and an average of four items per dimension in the AQoL-8D. QWB has three basic dimensions supplemented with 27 'symptom/problem' groups. Items have four to six response levels (e.g. the severity of pain or the level of mobility). Overall, items plus response levels define between 243 health states for the EQ-5D and 2.37×10^{23} for AQoL-8D. Larger instruments, particularly AQoL-8D, define numerous 'empty' states (e.g., 'bedridden' but 'no problems with self-care').

Different instruments include different dimensions (Table 1). Several are unique to a particular instrument and dimensions with similar titles include different items. Consequently, to appreciate the scope ('content') of an instrument an examination of the items is required. These are compared in Table 2, which indicates that the scope of instruments varies significantly, in part, because of the differing conceptual bases and, in part, from the level of descriptive detail contained in the items. In principle, instruments with fewer items may indirectly capture the same – or even more – information as the larger instruments by using items with broader descriptions. Alternatively, they may be omitting content to

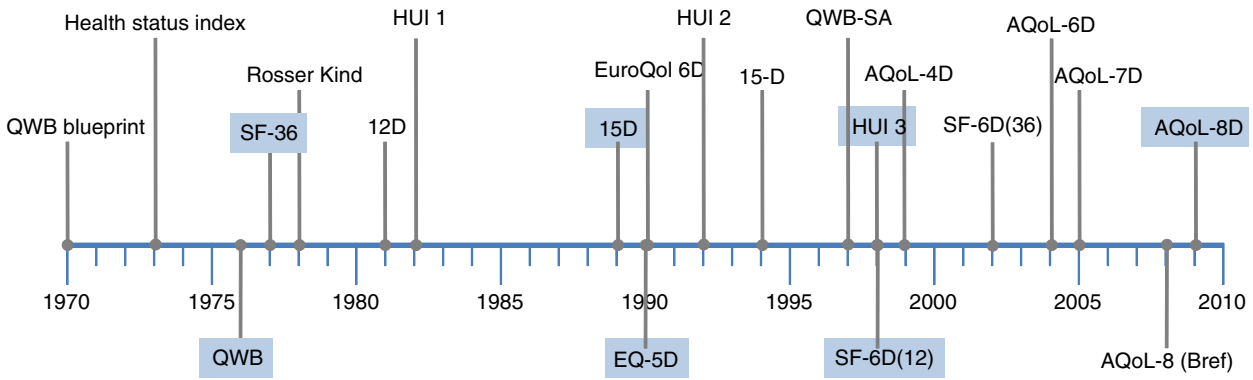


Figure 1 History of MAU instruments.

Table 1 Instrument descriptive systems

Descriptive system	*QWB*	*15D*	*EQ-5D*	*HUI 3*	*SF-6D*	*AQoL-8D*
Conceptual type	Handicap disability impairment	Disability (handicap)	Handicap (disability)	Disability	Handicap (disability)	Handicap (disability)
Selection of content	Medical literature matched with Health Interview Surveys	Medical + psychometrics	Consensus	Survey; importance ranking	SF-36, SF-6D, psychometrics	Focus groups, medical and psychometrics
Dimensions	3 + 27 symptoms/ problems	15	5	8	6	8
Items		15	5	8	6	35
Response levels	2, 3(2)	4–5	3	5–6	4–6	4–6
States defined	945	3.1×10^{10}	243	972 000	18 000	2.37×10^{23}
Completion time	na	4 min	1 min	3 min	2.5 min	5.5 min
Cronbach's α[b]	0.94	0.81	0.69	0.74–0.81		Dimensions 0.82–0.92 AQoL-8D 0.97
Test–retest (P)	0.93–0.98[a]	0.9–0.94[b]	0.73[b]	0.77[b]	0.88[b]	0.91–0.89[c]

[a]Fryback *et al.* (2010).
[b]2 months.
[c]2 weeks, 4 weeks.

Table 2 Comparison of the dimensions and content of 6 MAU instruments

		Number of symptoms (.) and items (∗)					
	Dimension	*QWB*[a]	*15D*[b]	*EQ-5D*	*HUI 3*	*SF-6D (36)*	*AQoL-8D*
Physical	Physical ability/vitality/coping/control	……..	∗			∗	∗∗
	Bodily function/self care	……	∗∗∗	∗			∗
	Dexterity				∗		
	Pain/discomfort	……	∗	∗	∗	∗	∗∗
	Senses	…..	∗∗		∗∗		∗∗
	Usual activities/work function	……...	∗	∗		∗	∗∗∗∗
	Mobility/walking	……..	∗	∗	∗		∗
	Communication	..	∗		∗		∗
Psychosocial	Sleeping	.	∗				∗
	Psychological: Depression/anxiety/anger	….	∗∗∗	∗	∗	∗	∗∗∗∗∗∗∗
	General satisfaction						∗∗∗∗
	Self-esteem						∗∗
	Cognition/memory ability	.			∗		
	Social function/relationships					∗	∗∗∗∗∗∗
	(Family) role					∗	∗
	Intimacy/sexual relationships	.	∗				∗
			15 items	5 items	8 items	12 items	35 items

[a]Symptom problem groups associated with consciousness, burns, pain, stomach, cough, fever, depression, headache, itching, talking, eyes, weight, teeth, ears, hearing, throat, breathing, sleeping, intoxication, sex, anxiety, eyeglasses, and use of medication.
[b]15D also includes breathing, sleeping, eating, elimination, and sexual activity.

achieve some other goal such as brevity. The differences are potentially important for the validity of the instruments and are discussed further below.

In addition to differences in their descriptive systems, different scaling techniques have been used to measure utilities (in particular, the TTO, RS, and SG) and different models have been used to extrapolate utility scores over the full range of health states (Table 3). Three instruments have adopted models based on MAU theory. Two have used statistical analysis and one (AQoL-8D) uses both techniques (see section Theory and Evaluation). Best health states are described differently by each of the scales, but all assign 'best health'

Table 3 Properties of the combination model and the predicted utilities

	QWB	*15D*	*EQ-5D*	*HUI 3*	*SF-6D*	*AQoL-8D*
Theory[a]	MAUT	MAUT	Statistical	MAUT	Statistical	MAUT/statistical
Model type	Additive	Additive	Additive	Multiplicative	Additive	Multiplicative/exponential
Scaling[b]	RS	RS	TTO and RS	SG/RS	SG	TTO
Best health[c]	1.00	1.00	1.00	1.00	1.00	1.00
Worst health[c]	0.320	0.11	−0.59	−0.36	0.203	−0.04
Utility at age 1[d]						
34–44	0.67[f]	0.95	0.89[f]	0.83[f]	0.80[f]	0.81[g]
60–64	0.64[f]	0.87	0.86[f]	0.80[f]	0.78[f]	0.84[i]
Test–retest[e] (correlation)	0.59[a]	Very high[c]	0.61	0.75	0.66[h]	0.89[i]

[a]MAUT, MAU theory.
[b]RS, rating scale; TTO, time trade off; and SG, standard gamble.
[c]Best/worst health utilities which are theoretically possible in the model.
[d]Values predicted for the general population.
[e](Intraclass) correlation between scores obtained.
[f]US data $n=462$ (35–44); 965 (65–74) (Fryback *et al.*, 2010).
[g]Australian data $n=225$ (35–44); 340 (60+) (Hawthorne *et al.*, 2001).
[h](Intraclass) correlation between scores obtained after 5 months.
[i](Intraclass) correlation between scores obtained after 1 month.

a numerical value of 1.00. This implies that 1.00 corresponds with different levels of real utility. The utility of the worst health state varies in the instruments from 0.32 (QWB) to −0.59 (EQ-5D), similarly implying differences in the numerical scale.

Instrument Construction

QWB Index: The QWB descriptive system was derived from the Health State Index Questionnaire. Items for this were selected from 343 'core descriptions' (items) derived from the literature and from existing health surveys.

The three multiresponse items of the QWB (mobility, social, and physical activity) define 47 health states. In combination with 27 symptom/problem groups, this rises to 945 states (**Table 1**). Although these contain no explicit mental health components, the instrument has been used for patients with psychiatric problems as the general items are sensitive to psychiatric problems. Items were scaled using RS responses from the general population of San Diego ($n=866$). An additive model was used in which the disutility from each dimension and from the worst symptom is subtracted from 1.00 (the utility of full health). The distribution of scores for the general population is approximately normal – bell-shaped – with responses distributed symmetrically around a central point. Perfect scores are rare and there are neither significant ceiling nor floor effects – that is, the instrument is sensitive at both ends of the value scale and can discriminate between states close to full health and between very poor health states.

QWB was the first MAU instrument. Originally administered by trained interviewers, a self-administered version (QWB^{SA}) was created in 1997. Translations exist into Spanish, German, Italian, Swedish, French-Canadian, and Dutch. Information and the user manual may be obtained at https://hoap.ucsd.edu/qwb-info/.

15D: The descriptive system of the 15D is based on a review of Finnish health policy documents. Scores were obtained from a sample of the Finnish population. The instrument has 15 items, 14 relating to disability (mobility, mental function, etc.) and one to handicap ('usual activities'). The 1981 version was revised following feedback from the medical profession in 1986 and again in 1992 following further user feedback and factor analysis. Utilities were obtained using a RS. Each level of each dimension is given a value and each dimension given an importance weight. Utility is calculated by adding the weighted dimension scores together.

Five separate models were subsequently used to re-estimate utilities. These used published econometric formulae to convert RS values into 'utility' scores ($n=2500$). Results demonstrated convergent validity of 15D values (i.e., the different models produced similar results).

Few people have perfect scores on the 15D, but few obtain scores below 0.4; that is, there are no serious ceiling effects, but the instrument does not identify health states with very low utility scores, at least, as measured by other scales.

The 15D has been modified for children (16D) and has been translated into 25 languages with 4 in preparation. The 15D website is: http://www.15d-instrument.net/15d

HUI 3: The HUI 3 descriptive system is an adaptation of HUI 2 and reflects the importance ranking assigned to a list of 15 symptoms in a Canadian survey of hospital patients. It consists of eight items with either five or six response levels. The 'within-the-skin' – i.e. disability based – descriptive system has no explicit social or handicap-based dimensions. An RS was used with 504 residents of Ontario, Canada, and the scores were converted to a SG (utility) equivalent score using an equation (a power function), which was calibrated to predict three SG scores from their corresponding score on a RS.

The HUI 3 model for combining items was based on the assumption of 'structural independence' – that is, the assumption that a single attribute is not measured in more than one way (as this will result in 'redundancy' or 'double counting' of disutility). According to one study, the correlation

between items varies between 0.02 and 0.35, which is consistent with the conventional psychometric definition of 'independence.' (When item correlation is low, it is assumed in psychometrics that items are picking up different aspects of a construct.) Instead of combining dimensions with an additive model, the HUI 3 employs the multiplicative formula recommended by Decision Analytic (Multiattribute Utility) theory.

The actual formula is deceptively complex and constructed from disutilities. As an example, with a three-dimensional instrument the formula might take the form:

$$\mathrm{DU} = 1.06\,[1 - (1 - 0.7\mathrm{DU}_1)(1 - 0.6\mathrm{DU}_2)(1 - 0.5\mathrm{DU}_3)]$$

where 1/1.06 is the scaling constant, and 0.7, 0.6, and 0.5 are dimension importance weights times the scaling constant, $U = 1 - \mathrm{DU}$. The scaling constant constrains the scale to the range of 0.00–1.00. When the dimension disutility scores for the three dimensions DU_1, DU_2, and DU_3 are all 0.00 (or 1.00), the utility score will be 1.00 (or 0.00).

The utilities predicted by HUI 3 fall below zero (worse than death), indicating the absence of floor effects. However, approximately 30% of scores from the general population exceed 0.95 indicating the likelihood of ceiling effects.

HUI 3 questionnaires are available in English, Chinese, Japanese, Russian, Dutch, French, German, Italian, Portuguese, Spanish, Czech, Polish, Finnish, Norwegian, and Danish. There are sixteen English versions, which differ in their mode of administration, the assessment viewpoint, and duration of assessment period. The website is http://fhs.mcmaster.ca/hug/

EQ-5D: The five item, three level EQ-5D defines 243 health states. It was originally designed as a brief 'linkage tool' to be used alongside more comprehensive MAU instruments and to facilitate comparison between studies that had used different instruments. Following the development of preference weights at the University of York, it became widely accepted as a standalone generic MAU instrument and eventually became the preferred instrument of the UK National Institute of Health and Clinical Excellence (NICE).

The UK weights, which are the most widely used, employ TTO data from a survey of 2997 adult members of the general UK population. TTO values were obtained for a number of holistic health states, which were created by combining different response levels from the EQ-5D descriptive system. These were regressed on item levels, and the best fitting regression equation was used to generate a score for all the health states defined by the descriptive system. Linear regression was used, so the final model, as with the QWB and 15D, is additive. Models were created for different sociodemographic groups with eight algorithms estimated using both TTO and RS. However, only the general population TTO formula is normally used (the formula was reported earlier in **Box 1**). The utilities for the 243 health states can be obtained directly from a table.

The correlation between EQ-5D dimensions varies, typically, from approximately 0.20 to 0.60 indicating structural dependence – that is, some aspects of HR-QoL are picked up by more than one item. However, econometric scaling may, potentially, overcome this problem (see section Theory and Evaluation) ensuring that predicted values and actual values are equal at the mean.

Negative scores are predicted for some of the general population indicating the absence of floor effects. However, approximately 35% of the general population obtain a score above 0.95 indicating the presence of ceiling effects.

The EQ-5D has been translated into 150 languages. A version for children aged 7–12 has been translated into 12 languages. A scoring algorithm has been estimated in the USA and 9 other countries (Belgium, Denmark, Finland, Germany, Japan, the Netherlands, Slovenia, Spain, and Zimbabwe). In 2009, the EQ-5 L, a 5-response level instrument (with the same items) was published and the EuroQol Group executive approved the use of 'bolt-ons' to increase instrument sensitivity for particular health states. The website is http://www.euroqol.org/.

SF-6D: Two versions of the SF-6D instrument are available – one derived from the SF-36, the most widely used generic (nonutility) HR-QoL instrument, and the other from its derivative, the SF-12. Consequently, utility scores may be derived from any study reporting values from these instruments. 'SF-6D (12)' and 'SF-6D (36)' are similar except for a reduction in the response categories for two items in SF-6D (12), which reduces the number of possible health states from 18 000 to 7500.

The items of the descriptive system of the SF-6D were derived from the factor analysis undertaken in developing the SF-36 and other psychometric evidence.

Utility scores for 249 health states were obtained from 611 respondents using the SG. These were regressed on item levels and the resulting linear equation used to predict utility scores for other health states. The resulting (0.5) formula took the form:

$$\mathrm{Utility} = 1 - 0.009\ \mathrm{PF2} + 0.008\ \mathrm{PF3} + \ldots - 0.007\ \mathrm{VIT5}$$

where PF2 and PF3 are the second and third response categories on the physical functioning dimension scale, and VIT5 is the fifth response category on the vitality dimension scale.

Several different models were used to estimate utilities (based on random effects linear regression, rank estimation data, and a nonparametric Baysian approach). The best fitting model predicted a minimum utility score of 0.203. Approximately 5% of the general population obtain scores below 0.5, indicating possible floor effects. In contrast only approximately 8% scored above 0.95, indicating the absence of ceiling effects.

Versions of the instrument have been developed in Australia, Brazil, Hong Kong, Japan, Portugal, and Singapore. More information about SF-6D can be accessed at: http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d.

AQoL: AQoL descriptive systems were constructed from reviews of existing instruments, the HR-QoL literature, from focus groups and 'construction surveys'. The latter involved administering large numbers of items to selected patients and the public. Factor analyses and structural equation modeling (SEM) were used to obtain a multilevel model. AQoL-8D has 35 items, which combine to form eight dimensions, which, in turn, combine into the two 'super dimensions' of physical and psychosocial ('mental') health.

Utility (TTO) scores are estimated from a multistage procedure, which employed both the multiplicative model described earlier for the HUI 3 and econometric modeling,

similar to the SF-6D (except that exponential net linear models were used).

The estimation procedure has four steps: (1) estimation of dimension from item scores with multiplicative models; (2) econometric correction of these dimension estimates; (3) combination of corrected dimensions into a (single) multiplicative model; and (4) econometric correction of the final multiplicative model.

AQoL-8D used a sample of 712 people aged 18–70 years to construct the descriptive system and a second sample of 628 to obtain TTO scale values (322 patients and 306 public). The scaling survey obtained values for 174 'within dimension,' multiitem health states and 375 multidimensional health states. Transformations have been created between AQoL-4D, 6D, and 8D. AQoL-4D (the original AQoL instrument without the original dimension for symptoms) has been reduced to an 8-item AQoL-Bref or AQoL-8 (which should not be confused with the AQoL-8D).

In a general population, few people score below 0.25, approximately 1.5% have perfect scores, and approximately 14% score above 0.95. Floor effects are therefore closer to 15D and SF-6D than to HUI 3 and EQ-5D, but there are no significant ceiling effects. The AQoL instruments have been translated into traditional and simplified Chinese, Spanish, German, Arabic, Norwegian, and Danish. The AQoL website is: http://www.aqol.com.au/.

Instrument Use and Acceptance

Instrument Use

Information on the use of each of the MAU instruments was obtained from the Web of Science database for the period 2005–10 and supplemented by references provided to the authors or from the instrument websites. The search identified 1682 studies, which employed at least one of the MAU instruments. These were used to construct Tables 4–6.

Table 4 indicates that EQ-5D was the most popular instrument by a significant margin, with 63.2% of the 1682 studies using it. This was followed by HUI 3 (9.8%) and SF-6D (8.8%). At the other end of the scale, 15D and AQoL were included in 6.9% and 4.3% of studies, respectively, and the QWB, the earliest widely used instrument, accounted for only 2.4% of total use.

The EQ-5D also dominated use in most countries and was only exceeded in Canada by the HUI 3 and in Finland by the 15D. Table 4 reveals significant 'local loyalty' with the use of all instruments peaking in their country of origin. Apart from EQ-5D, only HUI 3 and SF-6D achieved significant use in other countries.

Use of the instruments was also very concentrated. European studies accounted for 55% of the total, and the addition of USA and Canadian studies raises this to 80.5%. Within Europe, use was also concentrated, with Finland and Netherlands each accounting for more than 8% of the total, or double the usage by Germany, despite its much larger population and more than 65% of the usage by all other European

Table 4 Number of studies using the 6 MAU instruments

Instrument	*Country of Study Population*											*Economic evaluation*	*Total studies*	*%*
	USA	*Canada*	*UK*	*Finland*	*Germany*	*Spain*	*Sweden*	*Netherlands*	*Other Europe*	*Australasia*	*Multinationals*	*Other*		
QWB	31	4			1		1					4	6	41	2.4
15D		1		93	1		3		17	1			18	116	6.9
EQ-5D	133	52	181	24	62	57	67	103	181	34	97	72	166	1063	63.2
HUI 2	27	25	9		3		1	7				6	8	78	4.6
HUI 3	43	60		21	3		2	15	2	2	6	10	22	164	9.8
SF-6D	30	16	27	1	2	6	2	16	23	6	6	13	27	148	8.8
AQoL[a]		1								69	1	1	6	72	4.3
Total	264	159	217	139	72	63	76	141	223	112	110	106	253	1682	100
Percent	15.7	9.5	12.9	8.3	4.3	3.7	4.5	8.4	13.3	6.7	6.5	6.3	15.0	100.0	

[a]Combines AQoL 4D, 8D; 5 studies were pre 2005.
Data source: Web of Science, 2011.

Table 5 MAU instrument use by disease subgroup 2005–10

Disease subgroups	*QWB*	*15-D*	*EQ-5D*	*HUI 2*	*HUI 3*	*SF-6D*	*AQoL*[a]	*Total*	*%*
Muscular skeletal	4	12	107	3	4	17	5	152	9.1
General population	7	4	87	18	19	15	4	154	9.3
Cardio	2	15	84	4	7	10	11	133	8.0
Arthritis	0	8	71	5	11	21	5	121	7.3
Cancer I	4	6	69	5	16	2	2	104	6.3
Degenerative and elderly	3	3	69	6	14	3	8	106	6.4
Internal organs	0	10	67	5	5	10	1	98	5.9
Psychiatric	3	8	66	2	6	8	6	99	6.0
Diabetes mellitus	1	2	51	2	10	2	1	69	4.1
Other	1	2	51	3	7	8	1	73	4.4
Medical patients	3	8	49	3	11	9	1	84	5.1
Injury	1	6	44	3	4	6	8	72	4.3
Eating/obesity	2	5	29	0	2	6	5	49	2.9
Respiratory	2	1	27	2	6	2	0	40	2.4
Vision	1	4	26	0	6	1	0	38	2.3
Neurological	0	8	20	6	7	0	3	44	2.6
Skin	0	0	20	0	1	1	0	22	1.3
Female conditions	2	4	19	1	4	3	1	34	2.0
Trauma	0	0	19	0	0	2	3	24	1.4
Chronic condition	0	3	17	0	4	2	0	26	1.6
HIV[b]	1	1	15	1	4	2		24	1.4
ENT[c]	3	2	15	6	11	2	0	39	2.3
Renal	1	3	11	1	2	6	1	25	1.5
Autoimmune	0	1	9	0	2	7	3	22	1.3
Rheumatic		1	5	1	1	2		10	0.6
Total	41	117	1047	77	164	147	70	1663	100.0
%	2.5	7.0	63.0	4.6	9.9	8.8	4.2	100.0	

[a]AQoL-4D, 7 AQoL-6D and 2 AQoL-8D.
[b]HIV–Human immunodeficiency virus.
[c]ENT–Ear, nose and throat.
Data source: Web of Science, 2011.

Table 6 Validation studies (2005–10) comparison with other scales

Instrument	*Type of scale*[a]			*Head to head comparisons*[b]							*Total MAU comparisons*
	Disease specific instrument	*Nonutility instrument*	*Generic MAU instrument*	*QWB*	*EQ-5D*	*SF-6D*	*HUI 2*	*HUI 3*	*15D*	*AQoL*	
QWB	10	0	28	–	7	6	6	8	1	0	28
EQ-5D	137	53	76	7		57	16	26	9	5	120
SF-6D	21	9	57	6	57	–	10	16	3	3	95
HUI 2	22	3	52	6	16	10	–	18	1	0	51
HUI 3	37	11	71	8	26	16	18	–	1	2	71
15D	6	3	15	1	9	3	1	1	–	1	16
AQoL[c]	5	5	11	0	5	3	0	2	1	–	11
Total	238	84	310	28	120	95	51	71	16	11	392

[a]Number of separate publications classified by the instrument, which was the principal focus of the study.
[b]Number of comparisons. Studies with (3+ instruments) are entered multiple (2+) times.
[c]Combines AQoL 4D, 8D; 5 studies were pre 2005.
Data Source: Web of Science, 2011.

countries combined. The extent to which this is attributable to language and publication bias is unknown.

Only 15% of the studies included in Table 4 were primarily concerned with economic evaluation (which need utility scores) as distinct from their use as generic tools for the measurement of HR-QoL (which does not require scores to be 'utilities'). The disease categories in which they were used are reported in Table 5. This reflects a broad acceptance

Box 4 International pharmacoeconomic guidelines

References	*QWB*	*15D*	*EQ-5D*	*HUI*	*SF-6D*	*AQoL-8D*
Hungary	Noted as internationally recommended		Noted as internationally recommended	Noted as internationally recommended		
Poland			Recommended for measuring generic quality of life and the utility of health states	Recommended for determining the utility of health states		
Belgium			'As long as Belgian valuation sets for other instruments are not available, the use of the Flemish valuations for the EQ-5D health states is recommended'			
France	Recommended		Recommends QWB, HUI and EuroQoL: 'validations of French versions of the latter two are proposed'	Recommended		
Netherlands			Recommended	Recommended		
UK			Preferred, but 'may not be an appropriate measure of health-related utility in all circumstances'			
Ireland			Recommended		Recommended	
Scotland			Recommended, but 'it would be inappropriate to require the use of the EQ-5D to the exclusion of any other valid generic utility measures'			
Sweden			Recommended as an indirect measure for QALY-weightings			
Italy					Recommended	
Canada		Noted as widely used	Noted as widely used	Noted as widely used	Noted as widely used	
USA			Recommended	Recommended		
New Zealand			'The New Zealand EQ-5D Tariff 2 recommended. 'Other instruments can be used, however, their use should be well justified'			
Australia			Acceptable	Acceptable	Acceptable	Acceptable

of MAU instruments across the spectrum of disease categories, possibly reflecting the widespread use of self-reported disease-specific instruments in medicine. Given the scope of the literature search, however, the number of studies published in most of the disease areas is relatively small.

Acceptance by Government Health Authorities

The different instruments enjoy varying degrees of acceptability by health authorities and in national pharmaceutical guidelines. Examples of this are given in Box 4. Each of the instruments has been used in government health surveys. The 15D and AQoL instruments have only been adapted in this capacity in Finland and Australia, respectively.

Comparison of Instruments

Table 6 reports the number of studies which compare instruments and the head-to-head comparisons between MAU instruments between 2005 and 2010. From the first three columns, the majority of comparative studies involved a disease-specific instrument (238) or a generic nonutility instrument (84). There were 310 comparative studies of MAU instruments. Because these included multi-instrument comparisons, the number of head-to-head comparisons was greater than the number of studies (392). The largest number of these comparisons involved the EQ-5D (120) closely followed by the SF-6D (95) and HUI 3 (71). Comparisons primarily consisted of a Pearson correlation. Intraclass correlation, the preferred statistic even in simple comparisons, was relatively

Table 7 Proportion of variance in one instrument explained by another instrument (R^2): Australia and USA

Australia[a]	*15D*	*EQ-5D*	*HUI 3*	*SF-6D*	*AQoL-4D*
15D	1.00	0.58	0.55	0.59	0.64
EQ-5D		1.00	0.41	0.56	0.53
HUI 3			1.00	0.44	0.55
SF-6D				1.00	0.55
AQoL-4D					1.00
Mean	0.59	0.52	0.49	0.54	0.57

USA[b]	*QWB SA*	*EQ-5D*	*HUI 3*	*SF-6D*
QWB SA	1.00	0.41	0.45	0.43
EQ-5D		1.00	0.49	0.50
HUI 3			1.00	0.52
SF-6D				1.00
Mean	0.43	0.47	0.49	0.45

[a]Hawthorne, G., Richardson, J., and Day, N. A. (2001). A comparison of the assessment of quality of life (AQoL) with four other generic utility instruments. *Annals of Medicine* **33**(5), 358–370.
[b]Kaplan, R. M., Tally, S., Hays, R. D. *et al.* (2010). Five preference based indexes in cataract and heart failure patients were not equally responsive to change. *Journal of Clinical Epidemiology*, doi:10.1016/j.jclinepi.2010.04.010.

uncommon and psychometric analyses were rare. This is discussed further below.

Despite the large number of comparisons between MAU instruments reported in **Table 6,** only two large and two smaller studies have included five instruments. In an early Australian comparison (Hawthorne *et al.*, 2001), 956 hospital patients and general population respondents were administered the EQ-5D, SF-6D, 15D, HUI 3, and AQoL-4D. The proportion of instrument variation explained by other instruments varied from 41% to 59% leaving an average of 44% unexplained. The highest explanatory power was achieved by 15D, followed by AQoL-4D (**Table 7**). In a recent US study (Fryback *et al.*, 2010), 3844 adults were surveyed to compare the EQ-5D, QWB, HUI 2, HUI 3, and SF-6D. A weaker association was found than in Australia (reflecting the inclusion of only general population respondents). Overall 53% of instrument variance was not explained (**Table 7**). Recent work indicates that the strength of the association between instruments varies across the health spectrum.

Frequency distributions from the US study are reproduced in **Figure 2**. As the individuals included in each distribution are the same, the distributions would be identical (subject to respondent or transcription errors) if the instruments were measuring the same construct on the same scale. However, the figures indicate significant differences. A more recent and smaller Australian study (Khan and Richardson, 2011) reinforces this conclusion by creating the pair-wise comparisons of instrument frequencies shown in **Figure 3**. If the different instruments predicted identical utilities, the points in **Figure 3** would lie on the 45° line, i.e., Instrument A = 0.00 + 1.00 Instrument B. This does not occur indicating significant discrepancies in the predicted utilities.

Generally, researchers conducting multi-instrument comparisons have concluded that the utilities derived from the instruments are 'not equivalent,' that translation between them will result in 'low precision,' and that comparisons between them 'warrant caution.'

Theoretical reasons for these differences are discussed in section Theory and Evaluation and section Construct and Content Validity. However, one proximate cause is the difference in upper and lower end sensitivity, i.e., in 'ceiling' and 'floor' effects. For example, in the five-instrument US study cited earlier, the percentages of scores above 0.95 were 37.0 for EQ-5D, 36.9 for HUI 2, 36.2 for HUI 3, 1.7 for SF-6D, and 2.3 for QWB. **Figure 3** also reflects the strong ceiling effect of the EQ-5D (the horizontal scale in the three left hand diagrams) and the HUI 3. The SF-6D and EQ-5D have the strongest floor effect(s) with no values below 0.6. The AQoL-8D and HUI 3 had minimum values of 0.42 and −0.04, respectively. Additionally, for each value on one instrument there is significant variation in the value of other instruments (for the same person). For example, when SF-6D = 0.6, HUI 3, and AQoL-8D values varied from 0.25 to 1.00 and 0.55 to 0.95, respectively; when AQoL-8D = 0.8, HUI 3, and SF-6D varied from 0.25 to 1.00 and 0.10 to 1.00, respectively. Some of these variation will be undoubtedly random. Some may be attributable to the choice of scaling instrument as TTO, SG, and RS give slightly different values. The remainder must be attributable to the instrument descriptive systems and models.

The variation in instrument scores raises the question of which instruments yield the most appropriate utilities for use in economic evaluation and the appropriate contexts in which they might be used. There is no agreement about this. The theoretical foundations of the instruments are discussed in the next section. Instrument performance and validity is discussed in sections Construct and Content Validity, and Criterion Validity.

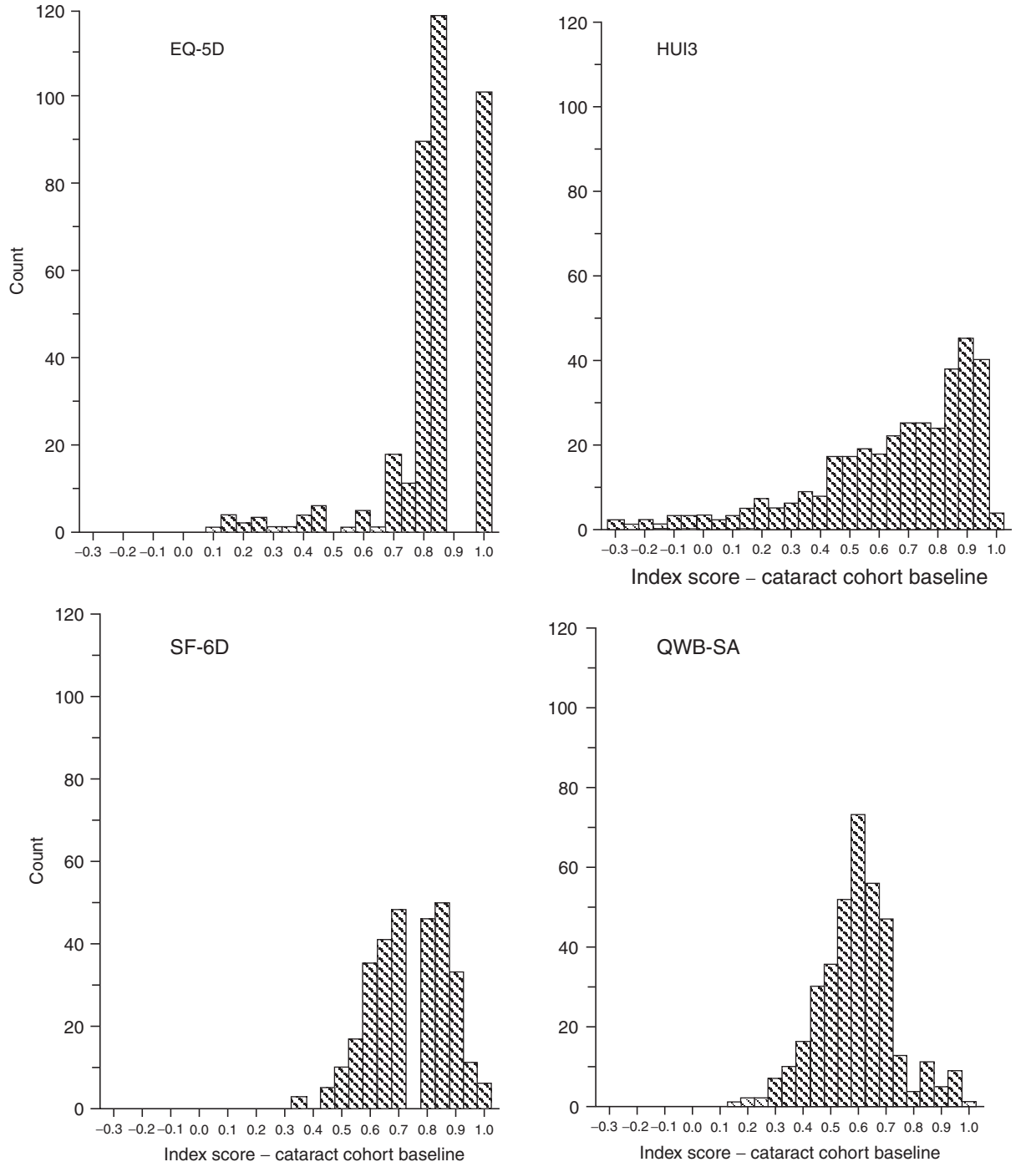


Figure 2 Distributions of baseline scores on 5 indexes at baseline for cataract. Reproduced from Kaplan, R. M., Tally, S., Hays, R. D., *et al.* (2010). Five preference based indexes in cataract and heart failure patients were not equally responsive to change. *Journal of Clinical Epidemiology*, doi:10.1016/j.jclinepi.2010.04.010.

Theory and Evaluation

Theoretical Foundations

Present MAU instruments draw on theory from three relatively distinct disciplines: decision analysis (DA), psychometrics, and economics/econometrics. The traditions in these areas are not always consistent, reflecting the context from which they arose. This has received relatively little explicit discussion, possibly because decision analysis and psychometrics have played only a limited role in mainstream economics. Nevertheless, they are of fundamental importance for the methods adopted in the construction of MAU instruments and their validity.

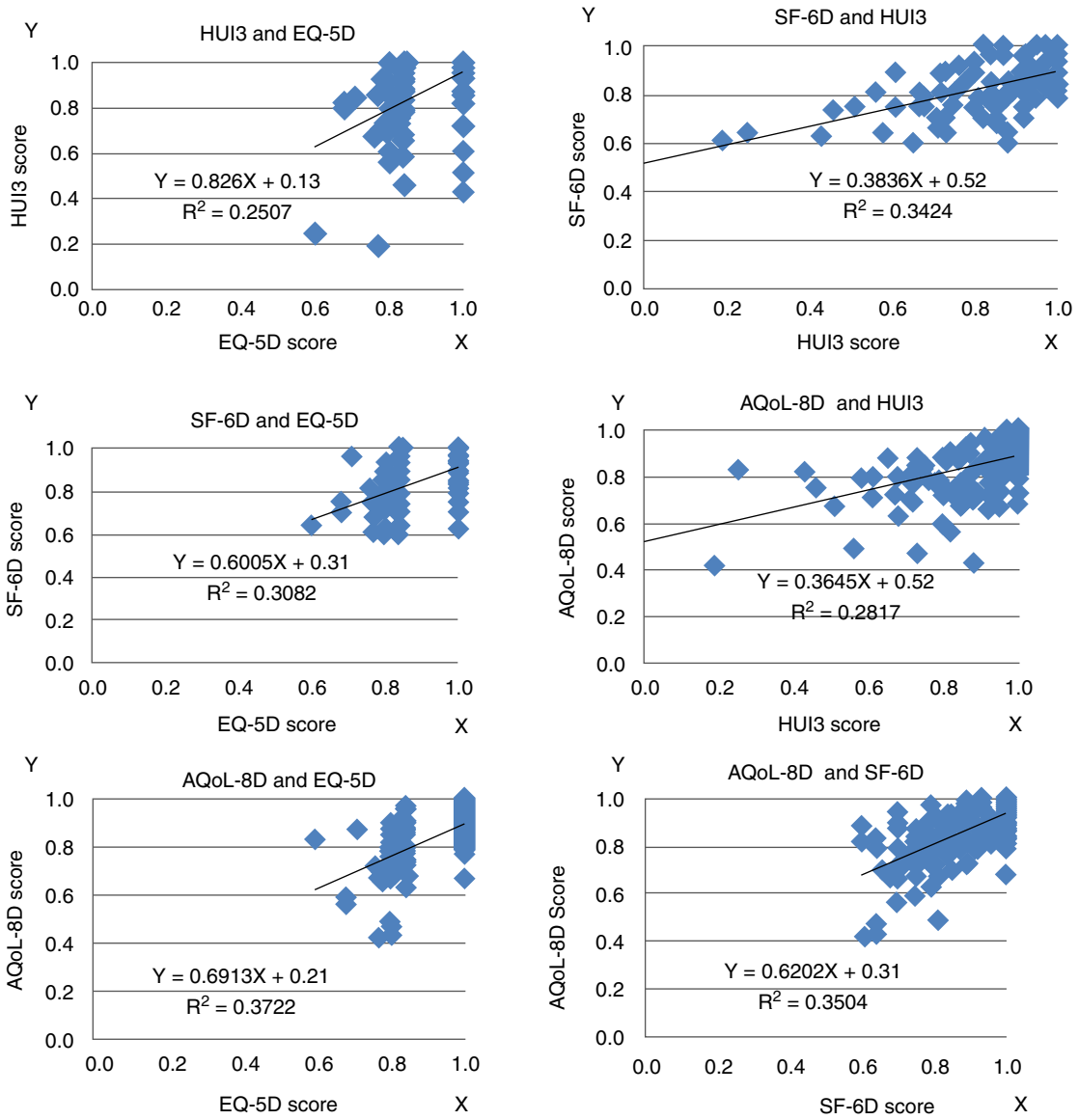


Figure 3 Pair-wise comparison of 4 MAU instruments. Reproduced from Khan, M. A. and Richardson, J. (2011). A comparison of 7 instruments in a small, general population, Research Paper 60. Melbourne: Centre for Health Economics, Monash University.

DA: The 15D, HUI 3, and AQoL-8D all seek theoretical justification, at least in part, from MAU theory, a subset of DA theory. This recommends that complex outcomes (in the present case, 'health states') should be decomposed into attributes (dimensions) such as pain and vision. Utility scores should be assigned to each of the attributes and a model used to combine attribute utilities into a total utility score.

Importantly, the theory requires that descriptive attributes should be structurally independent. For example, a business model optimizing output as a function of total revenue, total cost, and profit would result in 'double counting' as the first attribute is the sum of the other two. Depending on the nature of preferences (for the attributes) DA models may be additive, multiplicative, or multilinear. The QWB and 15D assume additive independence. Empirical results for the HUI 3 and AQoL-8D implied the need for multiplicative models.

Psychometric theory: Psychometrics is the basis of measurement theory in education and psychology, which quantifies unobserved 'constructs' (such as educational attainment, IQ, and personality). Its potential contribution in the present context is threefold. First, it prescribes methods for constructing instruments; second, it describes criteria for their evaluation; and third, it describes numerous forms of bias and other sources of measurement error.

A tension exists between the decision analytic and psychometric approaches. The former requires independence between items and between dimensions to avoid double counting of disutilities. The latter approach assumes that items will correlate to some extent and that the scale for a satisfactory construct requires a minimum of 3 and

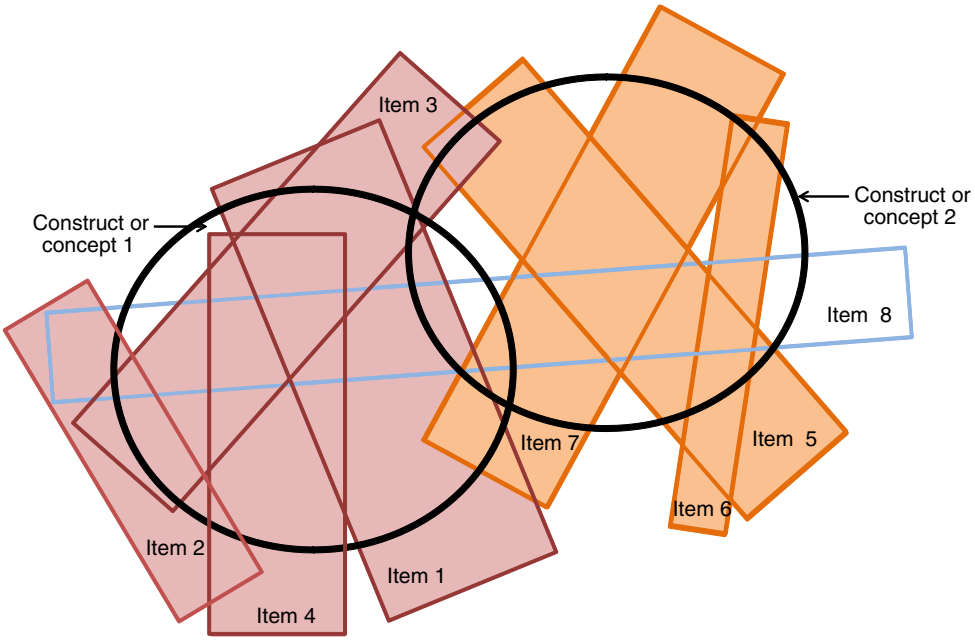


Figure 4 Item: Question with a series of possible response levels (e.g., how often do you feel sad? (a) never, (b) rarely, (c) some of the time, (d) usually, and (e) nearly all the time). Concept 1: An abstract idea concerning some hypothesized attribute or characteristic, mental health. Concept 2: A mini theory or created construct to explain observed behavior.

preferably 4 items for content validity. The reason for this is illustrated in Figure 4. Two constructs are represented by bolded circles and a number of items – linguistic statements – are represented by rectangles. Reflecting the imprecision of language, no one item exactly corresponds with a construct. Items 1–4 are required to measure construct 1 and items 5–7 to measure construct 2. Factor analyses may be used to obtain the efficient set of items and to omit items that 'cross load' (item 8). Confirmatory factor analysis or SEM may be used to achieve this goal while forcing the retention of theoretically desired constructs. However, the resulting instrument structure achieves content validity by violating the DA requirement of structural independence, which is needed to avoid the double counting of disutilities. This issue, however, has received little attention.

With the exception of the AQoL instruments, full psychometric analysis has played little or no role in MAU instrument construction. The SF-6D employed psychometric evidence from the SF36 in the selection of items, and the 15D was revised following psychometric analysis. Other instruments selected items by other means (see section History, Description and Construction of MAU Instruments). One explanation offered for this is the belief that it is preferences that are important for MAU instruments, not description. However, valid preference measurement requires valid description of what is to be evaluated. If a descriptive element is unimportant, then the preference weight will be lower or zero. Use of the best preference methods and modeling cannot compensate for the absence of a nontrivial descriptive element. Content validity is discussed further in the section Construct and Content Validity.

Economics and econometrics: The gold standard for evaluating an MAU instrument is whether or not it measures 'utility' as envisaged in economic evaluation studies (see section Criterion Validity). This implies that a preference-based instrument should be used for scaling, and this is generally interpreted as implying the use of the SG or TTO. However, the subject is controversial, and some argue that there are insufficient reasons for excluding the RS. Recently, weights have been assigned using ranking techniques and item response theory, and the use of 'best–worst scaling' has been foreshadowed.

As noted above, the decision-theoretic requirement of item or dimension independence is difficult to achieve, and the resolution of this problem in the EQ-5D, SF-6D, and AQoL-8D has been to use statistical methods to determine the importance of items. Multiattribute health states are evaluated with the SG or TTO and regressed on item scores or on dummy variables for the response levels. The regression assigns the most efficient weights to the items, i.e. the regression coefficients – which best explain variation in the health state values. The predicted values 'fit the data' as a regression line passes through the mean value of observations. Consequently, at least at the mean, the effect of 'double counting' is mitigated. The choice of regression model, however, is contentious. From MAU theory, linear (additive) models may be inappropriate, and utilities predicted from linear models may therefore incorporate systematic bias as they move away from the mean. However, these models are employed by the QWB, SF-6D, and EQ-5D.

Competing claims have been made about the use of decision analytic and econometric techniques, but the evidence is limited. Both approaches are based on a set of assumptions and constraints that are violated to a greater or lesser extent depending on the population group and

Box 5 Validity reliability-related definitions

Validity: Measurement of what is intended
Validation: A process of determining (the appropriate level of) confidence in the inferences drawn from instrument values
Construct: A concept created to explain observed relationships
Construct validity: The construct measures what is intended
  a. Convergent validity: Correlation with other measures expected to correlate with the construct
  b. Discriminant validity: Non-correlation with measures of different constructs (e.g., MAU instruments and blood pressure
  c. Discriminative (extreme group) validity: Discrimination between different groups (patients and public)
Content validity: There is a representative sample of target elements in the descriptive system (i.e., outcomes, behaviors, symptoms, etc.) or elements, which vary directly with the elements of interest
Face validity: The content appears adequate on inspection
Criterion validity: Constructs behaviors expected as judged by external criteria
  a. Gold standard validity: The instrument correlates with the gold standard measure
  b. Concurrent validity: The instrument correlates with the criterion
  c. Predictive validity: The instrument predicts other (criterion) variables as expected
Reliability: A measure of consistency. It is the proportion of the total variability in scores, which is accounted for by the differences in the average values across observations. It applies to the interval consistency of the items of an instrument and to the test-retest consistency of the instrument over time.

disease. This suggests that validation requires context-specific evidence.

Evaluation Criteria

Evaluation criteria for assessing MAU instruments include practicality and reliability (measurement error should be a small fraction of total variability as judged, for example, by test–retest and by Cronbach's α). The MAU instruments reviewed here have evidence of these properties, which is short in terms of most questionnaire-based research. The largest instrument – AQoL-8D – takes an average of 5.4 min to complete in its online version. Test–retest and Cronbach's α coefficients are satisfactory according to accepted norms.

The most contentious criterion is validity, whether or not an instrument measures what it purports to measure. The lack of agreement between instruments noted earlier implies that some or all of the MAU instruments are not universally valid or that they seem to measure differing concepts, although this possibility has not been suggested in the literature.

Different types of 'validity' are defined in Box 5. The common element is that each is a test that justifies greater or lesser confidence in the instrument's predictions. This means that in practice an instrument is never (fully) 'validated' in the sense that it has been 'proven universally correct' and the statement that 'an instrument has been validated' is misleading in implying this. Rather, instruments are more or less supported both empirically and theoretically (an interplay sometimes described as a 'nomological net'). Importantly, the strength of the evidence depends on the stringency as well as the outcome of the test.

Construct and Content Validity

The validity of an MAU instrument depends on the validity of its three components: the descriptive system (items and dimensions), the scaling method (TTO, SG, etc.), and the model used to combine items (additive, multiplicative, etc.). There is no consensus concerning the scaling method (TTO, SG, etc.). However, there is a relatively high level of agreement between scores from the chief scaling methods, and differences in between these could not explain the observed variation between instrument scores.

Combination models also differ and the evidence for the assumptions behind them is incomplete. Validity could be tested by comparing estimates of health state utilities from different models with independent holistic estimates of the same health states. Few such studies have been undertaken. Descriptive systems also differ very significantly in size, item content and syntax (Tables 1–3). The effect of this on MAU instrument scores and validity is an unresolved, although critical, issue. Some of the evidence is outlined in section Criterion Validity below.

The great majority of the validation studies in Table 6 are concerned with 'construct validity' and primarily 'convergent validity'. These correlation-based studies are relatively weak tests, which are necessary but not sufficient for confidence that an instrument measures the utilities needed for economic evaluation. Correlation will occur as long as an instrument can, minimally, detect extreme values. It does not indicate that values have the properties needed for economic evaluation or even that both instruments use the same scale. In the linear relationship, $U = a + bI$, where I is an instrument's estimate of true utility U, instrument validity would imply that $a = 0$; $b = 1.0$. For this reason, a better measure of association than correlation is the intraclass correlation (ICC), which tests the equivalence of absolute values. However, only a minority of the studies use this technique. The difference is potentially important. In the early five-instrument Australian study, the 15D had the highest average correlation with other instruments (construct validity). However, incremental changes in 15D were about half the magnitude of corresponding changes in other instruments indicating a low ICC. Similarly, the ability to discriminate between extreme groups is a weak test of the validity of the numerical values produced by an instrument.

Differences between instrument descriptive systems, summarized in Tables 1 and 2, indicate the potential for different levels of content validity. The early five-instrument Australian study anecdotally illustrated the importance of these differences when the same respondent scored 0.14 and 0.8 for the HUI 3 and EQ-5D, respectively. When the HUI 3 items for sense perception were altered from their reported scores to the highest HUI 3 item score (effectively removing them from the instrument), the predicted HUI 3 utility score rose to 0.74; that is, 91% of the original difference was attributable to items in HUI 3 which are not included in the EQ-5D. If the items in the EQ-5D had the same descriptive power – 'content' – in the context of sense perception, this would not have occurred.

Evidence for content validity is commonly obtained from the psychometric analyses, which led to the selection of the instrument's item structure and, in particular, evidence of whether or not additional items were redundant or added new content. However, except for AQoL, generic MAU instruments have not been developed in this way.

Few other tests of content validity have been reported. The most common have been comparisons of ceiling and floor effects (i.e. 'insensitivity' close to 'best' and 'worst' health). The test is limited as it applies only to extremities of the scale and results will vary with either the item structure or the weights attached to a given set of items. As noted in the section History, Description and Construction of MAU Instruments, ceiling effects vary significantly between instruments.

In one recent test, scores for each attribute of the EQ-5D, HUI 2, and SF-36 were individually predicted from the scores obtained by the other two instruments using data from 264 German patients. Adjusted R^2 values were between 0.01 and 0.57. The SF-6D attribute 'role limitation' and HUI 2 'sensation' were 'virtually unrelated to the other instruments.' The authors concluded that the instrument content differs 'so much that... (they) would produce different valuations even if other components of the instruments were the same' (Konerding *et al.*, 2009).

Rather than demonstrate differences, the authors of the Australian study reported in **Figure 3** (Khan and Richardson, 2011) attempted to identify missing content. Instruments on the vertical axis in **Figure 3,** which are relatively sensitive to a particular dimension, will, on average, have lower scores than predicted. Points will be below the line. The ratio of dimension scores of individuals from above to below the line therefore indicates the relative sensitivity of the instrument to a dimension. Results suggest that, at least in the relatively healthy population surveyed, HUI 3 has less content than other MAU instruments in the domains of mental health and relationships, and that AQoL-8D has greater content for all of the mental and social dimensions. EQ-5D is relatively sensitive to pain. Unexpectedly, HUI 3 was not significantly more sensitive with respect to senses, but this is probably because the sample was small ($n=158$) and there were few respondents with impaired senses.

Table 8 Predictive validity: prediction from utility scores

Instrument	*Permanent problem cured (health states of similar severity chosen for each instrument)*	*Increase in utility*[a]	*Equivalent*		
	= return to good health for 20 years	*Value per annum*	*Cures = 1 life saved*[b]	*Life extension with original QoL*[c]	
				Rate of time preference = 0%	*Rate of time preference = 2%*
QWB	Headache, dizziness, ringing in ears, and spells of feeling hot, nervous, or shaky	0.244	4	6.5 years	9.6 years
15D	Mild physical discomfort...pain, ache, nausea, itching, etc.	0.023	43	5.6 months	8.3 months
EQ-5D	Moderate pain or discomfort and some problem walking	0.273	4	7.5 years	11.1 years
HUI 3	Moderate pain that prevents a few activities	0.137	7	3.2 years	4.7 years
SF-6D	Pain, which interferes with normal work...a little bit	0.07	14	1.5 years	2.2 years
AQoL-8D[d]	Moderate pain...which sometimes interferes with usual activities	0.02	50	4.9 months	7.3 months

[a]Increase in utility if an individual is cured from the permanent problem and returned to best health on the scale or in the case of AQoL-8D to normal health as this corresponds approximately to best health on other scales. The seven items set at 'normal' levels are jobs around house, getting around the house, mobility, toileting, coping, relationships, content with life, and enthusiasm.

[b]The number of cures, *n*, equivalent to saving one life is calculated as $n=1/(\text{increase in utility})$. Therefore, cures times value of cure $= n \times$ increase in utility $= 1.00$.

[c]The number of years of life extension, *n*, is calculated from QALY gain $= 20(\text{utility gain}) = n(\text{original utility})$.

[d]AQoL-8D is at 'normal' (not best) levels for seven items, viz, jobs around the house, getting around the house, mobility, toileting, coping, relationships, content with life, and enthusiasm.

Criterion Validity

Comparisons of an instrument with external criteria are usually classified as tests of concurrent or predictive validity. They raise the question of what represents an acceptable external criterion. The theoretical questions include the choice between (*inter alia*) 'decision' and 'experience' utilities (i.e. evaluation before and after experiencing a state); the selection of the appropriate judge of utility – the public or patients – and the measurement perspective – individual or social. These issues are unresolved in the literature. However, instruments (of necessity) incorporate judgments with respect to each of these questions.

All of the instruments in the present review incorporated an individual perspective. TTO or SG values were obtained from people imagining they were, personally, in a health state. They were not asked to be representatives of the society. Most instruments embody answers from members of the public, not from patients (AQoL-7D and AQoL-8D obtained answers from both). Perhaps the most important neglected question is the scope of the construct 'health-related quality of life' to be incorporated in the instrument: should it include social dimensions, be restricted to a 'within the skin' concept or be defined by what individuals have in mind when asked about 'health?'

Subject to these caveats, a limited number of tests of criterion validity have been reported. One approach has been to ask respondents to directly value their own health state using a scaling instrument (TTO, SG, etc.) and to compare the result with the value predicted from an MAU instrument. Testing individual instruments this way has provided supporting evidence for the validity of the HUI and 15D.

The early Australian study (Hawthorne *et al.*, 2001) tested the validity of 15D, SF-5D, AQoL-4D, HUI 3, and EQ-5D using the 'self TTO' – i.e. the reduction people would accept in their own life expectancy in exchange for perfect quality of life. The aim of the test was two-fold: to determine (1) which instrument explained most variation in self TTO and (2) which instrument best explained what other instruments failed to explain, i.e. the residual from the first stage analysis. The results of both tests were similar: 15D demonstrated the greatest explanatory power followed, respectively, by SF-6D, AQoL-4D, HUI 3, and EQ-5D.

The results of this study are similar to those of a recent five instrument Finnish study (Honkalampi and Sintonen, 2010). However, the properties of self-referential measures have not been discussed in the economics literature and interpreting these results is difficult. A further suggested test is the use of willingness to pay as a criterion for evaluating MAU instruments. However, the technique is controversial in the context of QALYs, and no one has adopted the suggestion empirically.

A weak test of preferences is to determine whether most people agree that improvement has occurred when MAU instrument scores increase. Applying this test, Roberts and Dolan (2004) found that a 0.20 increase in the EQ-5D score was necessary before 70% of respondents agreed that any improvement had occurred.

The logic of this test of predictive validity was to use MAU instrument scores to predict what people would choose. Similar logic was used earlier in a study by Nord *et al.* (1993) drawing on the idea that QALYs are the product of utility, life years, and the number of people affected. From this, MAU instrument scores were used to predict the number of people moving from a health state to full health, which would be equivalent to saving one life. Results from the QWB and HUI 1 were so implausible that they suggested a lack of predictive validity.

An objection to this latter approach is that the test alters the distribution of utility and introduces a consideration of equity, which may (or may not) invalidate the conclusion. However, the method could have been applied with respect to individuals' personal trade-offs between quality of life and length of life, the right hand columns of **Table 8**. According to this, the life extension that is equivalent to a moderate improvement in the quality of life (described in somewhat different ways across instruments) is 38 times greater for the EQ-5D (11.1 years) than for AQoL-8D (3.5 months) and 14 times greater for the QWB (9.6 years) than for the 15D (8.3 months). As with the Roberts and Dolan (2004) study, (dis)agreement with the predictions from each instrument could be obtained independently from the population. Because the test is simple, it is a potentially powerful and rigorous test of criterion validity in the context of economic evaluation.

Conclusions

Numerous questions have not been considered here. Foremost is disagreement about what is to be measured. 'Health' like 'beauty' is a vague concept and has been operationalized very differently. In effect, each MAU instrument has provided its own unique definition, which has generally been unchallenged. The chief decision concerns the breadth and content of the definition. If an MAU instrument is intended strictly for use within a Government Health Service, the definition may remain narrow and exclude items extraneous to government funding, for example, social or dental-specific dimensions. The values embodied in the WHO definition of health and in orthodox economics would suggest a broader, more encompassing approach. That is, anything affecting preferences should be included.

Other omitted issues include perspective and the concept of utility. Present MAU instruments seek to measure personal, not social, preferences. These are generally measured on the basis of descriptions given to the public or to individuals who have experienced the health state and without any consideration given to the distribution of benefits. Challenges include a proper demonstration that MAU instruments have construct validity in different disease areas and, more fundamentally, predictive validity using criteria relevant for economic evaluation.

The article has focused on the construction and validity of MAU instruments. It indicates that the scores obtained from different MAU instruments differ significantly and, consequently, QALY values, CUA ratios, and the likelihood of health service funding are all significantly affected by the choice of instrument. The numerical values obtained from the instruments depend on the validity of the descriptive system, model combining the items, and scaling instrument.

Of these, the evidence suggests greatest agreement about the scaling instruments. TTO, SG, and even RS values correspond fairly well. Despite this, the focus in the economics literature has been on this choice, with overall instrument validity often judged primarily on the basis of the scaling instrument.

There has been little empirical evidence published with respect to the choice of the MAU model. Authors of the HUI 3 and AQoL-8D both found evidence that additive models were less satisfactory than multiplicative models. However, these models permit double counting of the disutility. Econometric linear models are flexible and ensure that predicted values 'pass through' the observed utilities, at least at the mean. But extrapolation with an additive model is problematical. Apart from AQoL-8D, there has been little experimentation with nonlinear models. Least agreement exists between the items of the MAU instrument's descriptive system, but this is where there has been least discussion. Instruments have been created using different approaches and generally with little explicit regard for content validity. It is therefore, at this level that differences between instrument scores will most probably be found.

None of the studies evaluating MAU instruments (which have been published) appears to have concluded that a scale is invalid. This poses a problem for decision makers as the outcome of an evaluation may presently depend on the choice of instrument. The approach to this problem by the UK NICE has been to nominate a preferred instrument for use in all evaluations. This does not overcome the underlying problem because, as recognized by NICE, a single instrument may not be an appropriate measure of health-related utility in all circumstances and, at present, there are no evidence-based guidelines concerning which instrument to use or how to interpret results from instruments which conflict.

Instruments are neither right nor wrong. The evidence suggests that they are more or less sensitive in different contexts. Use of a single instrument will favor interventions affecting health states where the instrument is sensitive (and the intervention efficacious) and disadvantage interventions where sensitivity is low. Researchers presently have little choice but to select from available instruments and to evaluate their ability to measure the health states which are of relevance to them. In the longer term, there is a need for a significant research program to determine which instruments should be used in which contexts and how to compare their values.

See also: Disability-Adjusted Life Years. Multiattribute Utility Instruments: Condition-Specific Versions. Quality-Adjusted Life-Years

References

Fryback, D. G., Palta, M., Cherepanov, D., Bolt, D. and Kim, J. (2010). Comparison of 5 health related quality of life indexes using item response theory analysis. *Medical Decision Making* **30**(1), 5–15.

Hawthorne, G., Richardson, J. and Day, N. A. (2001). A comparison of the assessment of quality of life (AQoL) with four other generic utility instruments. *Annals of Medicine* **33**(5), 358–370.

Honkalampi, T. and Sintonen, H. (2010). Do the 15D scores and time trade-off (TTO) values of hospital patients' own health agree? *International Journal of Technology Assessment in Health Care* **26**(1), 117–123.

Khan, M. A. and Richardson, J. (2011). A comparison of 7 instruments in a small, general population. Research Paper 60. Melbourne: Centre for Health Economics, Monash University.

Konerding, U., Moock, J. and Kohlmann, T. (2009). The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common? *Quality of Life Research* **18**, 1249–1261.

Nord, E., Richardson, J. and Macarounas-Kirchmann, K. (1993). Social evaluation of health care versus personal evaluation of health states. *International Journal of Technology Assessment in Health Care* **9**(4), 463–478.

Roberts, J. and Dolan, P. (2004). To what extent do people prefer health states with higher values? A note on evidence from the EQ-5D valuation set. *Health Economics* **13**, 733–737.

Further Reading

Brazier, J., Ratcliffe, J., Tsuchiya, A. and Salomon, J. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.

Nord, E. (2001). Health state values from multi attribute utility instruments need correction. *Annals of Medicine* **33**, 371–374.

Richardson, J., McKie, J. and Bariola, E. (2011). Review and critique of health related multi attribute utility instruments. Research Paper (64). Melbourne: Centre for Health Economics, Monash University. ISBN 1 921187 63 8.

Streiner, D. L. and Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford University Press.

Multiattribute Utility Instruments: Condition-Specific Versions

D Rowen and J Brazier, School of Health and Related Research, University of Sheffield, Sheffield, UK

Introduction

A multiattribute utility (MAU) instrument consists of (1) a health questionnaire that establishes a health profile and (2) a scoring system that converts the multiattribute profile into an overall utility score. A generic MAU instrument can be used for all patients, whereas a condition-specific MAU instrument is intended only for patients with the condition of interest. Overall utility scores are used in economic evaluation of health care in terms of cost-utility analyses, where benefits are measured using quality-adjusted life-years (QALYs). They can also be used in monitoring of population health in terms of quality-adjusted life expectancy (QALE).

Some reimbursement agencies prefer the use of generic MAU instruments to generate QALYs. However, the generic measures may not always be available or appropriate. There is published evidence indicating that existing generic MAU instruments are valid, responsive, and reliable for many conditions. However, generic MAU instruments have been shown to perform poorly in terms of validity or responsiveness to change in some conditions. For example, the generic EQ-5D has been found to perform poorly in visual impairment in macular degeneration, hearing loss, leg ulcers, and schizophrenia. For these conditions the descriptive system of the generic MAU instruments either does not capture the impact of the change for that condition within its dimensions, or is not sufficiently able to capture small changes within a dimension. This means that the generic MAU instruments will not capture potentially important changes in utility across interventions because of the inadequacy of the descriptive system for that patient group. Condition-specific MAU instruments by contrast are designed to focus on functionings and symptoms that are affected both by the condition and treatments for the condition. This enables them to capture potentially important changes better across interventions for that condition.

In the following greater details on the construction and use of condition-specific MAU instruments are provided. Section 'Construction' summarizes their construction. Section 'Validity' examines their performance both in comparison to condition-specific instruments that are not preference-based and generic MAU instruments and addresses some further issues regarding validity. Section 'Future developments' outlines future developments. Additional readings are provided which contain information and references supporting the text in this article.

Construction

Background

Condition-specific measures vary in composition but typically have multiple domains, at least some of which may be highly correlated and not independent, and each of which has multiple items. The number of domains and items is often large (e.g., 30 items). The scoring process is often not based on preferences (i.e., non-preference-based), as item scores are typically summed to obtain a domain score and an overall score across all items. This simple summative scoring unrealistically assumes that all items, and the difference between all response choices, are equally important. Changes in scores do not necessarily reflect changes in quality of life that are valued by either patients or the general population. In the following these non-preference-based condition-specific measures are referred to as 'NPCS measures.'

Clinical studies often use NPCS measures (e.g., the cancer-specific EORTC QLQ-C30) and do not include generic MAU instruments. This is in part because of concerns about the appropriateness of generic measures in some conditions, but is also because of concerns surrounding patient burden and costs. Many trials are designed to provide information for multiple analyses (e.g., licensing and labeling claims) rather than economic evaluation alone, and often NPCS measures are used in these analyses rather than generic MAU instruments. This means that NPCS measures will continue to be an important source of evidence for economic evaluation. However, these measures cannot be used directly to generate QALYs. As shown by John Brazier and colleagues, they can be used indirectly to generate QALYs using mapping, which uses the statistical relationship between the NPCS measure and a MAU instrument to estimate utility scores using data from the NPCS measure alone. However, NPCS measures can be used directly if a condition-specific MAU instrument is derived from them. Section 'Development from existing non-preference-based condition-specific measure' describes this process.

Development from Existing Non-Preference-Based Condition-Specific Measure

For the majority of condition-specific MAU instruments the scoring system generates a score using responses to only a subset of the items included in the questionnaire. This is because the original questionnaire was often not designed as a MAU instrument and typically contains a large number of items across many domains. To derive utility scores from the questionnaire a subset of items is selected to form a descriptive system, which contains a small number of dimensions and severity levels for each dimension. The descriptive system is used to describe all possible health states with its associated utility score. What this means in practice is that for condition-specific MAU instruments the utility scores are usually generated by converting data for an existing questionnaire in the same way that SF-36 or SF-12 data are used to generate the SF-6D. For example, the cancer-specific EORTC QLQ-C30 data are used to generate the EORTC-8D cancer-specific utility score. This has the advantage that many condition-specific utility scores can be generated using existing data sets. Furthermore these data sets will contain both domain scores for

 doi:10.1016/B978-0-12-375678-7.00512-5

the original instrument and a utility score to provide detailed information for a range of analyses.

Figure 1 reports a six-stage process for deriving a condition-specific MAU instrument from an existing non-preference-based condition-specific measure that was first used to derive an overactive bladder-specific MAU instrument (AQL-5D) from the overactive bladder questionnaire (OABq). Stages 1 to 4 produce the descriptive system and stages 5 and 6 generate utility scores for all states defined by the descriptive system.

Stages 1 to 3 produce the descriptive system using existing data for the NPCS measure, and stage 4 validates the descriptive system using other data. Stages 1 to 3 provide the structure of the descriptive system by selecting dimensions and a minimum number of items per dimension to fully represent that dimension. This involves the use of psychometric techniques such as factor analysis, Rasch analysis, and Item Response Theory alongside classical psychometric techniques such as standardized response means and effect sizes. This process is undertaken to ensure that the selected descriptive system accurately represents the dimensionality and maintains the desirable psychometric properties of the original instrument. For example, the overactive bladder questionnaire has 33 items each with 6 severity levels covering 5 domains: symptom bother; sleep; coping; concern; and social interaction. The descriptive system of the OAB-5D MAU instrument derived from the questionnaire has 5 dimensions: urge to urinate; urine loss; sleep; coping; and concern each with 5 severity levels (Young *et al.*, 2009).

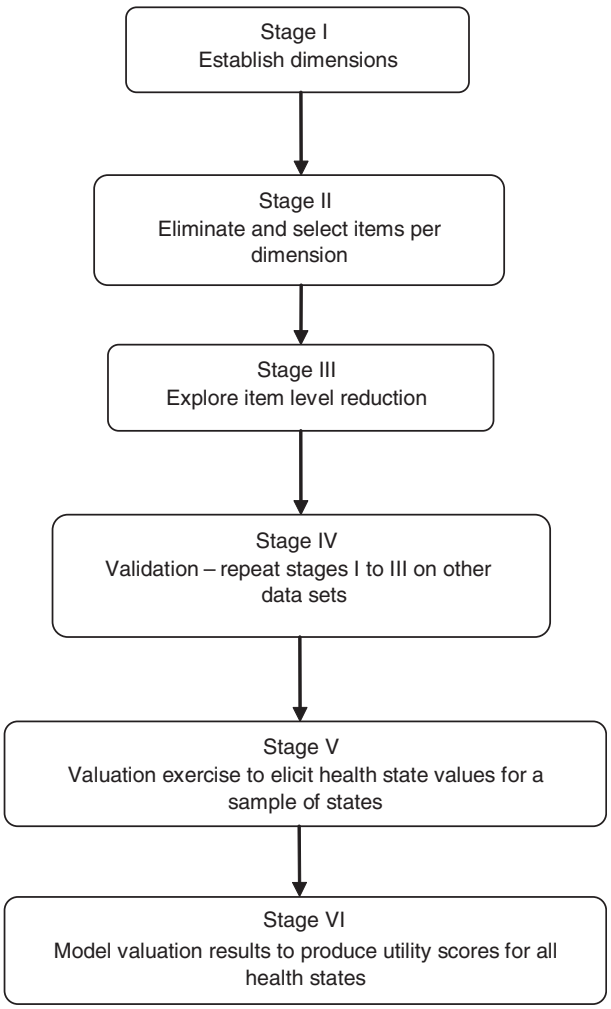


Figure 1 Six stages for deriving a condition-specific MAU instrument from an existing condition-specific non-preference-based measure. Reproduced with permission from Brazier, J., Rowen, D., Mavranezouli, I., et al. (2012). Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). *Health Technology Assessment* **16**(32), 1–114.

Stage 5 elicits health state utility scores for a sample of health states. Even health state descriptive systems like OAB-5D define thousands of health states meaning it is impractical and infeasible to value all states. A sample of health states are selected for valuation using a variety of techniques such as an orthogonal array or balanced design. The selection process differs by whether multiattribute utility theory (MAUT) or statistical inference is the proposed modeling strategy for stage 6 and the valuation technique used, but states are selected in order to enable utility scores for all states to be estimated from the elicited data for the sampled states. A variety of elicitation techniques is used, such as time trade-off (TTO), standard gamble (SG), visual analogue scale (VAS), and discrete choice experiment (DCE). Preferences can be elicited from a variety of sources including general population, patients, and carers. However, it should be noted that when patients value health states in such experiments they value hypothetical health states, not their own health state, to enable these values to be used to estimate scores for all health states described by the descriptive system. Currently, general population preferences are typically recommended by health economists if the measure will be used in economic evaluation submitted to policy decision makers. However, as noted by Mike Drummond and colleagues, this is being debated. Stage 6 models the utility data to produce utilities for every health state defined by the descriptive system. The modeling follows either a MAUT or statistical inference approach and will differ depending on the process used to select states and the elicitation technique. As noted by John Brazier and colleagues, the process will also differ if the instrument is not multiattribute with multiple independent attributes, but instead has one main attribute, such as flushing or mental health.

New Developments

The descriptive systems for some condition-specific MAU instruments are developed 'de novo' rather than derived from existing non-preference-based condition-specific measures, requiring a modified approach for stages 1 to 4 outlined above. The most rigorous method of development of a new descriptive system involves qualitative research to identify dimensions and items and validation of the generated descriptive system using psychometric analyses before valuation. Other approaches in the literature include the use of psychometric analyses on a battery of existing measures or a literature review. Recent US Food and Drug Administration (FDA) guidance for industry (U.S. Department of Health and Human

Services Food and Drug Administration (FDA), 2009) outlines guidelines for the development of dimensions and items for non-preference-based measures and emphasize the role of patients with the condition in generating and validating the content. Valuation follows the process as outlined in stages 5 and 6 above.

Description of Existing Instruments

Table 1 outlines the descriptive systems for existing condition-specific MAU instruments. Instruments were identified by John Brazier and colleagues in a recent review of published studies. Unpublished instruments that the authors were aware of were added to the list (see http://www.sheffield.ac.uk/scharr/sections/heds/mvh). There are 28 instruments covering 27 different conditions ranging from exact diagnoses, such as glaucoma and lung cancer, to more general conditions, such as visual impairment and cancer. The size of the descriptive system varies greatly by measure, with a range of possible health states for a system ranging from 10 to 390 625. The focus of the dimensions differs across instruments: (1) symptoms or health-related quality of life (HRQL) and (2) condition-specific dimensions alone or dimensions able to capture side effects and comorbidities. Indeed, dimensions vary for instruments within a condition or International Classification of Diseases.

Table 2 outlines the valuation processes used to produce utility scores for all states defined by the descriptive system of the various condition-specific instruments. Fourteen instruments use the statistical inference approach to select health states for valuation and model utilities, and eight instruments use a decomposed approach, which is either a pure MAUT approach or an approach combining statistical inference and MAUT. Three instruments value all health states defined by the descriptive system and thus require no modeling to produce utility scores for all health states. TTO, VAS, and SG and combinations of these were most commonly used to elicit utility scores. Twelve instruments have values elicited using TTO, six used VAS and SG, four used VAS alone, two used SG alone. Although VAS was used in 13 instruments and is the most commonly used technique, its usage differs across studies varying from valuing health states to valuing severity levels within a dimension or valuing different dimensions. Eight instruments cannot be used to estimate QALYs as they are not anchored onto the full health-dead 1–0 scale required to estimate QALYs. Half of the measures are valued using only a general population sample, with 10 measures valued by patients and one by patients and carers (using hypothetical states not own state). Valuation studies across all instruments have only been conducted in the UK, the US, Netherlands, and Canada with the majority of instruments valued only in the UK.

Validity

Comparison of Performance to Original Measure

There is little published analysis examining the extent of information loss arising from moving from the original non-preference-based condition-specific measure to a smaller condition-specific MAU instrument. However, evidence reviewed recently by John Brazier and colleagues suggests that there is either no information loss or a minimal degree of information loss when examining patient data sets for instruments produced for asthma, cancer, common mental health problems, and overactive bladder in terms of discrimination across severity group and responsiveness to change over time. This is reassuring given the rationale for deriving a MAU instrument from an existing measure is to benefit from its relevance and sensitivity, and means that this informational advantage is retained in the process of deriving a MAU instrument from the original measure.

Comparison of Performance to Generic MAU Instruments

Overall there is limited evidence comparing the performance of condition-specific and generic MAU instruments and it is mainly concerned with psychometric tests such as known group differences and responsiveness. However, analyses comparing generic EQ-5D to instruments produced for asthma, cancer, and common mental health problems found that the condition-specific MAU instruments performed better than EQ-5D regarding discriminative validity across severity groups. Figure 2 plots pair-wise comparisons of EQ-5D and these condition-specific MAU instruments, demonstrating that the condition-specific MAU instruments have a narrower range of utility scores than EQ-5D but there is a large dispersion of condition-specific MAU scores for each EQ-5D score (and vice versa). Unlike EQ-5D the condition-specific MAU instruments did not suffer from ceiling effects (where a large proportion of respondents report themselves as in full health and are clustered at the top end of the scale), suggesting they are more responsive for patients at the upper end of HRQL. It is also clear that there are differences between the condition-specific MAU instruments and EQ-5D at the observational level, and the condition-specific utility scores are typically higher with the exception of observations in full health on EQ-5D. Although research found that the condition-specific MAU instruments were better at discriminating between groups with different severity, they were comparable to EQ-5D regarding responsiveness to change following treatment (although for responsiveness this is based on little data owing to limited data availability). Typically, mean change over time and differences between severity groups were smaller for the condition-specific MAU instruments but with smaller standard deviation that improved precision in comparison to EQ-5D. This reduced uncertainty in utility scores for different time periods and severity groups is important for trials. However, the smaller mean change over time or across groups found using these condition-specific MAU instruments may potentially indicate that interventions are less cost–effective. Further research is needed to determine whether these findings are generalizable to all condition-specific MAU instruments. It should be noted that as a MAU instrument contains a subset of items from the original measure it will only offer an improvement on a generic MAU instrument if the original non-preference-based condition-specific measure offers an improvement. Therefore, the development of condition-specific MAU instruments from existing measures should be limited to measures already shown to be more responsive and valid.

Table 1 Condition-specific MAU instrument descriptive systems

Condition: name of MAU instrument (where available)	*Non-preference-based measure*	*Number of dimensions*	*Severity levels*	*No. of states defined by system*	*Dimensions*
Amyotrophic lateral sclerosis (ALS): ALS utility index	Amyotrophic Lateral Sclerosis Functioning Rating Scale – Revised (ALSFRS-R)	4	5–6	750	Speech and swallowing; eating, dressing and bathing; leg function; respiratory function
Asthma: Asthma Symptom Utility Index (ASUI)	N/A	5	10	100 000	Cough; wheeze; shortness of breath; awakening at night; side effects of asthma treatment
Asthma: AQL-5D	Asthma Quality of Life Questionnaire (AQLQ)	5	5	3 125	Concern about asthma; shortness of breath; weather and pollution stimuli; sleep problems; activity limitation
Cancer: EORTC-8D	EORTC QLQ-C30	8	4–5	81 920	Physical functioning; role functioning; pain; emotional functioning; social functioning; fatigue and sleep disturbance; nausea; constipation and diarrhea
Common mental health problems: CORE-6D	Clinical outcomes in routine evaluation – outcome measure (CORE-OM)	6	3	729	Functioning – close relationships; functioning – social relationships; functioning – general; symptoms – anxiety; risk/harm to self; physical health
Dementia: DEMQOL-U	DEMQOL (self-report)	5	4	1 024	Positive emotion; memory; relationships; negative emotion; loneliness
DEMQOL-Proxy-U	DEMQOL-Proxy (carer proxy-report)	4	4	256	Positive emotion; memory; appearance; negative emotion; loneliness
Diabetes: Diabetes Utility Index	Audit of diabetes-dependent quality of life (ADDQoL) plus additional items	5	3–4	768	Physical ability and energy level; relationships; mood and feelings; enjoyment of diet; satisfaction with managing diabetes
Epilepsy: NEWQOL-6D	Quality of life in newly diagnosed epilepsy measure (NEWQOL)	6	4	4 096	Worry about attacks; depression; memory; cognition; stigma; control
Erectile (dys)functioning	IIEF Index of erectile function	2	5	25	Ability to attain an erection sufficient for satisfactory sexual performance; ability to maintain an erection sufficient for satisfactory sexual performance
Flushing	Flushing symptoms questionnaire (FSQ)	5	4–5	2 500	Redness of skin; warmth of skin; tingling of skin; itching of skin; difficulty sleeping
Glaucoma: Glaucoma Utility Index (GUI)	Glaucoma profile instrument	6	4	4 096	Central and near vision; lighting and glare; mobility; activities of daily living; eye discomfort; other effects
Handicap: London Handicap Scale	International classification of impairments, disabilities, and handicaps (ICIDH)	6	6	46 656	Handicap mobility; occupation; physical independence; social integration; orientation; economic self-sufficiency
Head and neck cancer	N/A	8	5	390 625	Social function; pain; physical appearance; eating problems; speech problems; nausea; donor site problems; shoulder function

(*Continued*)

Table 1 Continued

Condition: name of MAU instrument (where available)	*Non-preference-based measure*	*Number of dimensions*	*Severity levels*	*No. of states defined by system*	*Dimensions*
Lung cancer	FACT-L	6	2	64	Physical; social/family; emotional; functional; symptoms – general: symptoms – specific
Menopause	Menopause-specific quality of life questionnaire	7	3–5	6 075	Hot flushes; aching joints/muscles; anxious/frightened feelings; breast tenderness; bleeding; vaginal dryness; undesirable androgenic signs
Menorrhagia	N/A	6	4	4 096	Practical difficulties; social life; psychological health; physical health; working life; family life
Minor oral surgery	N/A	5	4	1 024	General health and well being; health and comfort of mouth, teeth, and gums; impact on home/social life; impact on job/studies; appearance
Overactive bladder: OAB-5D	OABq overactive bladder questionnaire	5	5	3 125	Urge to urinate; urine loss; sleep; coping; concern
Pediatric asthma: Pediatric Asthma Health Outcome Measure (PAHOM)	N/A	3	2–3	12 – but only 10 are valid	Symptoms; emotion; activity
Pediatric atopic dermatitis	Unnamed questionnaire on atopic dermatitis	4	2	16	Activities; mood; settled; sleep
Parkinson's disease	N/A	2	2–5	10	Disease severity; proportion of the day with 'off-time' (impact on QOL due to condition covering domains: social function, ability to carry out daily activities, psychological function)
Pulmonary hypertension	Cambridge pulmonary hypertension outcome review (CAMPHOR)	4	2–3	36	Social activities; traveling; dependence; communication
Rhinitis: Rhinitis Symptom Utility Index (RSUI)	N/A	5	10	100 000	Stuffy or blocked nose; runny nose; sneezing; itchy watery eyes; itchy nose or throat
Sexual quality of life: SQOL-3D	Sexual quality of life questionnaire (SQOL)	3	4	64	Sexual performance, sexual relationship, sexual anxiety
Stroke	N/A	10	3	59 049	Walking; climbing stairs; physical activities/sports; recreational activities; work; driving; speech; memory; coping; self-esteem
Urinary incontinence	The King's health questionnaire (used for urinary incontinence and lower urinary tract symptoms)	5	4	1 024	Role limitations; physical limitations; social limitations/family life; emotions; sleep/energy
Venous ulceration	N/A	5	3–5	720	Pain; mobility; mood; smell; social activities
Vision/visual impairment: VisQoL/AQoL-7D	N/A	6	5–7	45 360	Physical well being; independence; social well being; emotional well being; self-actualization; planning and organization

Table 2 Condition-specific MAU instrument valuation

Condition: name of MAU instrument (where available)	*Theory and model type*	*Preference elicitation technique*	*Anchored on 1–0 full health-dead scale*	*Population*	*Country*
Amyotrophic lateral sclerosis (ALS): ALS Utility Index	Decomposed – multiplicative	VAS both for each level of each dimension alone and health states, and SG	Yes	General population	US
Asthma: Asthma Symptom Utility Index (ASUI)	Decomposed – multiplicative	VAS and SG both for states and for the levels per dimension	No	Patients	US
Asthma: AQL-5D	Statistical – additive	TTO	Yes	General population	UK
Cancer: EORTC-8D	Statistical – additive	TTO	Yes	General population	UK
Common mental health problems: CORE-6D	Statistical – additive	TTO	Yes	General population	UK
Dementia: DEMQOL-U DEMQOL-Proxy-U	Statistical – additive	TTO	Yes	General population	UK
Diabetes: Diabetes Utility Index	Decomposed – multiplicative	VAS and SG	Yes	Patients	US
Epilepsy: NEWQOL-6D	Statistical – additive	TTO	Yes	General population	UK
Erectile (dys)functioning	All states valued	TTO	Yes	General population and students	Netherlands
Flushing	Maps Rasch logit scores onto mean utilities – additive	TTO	Yes	General population	UK
Glaucoma: Glaucoma Utility Index (GUI)	Statistical – additive	DCE	No	Patients	UK
Handicap: London Handicap Scale	Statistical – additive	VAS	No	Patients	UK
Head and neck cancer	Decomposed – additive	VAS both for dimensions relative to each other and for the levels per dimension	No	Surgeons	UK
Lung cancer	Statistical – additive	VAS	Yes	General population	UK, Netherlands
Menopause	Statistical – additive	TTO	Yes	Patients	UK
Menorrhagia	Decomposed – additive	2 tasks: distribute 21 counters across the dimensions in proportion to their importance; VAS of the levels per dimension	No	Patients	UK
Minor oral surgery	Decomposed – additive	2 tasks: distribute 100 counters across the dimensions in proportion to their importance; VAS of the levels per dimension	No	Patients	UK
Overactive bladder: OAB-5D	Statistical – additive	TTO	Yes	General population	UK
Pediatric asthma: Pediatric Asthma Health Outcome Measure (PAHOM)	Power function used to convert VAS to SG, all states valued using VAS	VAS and SG	Yes	General population	US
Pediatric atopic dermatitis	All states valued	SG	Yes	General population	UK
Parkinson's disease	All states valued	VAS and SG	Yes	Patients	US
Pulmonary hypertension	Statistical – additive	TTO	Yes	General population	UK
Rhinitis: Rhinitis Symptom Utility Index (RSUI)	Decomposed – multiplicative	VAS and SG both for states and for the levels per dimension	No	Patients	US
Sexual quality of life: SQOL-3D	Statistical – additive	TTO	Yes	General population	UK
Stroke	Unclear	VAS	No	Patients and caregivers	Canada
Urinary incontinence	Statistical – additive	SG	Yes	Patients	UK
Venous ulceration	Statistical – additive	TTO	Yes	General population	UK
Vision/visual impairment: VisQoL/AQoL-7D	Decomposed – multiplicative	TTO, VAS for the levels per dimension	Yes	Unclear	Unclear

Notes: Statistical, statistical inference; decomposed, MAUT or combination of MAUT and statistical inference. Preference elicitation technique is reported only if it was used to produce the recommended utility scores for all health states.

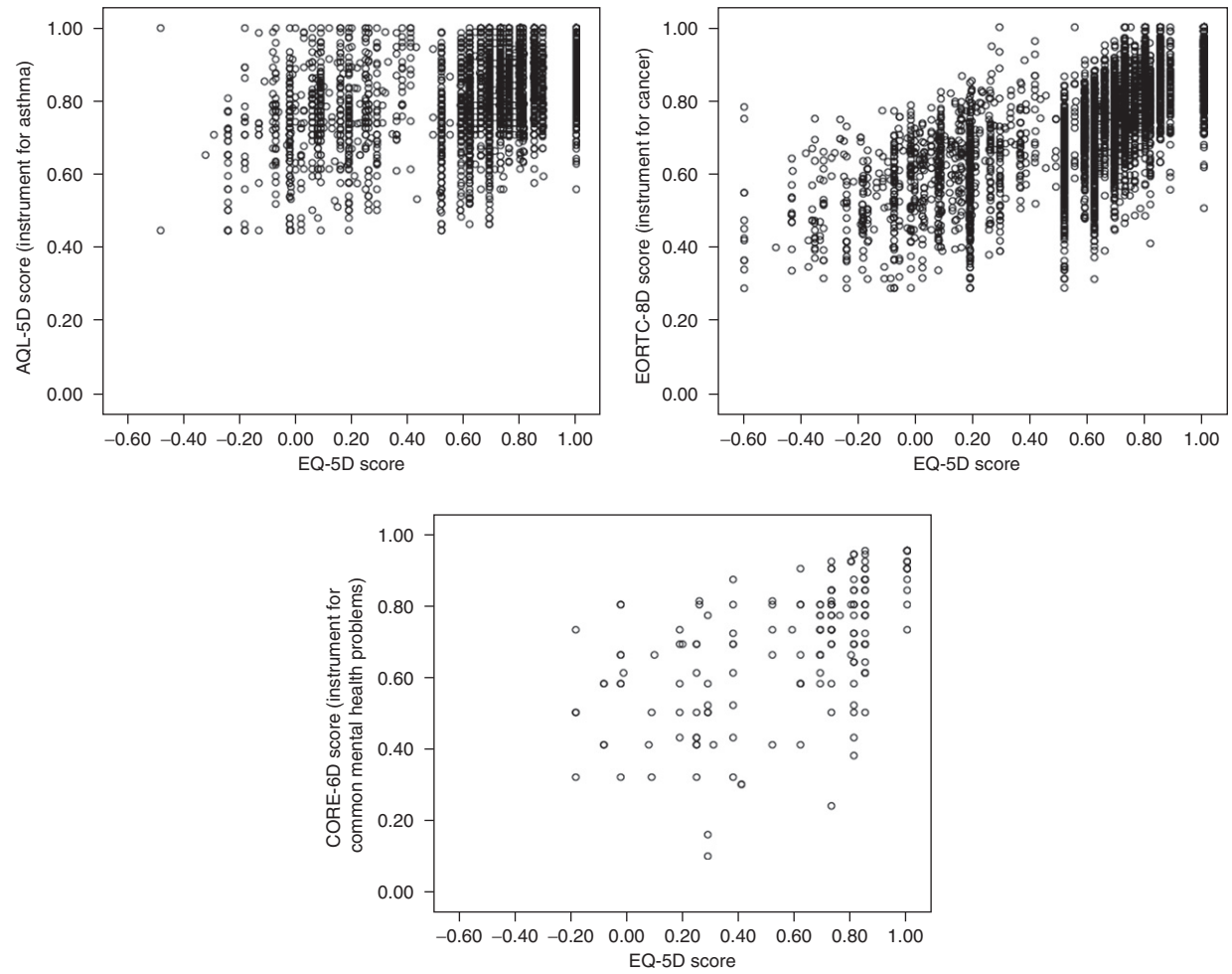


Figure 2 Pair-wise comparisons of 3 condition-specific MAU instruments and generic EQ-5D.

Other Issues

Condition-specific MAU instruments have been criticized on the grounds of focusing effects. It has been argued that respondents may provide lower utility scores for health states with only condition-specific dimensions as they are focusing on the problems presented rather than judging these relative to other generic dimensions of health that are not impaired, such as mobility or self-care. But this is not supported by evidence of the kind shown in Figure 2.

Some evidence suggests that the inclusion of the condition label in the health state descriptions used to elicit utility scores can itself affect results (e.g., a 'cancer' label was found to lower values). The inclusion of these condition labels is often unavoidable as the condition is embedded into the descriptive system (e.g., take the dimension 'concern about asthma') and is an important factor that may affect health state utility scores for condition-specific MAU instruments and question comparability to values produced by generic MAU instruments.

It has been argued that condition-specific MAU instruments are unable to capture comorbidities or side effects because of their focus on the condition. Table 1 indicates that whereas this may be a concern for some instruments with a descriptive system largely focused on symptoms related to the condition, this is unlikely to be an important concern for many instruments that cover a broad range of functionings (e.g., the EORTC-8D for cancer).

In relation to focusing effects and the inability to capture side effects and comorbidities, there is a concern that elicited utility scores may be affected if the descriptive system does not contain all important dimensions. There is some evidence supporting this concern, where it was found that adding a generic dimension to an existing condition-specific MAU instrument affected the utility scores for the condition-specific dimensions. This suggests that condition-specific MAU instruments should contain all important dimensions for that condition.

Future Developments

Ongoing research will identify where generic MAU instruments are insufficient or inappropriate and condition-specific MAU instruments are appropriate. Further research examining the impact of using either generic or condition-specific MAU instruments in economic evaluation is encouraged. Future

research should also examine the role of the descriptive system in valuation studies as this has largely been ignored to date. This research would determine the importance of focusing effects, condition labeling, and missing dimensions. These findings would indicate the comparability of valuation studies undertaken for generic and condition-specific MAU instruments and the accuracy of health state utility scores for condition-specific MAU instruments.

See also: Measuring Equality and Equity in Health and Health Care. Multiattribute Utility Instruments and Their Use. Quality-Adjusted Life-Years. Valuing Health States, Techniques for

References

U.S. Department of Health and Human Services Food and Drug Administration (FDA) (2009). *Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims*. MD: FDA.

Young, T., Yang, Y., Brazier, J. E., Tsuchiya, A. and Coyne, K. (2009). The first stage of developing preference-based measures: Constructing a health-state classification using Rasch analysis. *Quality of Life Research* **18**, 253–265.

Further Reading

Brazier, J. E., Ratcliffe, J., Tsuchiya, A. and Solomon, J. (2007). *Measuring and valuing health for economic evaluation*. Oxford: Oxford University Press.

Brazier, J., Rowen, D., Mavranezouli, I., et al. (2012). Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). *Health Technology Assessment* **16**(32), 1–114.

Brazier, J., Yang, Y., Tsuchiya, A. and Rowen, D. (2010). A review of studies mapping (or cross walking) from non-preference based measures of health to generic preference-based measures. *European Journal of Health Economics* **11**, 215–225.

Drummond, M., Brixner, D., Gold, M., et al. (2009). Toward a consensus on the QALY. *Value in Health* **12**(Supplement 1), S31–S35.

Revicki, D. A., Leidy, N. K., Brennan-Diemer, F., Thompson, C. and Togias, A. (1998). Development and preliminary validation of the multiattribute Rhinitis Symptom Utility Index. *Quality of Life Research* **7**, 693–702.

Yang, Y., Brazier, J., Tsuchiya, A. and Coyne, K. (2009). Estimating a preference-based single index from the overactive bladder questionnaire. *Value in Health* **12**, 159–166.

Noncommunicable Disease: The Case of Mental Health, Macroeconomic Effect of

M Knapp, London School of Economics and Political Science, London, UK, and King's College London, Institute of Psychiatry, London, UK
V Iemmi, London School of Economics and Political Science, London, UK

Mental Health

Mental health problems are among the most complicated and challenging of all illnesses, with considerable economic implications. It is conventional to distinguish between common mental disorders (including depression, generalized anxiety disorder, panic disorder, obsessive–compulsive disorder and post-traumatic stress, with an overall prevalence of 15–20%) and severe mental disorders (particularly schizophrenia and bipolar disorder, with a combined prevalence of approximately 2–3%). Alcohol and drug abuse are usually included under the mental health heading, as is suicide or suicidal ideation. The World Health Organization (WHO) also classifies epilepsy as a mental disorder, although many countries' national health systems do not (and it is not included in this article; intellectual disabilities and neurological disorders are also excluded). It is estimated that approximately 6% of children and young people aged under 18 years have behavioral problems serious enough to be classified as psychiatric disorders, and 4% have emotional disorders. Alzheimer's disease and other dementias become increasingly prevalent with age: 5% for people aged 65 years or over but 20% for those aged 85 years or over.

According to the WHO, mental, substance use, and neurological disorders as a group account for approximately 14% of the global burden of disease (measured in disability-adjusted life years), which is roughly 30% of the total global burden of noncommunicable disease. Both proportions are expected to grow over time, particularly in low- and middle-income countries (LAMICs).

The large disease burden follows from the high prevalence and chronic course of most mental disorders and is associated with high costs (see Economic Impacts). Few mental disorders can be cured, although symptom alleviation through evidence-based treatment is a realistic goal for many people. Primary prevention of illness is achievable for some common mental disorders and some people, although illness etiology remains only partially understood. What distinguishes mental from other illnesses is a particularly troublesome combination of at least four interconnected features: links to suicide and self-harm; associations with dangerous behavior; widespread public fear, stigma and discrimination; and restrictions to individual choice and liberty because of assumed or ascribed inability or danger.

There are wide gaps between underlying and treated prevalence, particularly in LAMICs where WHO estimate that 76–85% of serious cases have received no treatment in the previous 12 months; the figure in high-income countries is said to be 35–50%. This is why the WHO launched the Mental Health Gap Action Program (mhGAP) to support and encourage strategic planners and policy makers – as well as international donors – to tackle the 'burden' of untreated disorders.

Poor access to evidence-based care and treatment is a key factor in the overall disease burden. One reason is simply that policy makers fail to allocate sufficient funds to mental health programs, or that voluntary insurance coverage excludes some or all mental disorders. But there are other 'resource barriers' such as the unequal distribution of services and available treatments across regions, socioeconomic groups, genders, and age bands. Another is the inappropriateness of certain types of investment (e.g., the large, usually decrepit, dehumanizing institutions that still dominate mental health systems in some countries) and the consequent inflexibility of available resources to meet needs most cost-effectively. A further problem is a basic lack of information about what works in terms of symptom alleviation and quality-of-life enhancement or on the resource implications. In countries where health systems are primarily reliant on voluntary insurance or out-of-pocket payments, poverty is a major barrier.

Economic Impacts

Mental disorders are defined by their clinical symptoms and so impact heavily on health systems. But they have consequences across many life domains, leading to potentially wide-ranging needs for support from social care, housing, employment, criminal justice, income support, or other systems. The direct costs of treatment and care are high, but the indirect costs of mental disorders can be higher. For example, calculations for Europe in 2010 suggest that indirect costs (mainly from unemployment, absenteeism, and presenteeism) account for 38% of the total cost of anxiety disorders, 63% for mood disorders, and 69% for psychotic disorders.

There can be high opportunity costs for families because the need for unpaid care and support interferes with employment and leisure, as well as out-of-pocket payments if families subsidize treatment expenses. Estimates published by Alzheimer's Disease International in 2010 indicate that unpaid care from family members and other unpaid carers accounts for 58% of the total costs of dementia in low-income countries, 65% in lower middle-income countries, and 40% in high-income countries. These indirect costs are largely hidden but crucial inputs. They are often overlooked in policy frameworks; this is a dangerous strategy given rapidly ageing populations, the resulting rapidly growing prevalence of dementia, and the dwindling number of family carers (because of trends toward smaller families and higher female labor force participation rates).

 doi:10.1016/B978-0-12-375678-7.00613-1

For the wider society, economic impacts can include the victim, fear, and criminal justice system costs of acquisitive crime by people with serious substance misuse disorders, violent crime by people experiencing florid psychotic episodes, and suicide and self-harm by people experiencing severe depression. Although in some societies there is exaggerated attribution or fear of these criminal activities, it is nevertheless the case that, for instance, 5–15% of homicides in high-income countries are committed by people with psychosis. Responses can be both appropriate and inappropriate: a high proportion of people in prison have untreated mental disorders, in many cases before their incarceration, whereas most countries have legal structures in place to compulsorily detain and treat individuals at times of crisis. These powers can too easily be abused, as in the Soviet Union and Nazi Germany, but in less dramatic ways the basic human rights of people with mental disorders are still denied – often trampled – in many societies in the present day.

Because of the enduring nature of most mental disorders, economic impacts can be seen across much of the life course, including poor employment outcomes, low incomes, continuing high use of mental health services, continuing antisocial and criminal activity, and difficulties with personal relationships. Behavioral and emotional disorders in childhood have impacts well into middle adulthood and possibly beyond.

An immediate corollary of this multiplicity and durability of economic impacts – following from the multiplicity and chronicity of needs associated with some mental disorders – is for coordinated action across budgets and systems to avoid gaps and wasteful overlaps and particularly to ensure that resources work together effectively and efficiently. Silo budgets are almost clichéd but remain a substantial challenge, sometimes exacerbated by professional rivalry, narrowly framed performance assessment, and the slow churn of bureaucratic processes.

Mental Health, Employment, and Productivity

There are multiple, complex, two-way links between mental disorders and employment difficulties. People with mental disorders are at greater risk of unemployment, job insecurity, early retirement, absenteeism, presenteeism, and low salaries. However, stress, bullying, and other difficulties in the workplace are known risk factors for onset or exacerbation of common mental disorders. Psychoses, particularly schizophrenia, are most likely to emerge when people are in their late teenage years or early 20s, which is precisely when most people would make key investments in their human capital, and so these disorders have lifelong economic consequences. As it has been seen, lost productivity is the biggest contributor to the immediate costs of mental disorders but with wider consequences for household income, community prosperity, and national economic growth.

People with mental disorders – like most people – put great emphasis on employment, not only partly and obviously because it generates earnings, but also because it brings social status and social role, fosters social participation and networking, and is a major source of self-concept. Barriers to the employment of people with mental disorders will obviously include reduced abilities because of their symptoms, but endemic social stigma and widespread discrimination by employers are major challenges.

Given the close links between employment, income, personal debt, and poverty (see Mental Health, Poverty, and Debt), the complex downward spiraling relationship between mental disorders and work difficulties can be hard to break into. One set of responses, coming from strategic policy makers, is to create conditions so that employment opportunities are better, even if they are unlikely to match those available to the 'mentally well' population. There is also growing support for approaches that provide intensive support for employees with mental disorders and their employers in 'open employment' settings rather than traditional 'sheltered workshop' environments. The Individual Placement and Support approach, for example, has proved effective and cost-effective across many high-income countries, helping people with quite severe mental disorders to enter, remain, and sometimes thrive in the workforce, albeit with ongoing support in many cases.

Some countries have introduced legislation so that mental disorders are viewed in the same way as physical disability in terms of rights to employment and other opportunities. Some of these initiatives go hand in hand with attempts to address moral hazard concerns about criteria for income support eligibility and to counter what can easily become a dependency culture in groups with long-term conditions. Most people with mental disorders want to work are perfectly capable of working in appropriate settings, and derive therapeutic and other benefits from it. The alternatives are not only disabling, disempowering, and disadvantaging for those individuals but also costly for societies.

Employers face high productivity losses if key employees have mental disorders because losing skilled staff for short or long periods is both disruptive and expensive. Consequently, private businesses (and public and NGO sector employers, of course) have gains to make if there is appropriate preventive or ameliorative action. Employers in many countries have become more aware in recent years of the impacts on their profit margins of disrupted employment as a result of employees' poor health, particularly poor mental health. Perhaps more gradually, employers have also become more aware of the benefits of tackling some of the associated risk factors through workplace initiatives. Some risk factors for mental disorders are within their control, such as demands made on employees, opportunities for them to participate in decision-making, promotion prospects, harassment, and bullying. Workplace well-being programs and screening initiatives for stress and other approaches have a good evidence base in support. Although attention to the bottom line – the profit margin – should be enough of a motivation to introduce such programs in some companies (provided the risks and consequences are fully appreciated), small- and medium-sized enterprises may need support through tax incentives or health insurance deals.

Mental Health, Poverty, and Debt

Mental illness and poverty interact in vicious circles. Evidence from high- and low-income countries shows the close relationship between mental disorders and various measures of

individual economic disadvantage. Two hypotheses have been propounded. According to the social causation hypothesis, economic disadvantage such as poverty increases the risk of mental illness through augmented risk factors (e.g., financial stress, stigma, social exclusion, and malnutrition) and decreased protective factors (e.g., social capital, education). However, the social selection or 'drift' hypothesis argues that people with mental disorders have an increased risk of remaining or falling into poverty because of the costs of their treatment, lost or disrupted employment, and hence reduced earnings. The social causation explanation is more relevant for common mental disorders such as depression and anxiety, and the drift explanation more relevant for severe mental disorders such as schizophrenia, but the pathways are complex and evidence suggests causation in both directions for many people.

Poverty

The World Bank defines poverty as marked deprivation in well-being that includes monetary (e.g., income and consumption) and nonmonetary aspects (e.g., health, education, and housing), as captured by indices measuring one dimension (e.g., absolute and relative poverty) or many (e.g., the Multidimensional Poverty Index).

In high-income countries, poverty and unemployment are known to be associated with the maintenance of common mental disorders but apparently not their onset, whereas financial strain (personal debt) is associated with both. Several epidemiological studies have found positive associations between low socioeconomic status, on the one hand, and alcohol and substance misuse, and rates of schizophrenia and major depression, on the other hand. Suicide and parasuicide are strongly associated with socioeconomic deprivation. Children in the poorest households are much more likely to suffer from conduct disorders (severe antisocial behavior) and attention-deficit hyperactivity disorders than children from more affluent households. In LAMICs, common mental disorders have been found to be positively associated with low socioeconomic status, financial stress, low social class, low educational attainment, food insecurity, and bad housing conditions. More ambiguous associations have been found with low income, unemployment, underemployment, and low consumption.

Debt and Financial Instability

Following the recent global economic downturn, the relationships between mental disorders, personal debt, and financial instability have attracted considerable interest. Personal or household debt has a two-way association with mental health. In high-income countries, the impact of financial difficulties on depression depends in part on the type of debt and whether it is 'manageable', but there is apparently no direct association with anxiety disorders or nonspecific mental disorders. For example, a positive association has been found between the onset of mortgage indebtedness and rent arrears and poor mental health, with men more likely to be affected in the short term and women in the long term. A negative association has been found between outstanding nonsecured debt and psychological well-being, but no association for secured debt (such as a mortgage on housing). In a LAMIC context, one study found higher rates of distress and suicide among farmers who went into debt as a result of the agricultural crisis in India, whereas another found a positive association between personal debt and suicide attempts in Pakistan.

Macroeconomic recession affects numerous determinants of mental health, such as employment and job security, productivity and earnings, socioeconomic status, and social cohesion. In high-income countries, slower economic growth and associated increases in unemployment are associated with higher suicide rates. A positive association has also been found between unemployment *duration* and suicide, with a noticeable increase in suicide risk shortly after mass layoffs. Macroeconomic fluctuations that lead to worsening market conditions appear to be associated with poorer mental health in those people least likely to be employed, in ethnic minority groups, and in men with lower educational attainment. A macroeconomic downturn affects not only the mental health of some adults but also of their children. The Asian economic crisis of 1997 was associated with an increase in suicide rates in the years that followed, as well as a widening of income-related mental health inequalities. Sociopolitical crises in Serbia and Belarus increased suicide rates, and socioeconomic upheavals in the Russian Federation were followed by increased rates of suicide and alcohol-related deaths. Agricultural crises are strong risk factors for depression and suicide attempts in South Asia.

Solutions

Acknowledging the bi-directional relationship between mental illness and poverty, interventions have targeted both the causes of mental disorders and the causes of poverty. Poverty-alleviation interventions, such as cash transfers and microcredit, have as a by-product the potential to address the social causation of mental illness. Mental health interventions, such as drug treatments, psychotherapy and community rehabilitation, might be primarily focused on symptom alleviation but in so doing can also address the drift into poverty. In LAMICs, the effect of mental health interventions on poverty and other economic outcomes has been found to be positive but not always significant, whereas evidence is inconclusive on the effect of financial poverty-alleviation interventions on mental illness. Asset promotion programs have mental health benefits, and the evaluation of the conditional cash transfer program *Oportunidades* in Mexico found a significant reduction in both depressive symptoms in mothers and behavioral problems in children. A microcredit intervention in South Africa *increased* perceived stress levels among recipients of small loans (both men and women) but decreased depressive symptoms in men. Debt advice and counseling services can decrease the risk of developing or exacerbating mental disorders.

Macroeconomics and Mental Health

Mental disorders have large economic impacts that are often spread well beyond health care systems as conventionally constructed. There are also persistent economic impacts across

long time periods. Employment and associated productivity difficulties are common, both for people with mental disorders and their families, while disorders in childhood and adolescence reduce levels of educational attainment. Hence, the macroeconomic consequences flowing from untreated or inadequately treated mental disorders can be considerable. There is also the strong likelihood that social and economic inequalities will widen.

From the other direction, there is also strong evidence that economic difficulties experienced by individuals increase the risk of mental disorder so that countries experiencing prolonged recession will likely see growth in the prevalence of common mental disorders. Becoming unemployed, remaining unemployed for long durations, experiencing a drop in earnings or other income, moving into personal debt in ways that cannot be managed, and experiencing housing problems can all lead to lower psychological well-being and resilience, more mental health needs and alcohol misuse, higher suicide rates, greater social isolation, and worsened physical health.

Efforts are already being made to break, weaken, or respond to the links between the state of the macroeconomy and mental health-related needs, although in most countries a lot more could and should be done. The available options are many and various. They include poverty-alleviation strategies, programs to help individuals with mental disorders to get jobs, early intervention to head-off the most damaging of personal and economic consequences of disorders, investments to build community and social capital, stronger social safety nets, antistigma campaigns, workplace initiatives, alcohol price increases, school-based schemes to tackle bullying and build resilience, and personal health budgets. And, of course, the options also include the more conventional pharmacological and psychosocial treatments, family psychoeducation, respite care, liaison services between medical specialties, and reorganization of care arrangements, each in turn requiring commitment of budgets to mental health systems from governments or other funders, followed by investment in suitably trained staff and other treatment inputs.

See also: Efficiency and Equity in Health: Philosophical Considerations. Efficiency in Health Care, Concepts of. Health and Health Care, Macroeconomics of. Impact of Income Inequality on Health. Long-Term Care. Mental Health, Determinants of. Public Health: Overview. Valuing Informal Care for Economic Evaluation

Further Reading

Bond, G. R., Drake, R. E. and Becker, D. R. (2012). Generalizability of the individual placement and support (IPS) model of supported employment outside the US. *World Psychiatry* **11**(1), 32–39.

Fitch, C., Hamilton, S., Bassett, P. and Davey, R. (2011). The relationship between personal debt and mental health: A systematic review. *Mental Health Review Journal* **16**(4), 153–166.

Funk, M., Drew, N., Freeman, M. and Faydi, E. (2010). *Mental health and development: Targeting people with mental health conditions as a vulnerable group*. Geneva, Switzerland: World Health Organization.

Gustavsson, A., Svensson, M., Jacobi, F., et al. (2011). Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology* **21**(10), 718–779.

Knapp, M., Funk, M., Curran, C., et al. (2006). Mental health in low- and middle-income countries: Economic barriers to better practice and policy. *Health Policy and Planning* **21**(3), 157–170.

Knapp, M., McCrone, P., Fombonne, E., Beecham, J. and Wostear, G. (2002). The Maudsley long-term follow-up study of child and adolescent depression: Impact of comorbid conduct disorder on service use and costs in adulthood. *British Journal of Psychiatry* **180**(1), 19–23.

Lund, C., De Silva, M., Plagerson, S., et al. (2011). Poverty and mental disorders: Breaking the cycle in low-income and middle-income countries. *Lancet* **378**(9801), 1502–1514.

McDaid, D. (2011). *Background document for the thematic conference on promotion of mental health and well-being in workplaces*. Luxembourg: European Communities.

Saxena, S., Thornicroft, G., Knapp, M. and Whiteford, H. (2007). Resources for mental health: Scarcity, inequity and inefficiency. *Lancet* **370**(9590), 878–889.

Wahlbeck, K. and McDaid, D. (2012). Actions to alleviate the mental health impact of the economic crisis. *World Psychiatry* **11**(3), 139–145.

Weich, S. and Lewis, G. (1998). Poverty, unemployment, and common mental disorders: Population based cohort study. *British Medical Journal* **317**(7151), 115–119.

WHO Regional Office for Europe (2011). *Impact of economic crises on mental health*. Copenhagen, Denmark: WHO Regional Office for Europe.

Wimo, A. and Prince, M. (2010). *World Alzheimer Report 2010: The global economic impact of dementia*. London, UK: Alzheimer's Disease International.

World Health Organization (2008). *Mental health gap action programme: Scaling up care for mental, neurological, and substance use disorders*. Geneva, Switzerland: World Health Organization.

Nonparametric Matching and Propensity Scores

BA Griffin, RAND Corporation, Arlington, VA, USA
DF McCaffrey, ETS, Princeton, NJ, USA

Potential Outcomes Framework

In problems studying the effects of a given intervention or treatment on health outcomes or health costs, it is often the case that researchers and analysts are faced with using observational study data to draw inferences. The use of such data gives rise to various concerns about biases that may influence treatment effect estimates. In the ideal world, one would be able to observe how an individual or agent would act under two scenarios – one where the agent had been treated and the other where the agent had not been treated. From here, estimating the effect of treatment on both the individual agent and population would be straightforward. Unfortunately, this is never a possibility in real-world settings.

To illustrate in more detail the difference between the ideal and real world, we can utilize the potential outcomes framework that has become popular in both statistics and econometrics in the past two decades. First, it is assumed that the study sample has N individuals. For each individual, we can define two potential outcomes, namely $Y_i(1)$ and $Y_i(0)$, which denote individual i's outcome under the treatment and control conditions, respectively, for $i=1,\ldots,N$. In the ideal world, we would be able to observe both $Y_i(1)$ and $Y_i(0)$ and we would directly know the effect of the treatment on individual i (namely, $Y_i(1)-Y_i(0)$). However, in the real world, it is usually observed that $Y_i=Y_i(Z_i)$, where Z_i denotes whether the ith individual actually received treatment ($Z_i=1$) or control ($Z_i=0$) in practice. The value of $Y_i(1-Z_i)$, commonly referring to as the counterfactual for the ith individual, is not observed. In most studies, researchers also observe (in addition to Y_i and Z_i) a set of pretreatment covariates or exogenous variables, X_i, which are assumed to be unaffected by treatment but that may predict the outcome Y_i as well as whether the ith individual received the treatment or control condition. The variables in X_i are also commonly referred to as confounding factors or confounders because excluding them from an analysis examining the effect of Z_i on Y_i may lead to results from observational data in which the differences in the values of X_i between groups with $Z_i=0$ and $Z_i=1$ are combined with or confound the true causal effects of Z_i on Y_i.

With the potential outcomes framework in hand, we can now introduce a number of different types of treatment effects that a researcher may be interested in estimating in addition to the effect of the treatment on the ith individual ($Y_i(1)-Y_i(0)$). First, it is often the case that researchers are interested in estimating the population average treatment effect (PATE), which is defined as $E[Y(1)-Y(0)]$, and tells us the impact of treatment on average over the entire population to which inferences are being drawn. Alternatively, it may be of interest to estimate the PATE for the treated (PATT), which is defined as $E[Y(1)-Y(0)|Z=1]$, and tells us the average effect of treatment on individuals like those who actually receive treatment.

To estimate the PATE, the primary quantity of interest would be the sample average treatment effect (SATE), which equals $\frac{1}{N}\sum_{i=1}^{N}(Y_i(1)-Y_i(0))$, and takes the average over the effect of treatment on all individuals in the sample. To estimate the PATT, the primary quantity of interest would be the SATE for the treated (SATT), which equals $\frac{1}{N_T}\sum_{i:Z_i=1}(Y_i(1)-Y_i(0))$, where N_T denotes the number of individuals who are observed to receive treatment. SATT takes the average of the effect of treatment on all individuals in the sample who were observed to receive treatment. Both SATE and SATT are consistent (unbiased) estimates of PATE and PATT, respectively. However, it is never the case that one would be able to compute the SATE and SATT as only one of the two potential outcomes for each individual in the sample is usually observed.

A natural next step in trying to estimate PATE and/or PATT in practice would be to simply take the mean of Y_i among those with $Z_i=1$ and compare it to the mean of Y_i among those with $Z_i=0$. Unfortunately, this will only work if one has treatment and control conditions that are well balanced with regard to the distribution of all pretreatment factors that are known to be associated with the outcome. Such balance is hard to find when researchers are faced with using data from observational studies that do not randomize individuals in the sample to the treatment and control conditions. Moreover, usually only a handful of the pretreatment factors known to be associated with Y are actually available (i.e., those in X). Thus before estimating PATE and PATT, additional statistical and econometric methods must be utilized to achieve balance on the available pretreatment covariates or confounding factors, X, and additional assumptions must be made to ensure there is no residual confounding from imbalances in pretreatment covariates which were not observed. The rest of this article focuses on describing these assumptions and methods.

Critical Assumptions: Unconfoundedness and Overlap

Before introducing various methods that are available to researchers to achieve balance between individuals in the treatment and control conditions on the observed pretreatment covariates, X, when trying to estimate PATE or PATT, it is important to understand two key assumptions that are required to obtain consistent treatment effect estimates when faced with observational data.

The first key assumption goes by various names in the literature. It requires in layman's terms that one has matched or balanced the individuals in the treatment and control conditions on values of X in such a way that there are no unobservable differences between the individuals in the two conditions after conditioning on X. In notational terms, we write this as $(Y(1), Y(0))\perp Z|X$. The implication of this assumption is that systematic differences in outcomes between

 doi:10.1016/B978-0-12-375678-7.00711-2

individuals with the same values of the pretreatment covariates are solely attributable to treatment. This assumption is commonly referred to as either the unconfoundedness assumption, selection of the observables, the conditional independence assumption, or ignorable treatment assignment. Throughout the remainder of this article, we will refer to it as the unconfoundedness assumption.

In principle, the unconfoundedness assumption is untestable. However, a number of approaches have been proposed that are useful for addressing its credibility, which revolve around the idea of doing sensitivity analyses to assess how sensitivity treatment effect estimates are to specific types of deviations from the unconfoundedness assumption.

The second key assumption required for consistent estimates of PATE or PATT using the methods discussed is the overlap assumption. The overlap assumption is formally written as $0 < \Pr(Z = 1|X) < 1$. The implication is that one needs sufficient overlap in the distributions of the observed pretreatment covariates between individuals who are in the treatment and control conditions to be able to identify consistent estimates of PATE or PATT. Checking overlap is usually done by comparing the distribution of the estimated probability of receiving treatment conditional on X (commonly called the propensity score) among those in the treatment and control conditions to see if there is sufficient overlap in the estimated values of the measure; however, some promising work has been published which describes more formal ways to check overlap in the covariate distributions of the two conditions. Checking overlap is a critical step when estimating PATE and PATT. It is important to note that all is not necessarily lost if overlap is lacking. Assuming it is meaningful to do so, one may limit inferences to the average effect of treatment for the subset of the pretreatment covariate space where there is overlap.

Nonparametric Matching Methods

With our two key assumptions in hand, we now turn our attention to understanding ways in which researchers can try to balance the treatment and control conditions on the distributions of pretreatment covariates, X, so that consistent estimates PATE or PATT can be obtained. One of the most commonly used approaches for estimating PATE and PATT is to match individuals in the treatment and control conditions on values of X so that one can compare outcomes across pairs of matched treatment and control individuals with similar distributions of pretreatment covariates. Methods for creating meaningful matchings are numerous in the literature. This section focuses on the use of nonparametric matching techniques as nonparametric matching provides a method by which there are no restrictions on the functional form of the relationship between the outcome (Y), the treatment indicator (Z), and the pretreatment covariates (X). We begin by describing common techniques for one-to-one or pair matching (i.e., matching one control individual to one treatment individual) and then focus on the more elaborate techniques available for matching more than one individual to a given treatment or control individual.

The simplest case of one-to-one matching is exact matching which selects one control individual and one treatment individual so that they match exactly on X. Exact matching is only really feasible when the pretreatment covariates are all discrete (i.e., binary or categorical) and/or when one has only a handful of pretreatment covariates that need to be controlled for. In most other cases of one-to-one matching, the idea is to match one control individual with one treatment individual such that they have pretreatment covariates X whose values are within the same small neighborhood of each other. This type of matching (to no surprise) is commonly referred to as neighborhood matching or nearest neighbor matching.

The first step for any type of matching scheme (exact or neighborhood) is coming up with a meaningful distance metric that summarizes how close one individual is to another in terms of their values of X. There are generally three core types of distance metrics: categorical, caliper, and quadratic (including Mahalanobis distance). In each type, one begins by defining the distance between individual a in the treatment condition and individual b in the control condition using a specific formula. Here, the distance between individual a and individual b is denoted by δ_{ab}. Categorical distance metrics are usually defined such that $\delta_{ab} = 1$ if the covariates are an exact match and $\delta_{ab} = \infty$ otherwise. Caliper distance metrics are defined such that $\delta_{ab} = 1$ if $|X_{aj} - X_{bj}| \leq c_j$ for all j where $j = 1, \ldots, J$ and J denotes the total number of pretreatment covariates and $\delta_{ab} = \infty$ otherwise. Quadratic distance metrics are defined such that $\delta_{ab} = (X_a - X_b)D(X_a - X_b)^T$ and thus can take on any real number; Mahalanobis distance which defines D to the inverse of the variance–covariance matrix of X is an example of a quadratic distance.

Once the distance metric is defined, it is then computed for all possible pairs (a,b) of individuals, one in the treatment group and the other in the control condition and then an optimization algorithm selects the pairs that minimize the sum of δ_{ab} across all pairs. From there, PATE and PATT follow directly by computing the treatment effect estimate for each pair and averaging across these estimates. The key difference in obtaining an estimate of PATE and PATT is in how the matching is done. For obtaining an estimate of PATT, one would aim to match each treatment case to a single control case and throw away all other controls which were not matched. For obtaining PATE, one would try to optimally match across both conditions without requiring that one condition is favored more than the other.

There is an extensive number of methods available for implementing one-to-one matching, including canned programs in most standard statistical software packages. However, there has also been extensive work done on matching algorithms which go beyond one-to-one matching, including matching one treatment individual with multiple controls (say two or three for each treatment individual) or matching each treatment individual to a variable number of controls in such a way that all control individuals are utilized. The ideas for these methods build off the theory described earlier in the article for one-to-one matching. However, it is important to note that matching methods that match a fixed number of controls to a single treatment individual tend to be criticized for being overly restrictive and for discarding data (namely, the unmatched control individuals). Various authors have shown

that the use of a more flexible method, often called full matching, which makes use of all individuals in the data by forming a series of matched sets in which each set has either one treatment individual and multiple control individuals or one control individual and multiple treatment individuals is particularly effective at reducing bias due to confounding from pretreatment variables.

The section is ended by touching on the issue of how best to choose which pretreatment covariates that one should use in the matching algorithms described earlier in the article. It is generally recommended that variables chosen be predictive of both the outcome, Y, and the likelihood that the individual receives treatment, $\Pr(Z=1)$. Inclusion of irrelevant pretreatment covariates has the potential to impact both consistency and precision of treatment effect estimates.

Propensity Score-Based Methods

There is a second class of methods that are commonly used by researchers to handle the differences in the observed pretreatment covariate distributions among the treatment and control conditions. The class of methods differs from the methods presented in the previous section on Nonparametric Matching Methods because instead of matching on X directly, one matches on the predicted probability of receiving treatment, given the observed pretreatment covariates, commonly referred to as the propensity score. Matching on the propensity score as opposed to X greatly reduces the dimensionality problem of the nonparametric matching methods described earlier in the article. The theory shows that consistent estimation of PATE and PATT is possible when the treatment and control conditions are balanced with respect to the propensity score because balancing the two conditions on the propensity score in theory also balances the two conditions with respect to X.

Formally, the propensity score is denoted by $e(x) = \Pr(Z=1|X=x) = E[Z|X=x]$ and let $\hat{e}(x)$ denote the estimated propensity score (see below in this section for estimation). Various matching techniques (similar to those described earlier in the article) have become available for the case of trying to match treatment and control conditions on the propensity score in order to obtain consistent estimates of PATE and PATT. Each one is fundamentally based on the same ideas as the techniques for nonparametric matching on X, which are described earlier in the article except that things are much simpler when one only has to match on the single quantity of the propensity score rather than the entire vector of pretreatment covariates in X.

The most common approach for matching on the propensity score continues to be the nearest neighbor matching where a fixed number of controls, say k, are matched to one treatment individual and an optimal matching algorithm (as opposed to a greedy matching algorithm) are utilized to estimate PATT (as matching is driven by treatment individuals). The popular use of the nearest neighbor matching is likely a consequence of these algorithms being readily available in statistical software packages. However, as stated earlier in the article, such matching is suboptimal to full matching where a variable number of control individuals can be matched to a single treatment individual and a variable number of treatment individuals can be matched to a single control individual. The idea is that full matching forms the matched sets in an optimal way on the basis of the propensity score so that individuals from either condition who do not have many similar matches will not be forced to have bad matches as may happen when forcing all individuals to have a fixed number of controls. Unfortunately, software is not readily available to implement full matching in most software packages; the optmatch package available in R appears to be the only canned software currently available for implementing full matching.

The class of propensity score methods used to estimate PATE and PATT are not limited to matching. Commonly, both weighting by and subclassification on the propensity score are also used to obtain consistent estimates of PATE and PATT. With regard to weighting, one can consistently estimate PATE using the following formula:

$$\text{EATE} = \sum_{i=1}^{N} Y_i Z_i w_i / \sum_{i=1}^{N} Z_i w_i - \sum_{i=1}^{N} Y_i (1-Z_i) w_i / \sum_{i=1}^{N} (1-Z_i) w_i$$

where $w_i = 1/\hat{e}(X_i)$ for those with $Z_i=1$ and $w_i = 1/(1-\hat{e}(X_i))$ for those with $Z_i=0$. This estimate of PATE takes the difference in weighted means between those in the treatment and control conditions where the weights are set equal to the reciprocal of the estimated probability that an individual received the condition that they actually received. Such weights serve to make the distribution of observed pretreatment covariates, X, in each condition (treatment and control) look similar to the population's distribution of X. The corresponding estimate of PATT is

$$\text{EATT} = \sum_{i=1}^{N} Y_i Z_i / \sum_{i=1}^{N} Z_i - \sum_{i=1}^{N} Y_i (1-Z_i) w_i / \sum_{i=1}^{N} (1-Z_i) w_i$$

where $w_i = \hat{e}(X_i)/(1-\hat{e}(X_i))$. This estimate for PATT takes the unweighted mean for the treatment condition and subtracts off the weighted mean for the control condition where the weights for the control condition equal the odds of an individual in the control condition being in the treatment condition. This weight serves to upweight control individuals who look more like individuals in the treatment condition and downweight those who do not in order to yield a consistent estimate of PATT.

With regard to subclassification techniques based on the propensity score, the idea is to estimate the treatment effect of interest within subclasses that are defined based on the values of $\hat{e}(x)$ and to then aggregate across subclasses in a meaningful way to obtain final estimates of PATE or PATT. For example, it is generally the case that separate regression models on the outcome or mean differences between the two conditions are computed within each subclass and then results are aggregated across the subclasses by weighting those estimates by the subclass sample sizes if one is interested in PATE or the proportion of treatment cases in the subclass if one is interested in PATT.

A key issue for matching on, weighting by, or subclassification on the propensity score, is how best to estimate the propensity score. In most applications, simple logistic regression (or parametric) models are utilized to estimate the

propensity score. The treatment indicator, Z, is regressed on the pretreatment covariates X, possibly with or without interactions between the pretreatment variables and with limited attention paid to model selection procedures. In spite of the apparent ease of using logistic regression to estimate the propensity score, it is not the recommended method in the statistical and econometrics literature. Instead, use of more sophisticated semiparametric and nonparametric techniques are recommended. For example, it has been well established that machine learning methods outperform the use of simple logistic regression models in the binary treatment case. Machine learning methods work in an iterative fashion to fit nonparametric models to the predicted probability of receiving treatment. Use of conventional methods for estimating propensity scores such as logistic regression are less flexible than the machine learning techniques and typically require ad hoc variable selection procedures to reduce the number of degrees of freedom required by the pretreatment covariates and their interaction terms in the model. Such variable selection procedures risk biasing estimates of treatment effects because they can incorrectly omit covariates that are important to treatment selection. Similarly, variable selection procedures and the typical assumptions common in conventional modeling approaches risk model misspecification of the functional form of the relationship between the covariates and the treatment indicator, which can lead to very large biases in treatment effect estimate.

One example of a flexible, nonparametric machine learning technique that has been frequently utilized in the literature is the Generalized Boosted Model (GBM). GBM estimates the propensity score model using a flexible estimation method that adjusts for a large number of pretreatment covariates and which adaptively captures the functional form of the relationship between the pretreatment covariates and treatment selection with less bias than traditional approaches.

Use of propensity score weights has been shown to have a number of advantages over matching or subclassification techniques based on the propensity score. In comparison to one-to-one or k-to-one matching techniques which throw away observations that do not match, weighting includes all observations in the outcomes analysis. In comparison with subclassification techniques that require fitting multiple regression models to the outcome (one within each class), weighting techniques only fit one (weighted) regression model to the outcome which greatly minimizes variable selection issues and the need to tinker with the functional form. The main disadvantage of weighting is that it can be less efficient than matching or subclassification, particular when good weights are difficult to obtain.

Doubly Robust Methods

We end this article by describing doubly robust methods that are commonly used to consistently estimate PATE and PATT. To put the goal of doubly robust methods in context, one must note that historically, the traditional approach for estimating PATE and PATT using observational study data was to utilize multivariable parametric regression models of the outcome (Y), which regress Y on the treatment condition indicator (Z), and the observed pretreatment covariates (X), which were concerned to be potential confounders of the relationship between Y and Z. However, the assumptions of linearity (additivity) for the effects of the pretreatment covariates on Y was considered too rigid of an assumption and drew into question whether such models can yield consistent estimates of PATE and PATT giving raise to the nonparametric and propensity score methods described earlier in the article for obtaining consistent estimates of PATE and PATT. Although the methods described do not do any direct modeling of the outcome, they do come with their own assumptions, namely that the modeling done between the pretreatment covariates and treatment indicator are well specified. Doubly robust methods apply both techniques simultaneously (parametric regression modeling which control for X and matching or weighting the treatment and control conditions with respect to X) to produce consistent estimates of PATE or PATT as long as one of the two modeling approaches has been correctly specified. Thus, for example, use of both propensity score weighting and regression on the outcome with a parametric model leads to a doubly robust estimate of PATE or PATT because the estimates are consistent as long as one of the models is correct. Similarly, both matching on observed pretreatment covariates and controlling for those covariates in the outcome regression model fit to the matched pairs represent a doubly robust estimate of PATE and PATT assuming one of the two models is correct. Although doubly robust techniques can be less efficient if the parametric outcome model is correct, it is the final recommendation of this article that researchers go doubly robust whenever feasible to minimize the potential biases that could arise from using only a single technique.

See also: Models for Durations: A Guide to Empirical Applications in Health Economics. Survey Sampling and Weighting

Further Reading

Abadie, A. and Imbens, G. (2002). Simple and bias-corrected matching estimators for average treatment effects. *NBER Technical Working Paper No. 283*. Available at: http://www.nber.org/papers/t0283 (accessed 10.08.11).

Hansen, B. B. and Olsen Klopfer, S. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* **15**(3), 609–627.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review* **76**(1), 1–27.

Heckman, J. J., Ichimura, H. and Todd, R. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**, 261–294.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**(4), 1161–1189.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86**(1), 4–29.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**(4), 523–539.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* **29**, 337–346.

McCaffrey, D. F., Ridgeway, G. and Morrall, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9**(4), 423–434.

Ming, K. and Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics* **10**, 455–463.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association* **84**(408), 1024–1032.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Stuart, E. A. and Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* **44**(2), 395–406.

Zhao, Z. (2004). Using matching to estimate treatment effects data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* **86**(1), 91–107.

Relevant Websites

http://www.biostat.jhsph.edu/ ~ estuart/propensityscoresoftware.html
John Hopkins Bloomberg School of Public Health.

http://www.stat.lsa.umich.edu/ ~ bbh/#
University of Michigan.

Nurses' Unions

SA Kleiner, Cornell University, Ithaca, NY, USA

Introduction

Nurses' unions are widespread in the developed world. In the European Union, 90% of Denmark's nurses are members of a union (Danish Nurses' Organization, 2009), and in the UK, nurses (along with teachers and other professional workers) have the highest union density of any occupation (Metcalf, 2005). Organized labor for the nursing profession is prominent in non-EU countries as well, with nearly one-quarter of Australian nurses belonging to a union (Daly *et al.*, 2004), and 87% of Canadian nurses are represented by organized labor (Informetrica Limited, 2011), though OMalley (2012) reports the figure as lower (62%).

In the USA, the health care sector has been the most active sector of the economy for union organizing in recent years (NLRB, 2004). The percentage of health care practitioners reporting union membership has increased from 12.9% in 2000 to 13.3% in 2010 (Bureau of Labor Statistics, 2011), versus a corresponding decrease in overall union membership rates during that time period from 13.5% to 11.9% (Hirsch and Macpherson, 2013a). In addition, the number of unionized health care practitioners has grown by 38% to nearly 960 000 during this same period, versus a 9.5% decrease in the number of unionized members in the overall economy. Union representation among nurses is particularly strong; 18.7% of registered nurses (RNs) were represented by unions as of 2010 (Hirsch and Macpherson, 2013b) and as **Figure 1** indicates, these rates are even higher in states such as Washington (61%), Hawaii (55%), California (53%), and New York (46%).

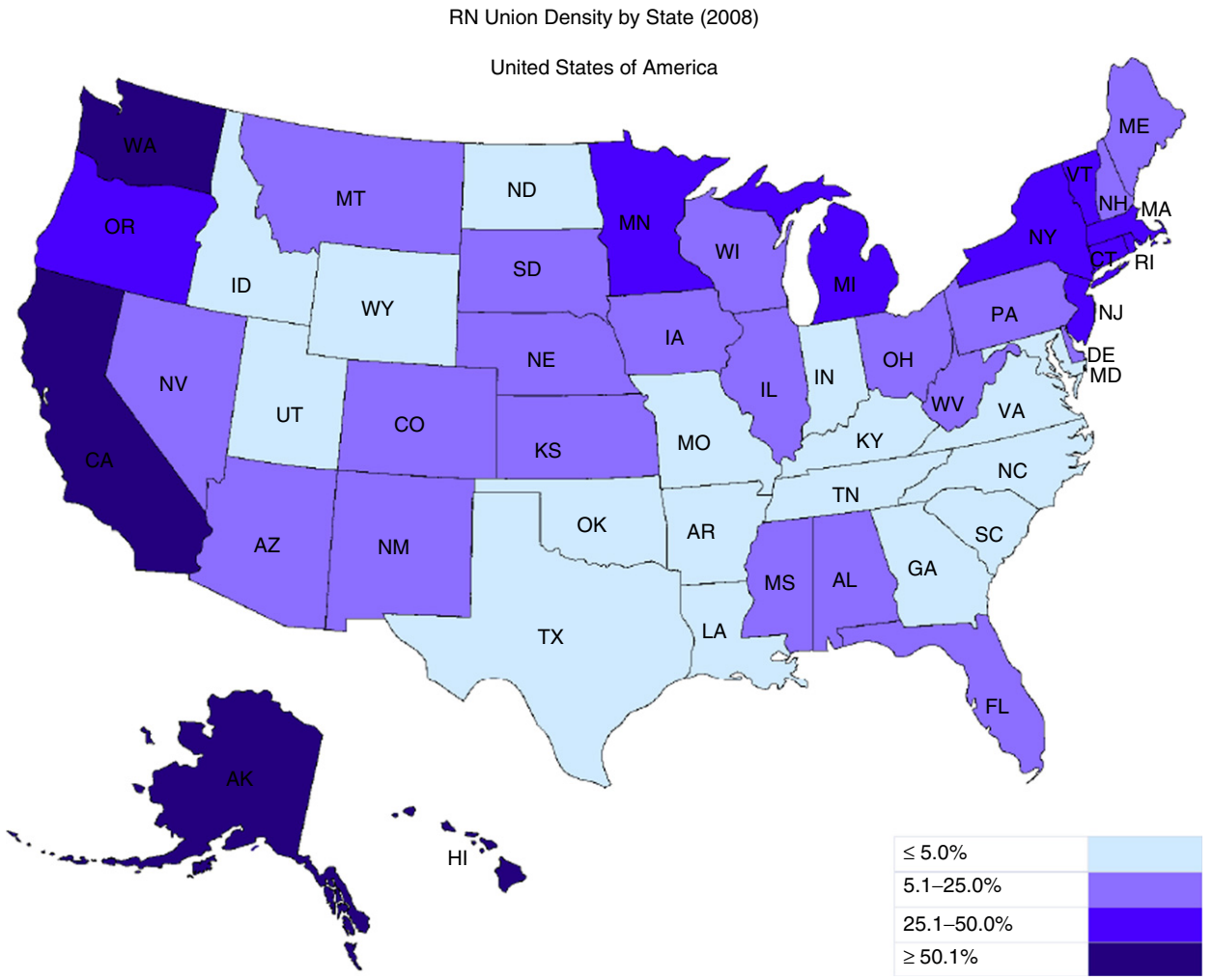


Figure 1 RN union density by state (2008). Calculated using the Current Population Survey Files for 2008. The author is grateful to Rebecca Givan for providing this figure.

Although a complete international comparison of trends in the union density rate for nurses is limited by data availability, the Amsterdam Institute for Advanced Labor Studies gathers data on the overall health care sector union density rate for a small group of countries, which is included in **Figure 2**. For each of the seven listed countries, the health care sector union density rate exceeds that of the overall trade union density rate. However, with the exception of New Zealand, the data show a slight downward trend in the union density rate for health care workers.

Nurses play a significant role in the delivery of health care. They are the most numerous health professionals in most Organization for Economic Cooperation and Development (OECD) countries, with particularly high numbers of nurses per capita in the Nordic countries, Belgium, and Ireland (OECD, 2011a). In the USA, the nursing profession comprises the largest group of health care employees, holding 2.6 million jobs in 2008. Employment in the nursing profession has grown substantially in the past three decades, and if current trends continue, nurse employment is expected to surpass 3 million by 2014 (Health Resources and Services Administration, 2010). In addition, health care production is particularly labor intensive; 60% of hospital costs are labor related (American Hospital Association, 2010) and nursing services constitute the single largest item in most hospital budgets (Public Policy Institute of California, 1996). Consequently, the presence of labor market institutions such as nurses' unions have the potential to substantially affect both the provision and cost of health care, an industry that accounted for an average of 9.6% of Gross Domestic Product in OECD nations (OECD, 2011b), and greater than 17% of spending in the US economy in 2009 (Martin *et al.*, 2011).

Central to the understanding of the effects of nurses' unions in the health care sector is the understanding of the two distinct 'faces' of unionism. The first, as characterized in Freeman and Medoff (1984), is to exercise market power when bargaining with employers in order to obtain more favorable working conditions for their members, including higher wages, and improved conditions of employment. The other, not necessarily incompatible role of unions is to facilitate a 'collective voice' that enables workers to channel their discontent into improved workplace conditions and productivity-enhancing industrial relations policies. Because these forces operate simultaneously within a unionized firm, the fundamental question for understanding the overall impact of unionism relates to the relative magnitude of each of these effects. Consequently, much of the economics literature on unions has focused on an empirical assessment of these two sides of unionism. Several research studies have established that unionized sector workers earn more and have better benefits than their nonunion counterparts (Mellow, 1979; Lewis, 1986; Freeman and Kleiner, 1990; Jakubson, 1991; Wunnava and Ewing, 1999; Hirsch and Schumacher, 2001), whereas evidence on the union productivity effect is less definitive (Fuchs *et al.*, 1998; Doucouliagos

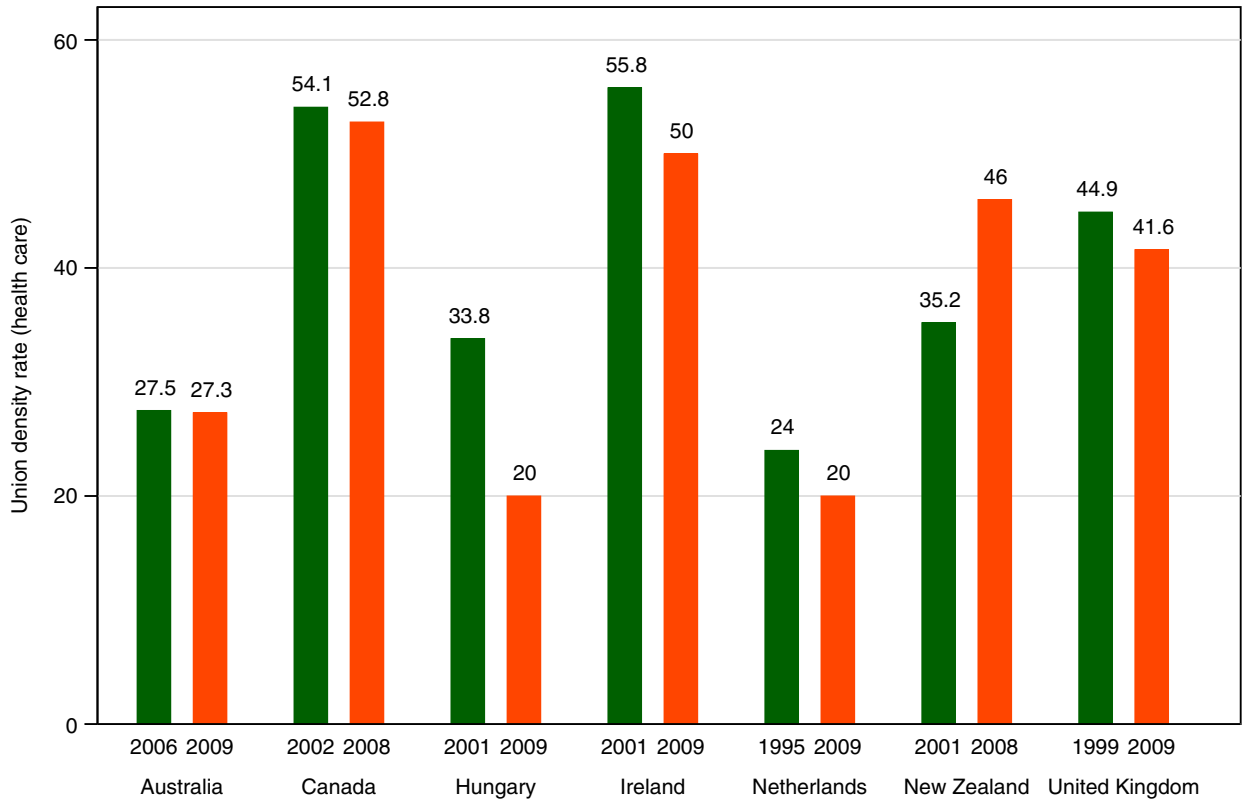


Figure 2 Health care union density rates for selected countries. Reproduced from Amsterdam Institute for Advanced Labor Studis, ICTWSS Database. Available at: http://www.uva-aias.net/208 (accessed 21.09.12). (Note: Overall trade union density rates for the listed countries in 2010 were as follows: Australia (18.0%), Canada (29.5%), Hungary (16.8% in 2008), Ireland (35.0%), the Netherlands (18.6%), New Zealand (20.8%), and the UK (26.5%). See http://stats.oecd.org/Index.aspxDatasetCode=UN_DEN (accessed 24.09.12).)

and Laroche, 2003; Hirsch, 2007, 2008). Furthermore, as Hirsch (2007) indicates, the importance of each side of unionism is very much dependent on the legal and economic environment in which unions and firms operate, as well as the skill level of the employees who are organized (Card, 1996).

This article critically reviews the literature on nurses' unions. Because most studies of nurses' unions rely on data from the USA, in what follows the focus is primarily on the role of these unions in the US economy (unless otherwise noted). The article proceeds in four steps. First, the section "Organized Labor in the US Health Care Industry" provides a brief overview of the legal and regulatory issues surrounding union organizing in the health care industry. The second section "Nurses' Unions and the Labor Market for Nurses," summarizes the evidence on the effects of nurses' unions on the labor market. In the section titled "Nurses' Unions and Firm Performance" I review evidence on the productivity effects of nurses' unions on the firms in which they are employed. I conclude by focusing on priorities for future research.

Organized Labor in the US Health Care Industry

The growth of organized labor in the US health care industry is a relatively recent phenomenon when compared with that of other traditionally unionized sectors of the economy. Although initially covered under the prounion Wagner Act of 1935, collective bargaining in health care institutions was limited by the National Labor Relations Act (NLRA) of 1947. This Act, which outlined unfair practices on the part of unions, also excluded both government and nonprofit hospitals from the right to unionize, asserting that unionization could disrupt the provision of necessary charitable services and open the way for "strikes, picketing and violence which could impede the delivery of health care" (Zacur, 1983, p. 10). Clark *et al.* (2002) note that although eight states enacted legislation granting some collective bargaining rights during this time, most employees in the sector did not have a right to unionization.

After intense lobbying efforts by hospital employee organizations, in 1974, President Nixon signed Public Law 93-360 (PL 93-360) reversing the 23-year exclusion. This law subjected all nongovernmental health care facilities to federal labor law, as governed by the NLRA, and in particular nurses, as 66% of nurses were employed by hospitals at the time (Aiken *et al.*, 1981). In the 4 years following this amendment, there were over 1000 hospital union certification elections (Scott and Simpson, 1989), and an increase in the percentage of hospitals with collective bargaining agreements to 23% in 1976 versus 3% in the early 1960s (Huszczo and Fried, 1988).

In the 15 years following the passage of PL 93-360, a protracted case-by-case bargaining unit determination process enabled hospitals to strategically challenge the bargaining unit determination in each union election, in order to reduce the resolve of employees voting to unionize (Keefe and Rakich, 2004). However, in a final rule published by the National Labor Relations Board in 1989 and affirmed by the US Supreme Court in 1991, eight separate bargaining units were clearly delineated, one of which included registered nurses. Tomey (2004) and Keefe and Rakich (2004) show that this ruling further opened the health care industry to unionization, and increased the number of union elections won in the years following this ruling.

Although these recent protections extended to health care workers have increased the ease with which workers are able to organize, despite these protections, the collective bargaining rights of nurses continue to be disputed. For example, recent cases have disputed the rights of nurses to organize on the grounds that they represent supervisory personnel and are thus exempt from the protections afforded by the NLRA (see *NLBR v. Kentucky River Community Care, Inc., 532 US 706, 2001* and *Oakwood Healthcare, Inc., 348 NLRB 686, 180 LRRM 1257, 2006*). Nonetheless, union representation among nurses remains strong, and is growing.

Nurses' Unions and the Labor Market for Nurses

One of the most important potential impacts of nurses' unions is their effect on the labor market for nurses. Unions have been shown to significantly raise wages in the overall economy (Fuchs *et al.*, 1998; Hirsch and Schumacher, 2004), and registered nurses have realized significant gains in wages in the three decades following the passage of PL 93-360. According to the 2010 National Sample Survey of Registered Nurses, real wages (in 1980 dollars) for registered nurses increased by 54% from 1980 to 2008, from US$17 400 in 1980 to nearly US$27 000 in 2008 (average nominal wages for RNs in 2008 were nearly US$67 000). Furthermore, nurses' unions have recently claimed credit for securing wage increases for their employees, as well as the implementation of employment-related regulation such as minimum nurse-to-patient ratios in California, and the introduction of federal minimum staffing legislation in Congress (California Nurses Association, 2009).

Effects on Employee Compensation

Despite the growing numbers of unionized nurses in the USA, most studies analyzing the effects of nurses' unions on wages are somewhat outdated. Before the passage of PL 93-360, evidence of only a small and sometimes statistically insignificant union wage effect was found in studies relying mostly on data from aggregate metropolitan areas or state-level data for their estimates (Feldman and Scheffler, 1982). Link and Landon's (1975) analysis of the interaction of union and monopsony effects was a notable exception during this period, in that it used a hand-collected data set from individual hospitals. Their results show a 5–10% gain in the wages for hospital nurses due to unionization, with particularly strong gains for lower skilled nurses.

Following the change in the NLRA in 1974, a number of studies attempted to quantify the effects of unions in hospitals, and most focused at least in part on the effects of union membership on nurses. Feldman and Scheffler (1982) use a national probability sample of hospitals drawn from the American Hospital Association (AHA) survey, and present results indicating that both hospitals and unions have market

power. Their findings indicate an overall effect of unions on nurses' wages of approximately 8%, with more substantial wage gains for unions established for at least 10 years at the time that their sample was collected. Adamache and Sloan (1982) in their study of a sample of hospitals in 1979 find a smaller union wage premium of 5% for hospital nurses, although they do not reject a union wage premium of up to 20%. Their study is also unique in that their estimates imply substantial union spillover effects, equating to a 10% wage premium at surrounding hospitals if 75% of hospitals in a market have formal collective bargaining agreements. Furthermore, their analysis of the effects of market structure on nurse wages finds no evidence that unions possess countervailing power that offsets the effects of monopsony. Groshen and Krueger (1990) also find evidence of a wage premium for nurses of approximately 4%, although their results suffer from similar limitations of earlier studies on union wage premiums in that their unit of analysis is a metropolitan area rather than a hospital.

Two more recent studies by Schumacher and Hirsch (1997) and Hirsch and Schumacher (1998) adopt a distinctly different approach to identifying the union wage premium for nurses, arguing that measurement of the union wage premium using firm-level data, as has been attempted in previous studies, may distort the true union wage premium. Specifically, they emphasize the potential for overstating the true union wage premium if such a wage differential corresponds to unmeasured skill differentials across unionized and nonunionized workers. Hirsch and Schumacher (1998) demonstrate that although cross-section regression estimates of the union wage premium in their sample produce a statistically significant union wage premium estimate roughly equivalent to those found in previous studies (approximately 3.2%), a specification that estimates the union wage premium using the change in wages for individuals who switch into or out of union membership over a 1-year period suggests a statistically insignificant estimate of 1.1%. This implies that the cross-section union wage differential may represent a compensating differential for unmeasured worker ability. Schumacher and Hirsch (1997) further this point in their work analyzing the magnitude of hospital wage premiums, finding a substantial wage premium realized by hospital nurses, with little of this premium due to union membership. Though the authors acknowledge that their identification strategy is likely to bias the results toward an underestimate of the true union wage effect, their findings parallel those of Bruggink *et al.* (1985) who find no direct impact of nurse unions on RN wages.

In the only analysis of the effects of unions on the distribution of wages, Spetz *et al.* (2010) find little difference in the pay structure of unionized versus nonunionized nurses. Though a number of their findings fail to achieve statistical significance, notable in their study are results indicating that nurses' unions decrease the disparities in income between nurses with dissimilar levels of education, and also lower the return to experience for unionized nurses. They conclude that, contrary to Freeman (1980, 1982), who found strong evidence that unions reduce wage dispersion and rationalize the wage structure, unions' primary effect in hospitals is to raise wages with no noticeable effects on the wage distribution.

Explanations for the Evidence on Union Wage Premiums for Nurses

Although the estimates of the union wage premium for nurses is substantially lower than the 15% average effect that persists in the overall economy, a number of factors likely contribute to the underestimation of the true premium. First, as Kaufman (2007) notes, the effects of unions may vary when wages and employment are not determined within a competitive labor market. Reports of nursing shortages in the USA in every decade since the 1950s, coupled with the concentration of nursing employment in relatively few settings, have led economists to characterize the nursing market as the "textbook example of monopsony" (Hirsch and Schumacher, 2004, p. 1). **Figure 3** indicates that approximately 60–70% of nurses are employed by hospitals, and more than 85% are employed by either hospitals, ambulatory care, or community health centers.

The existence of monopsony power can generate union wage effects that differ from those predicted by the standard competitive model where the market power of unions over labor supply is assumed to enable them to bargain for higher wages. Specifically, a union's market power in a monopsonistic labor market may be partially offset by an employer's dominance in the labor market, resulting in a wage level closer to that which would prevail in a competitive, nonunionized market. Thus, depending on the relative bargaining position of the firm versus the union within a market, collective bargaining in a monopsonistic labor market could lead to smaller wage premiums than would prevail in a competitive labor market, and studies estimating the union wage premium that omit variables that are correlated with hospital monopsony power could tend to underestimate the true union wage effect. Although empirical support for monopsony is mixed (Adamache and Sloan, 1982; Hirsch and Schumacher, 1995, 2005; Matsudaira, 2010), a sufficient number of studies have found evidence consistent with the presence of monopsony power for firms that employ nurses, which lends credibility to this argument (Hurd, 1973; Link and Landon, 1975; Bruggink *et al.*, 1985; Robinson, 1988; Sullivan, 1989; Currie *et al.*, 2005; Staiger *et al.*, 2010).

Second, both Sloan and Steinwald (1980) and Salkever (1984) find that union effects vary depending on the number of years a union has existed, with greater wage effects found for older, more established unions. Given the relatively new status of nurses' unions during the period when many of these studies were conducted, the full impact of bargaining may not yet be fully established. Third, as Nicholson (2003) notes, economists such as Pencavel (1984) and MaCurdy and Pencavel (1986) have modeled unions as utility-maximizing entities that negotiate with firms over both worker rents as well as the quantity of union members employed. Although no studies have explicitly focused on the employment effects of nurses' unions, the inclusion of minimum staffing language in most union contracts (Clark and Clark, 2006), as well as recent efforts by unions to mandate staffing levels at hospitals and nursing homes, is consistent with these models. Finally, Adamache and Sloan (1982), and Hirsch and Schumacher (1998) find evidence of a substantial union threat effect, wherein nonunion employers may offer higher wages to their workers to reduce the threat of unionization. As Ichniowski

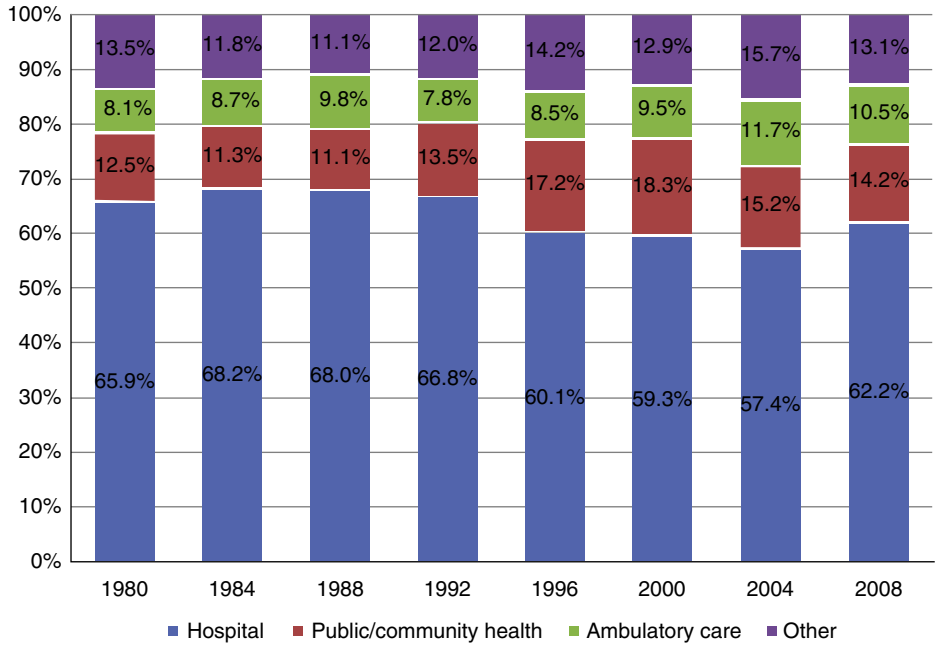


Figure 3 Employment settings of registered nurses, 1980–2008. Data from Health Resources and Services Administration (2010). *The registered nurse population findings from the 2008 National Sample Survey of Registered Nurses*, U.S. Department of Health and Human Services. Available at: http://bhpr.hrsa.gov/healthworkforce/rnsurveys/rnsurveyfinal.pdf (accessed 27.06.13).

et al. (1989) indicate, failure to adequately control for threat effects may underestimate the full effect of unions on employee compensation.

Nurses' Unions and Firm Performance

What unions do to productivity is one of the key factors in assessing the overall economic impact of unions. Given the importance of nurses for the firms in which they are employed, nurses' unions have the potential to substantially affect firm performance. Health care production is particularly labor intensive, with nurses accounting for 30% of hospital costs (McCue *et al.*, 2003). Nurses are a crucial part of the hospital production function and are, as one hospital Chief Executive Officer said, "the heart and soul of the hospital" (Draper *et al.*, 2008, p. 2). The nursing staff in a hospital is by far the most productive labor input, with a marginal product nearly three times as large as that of any other hospital input (Jensen and Morrisey, 1986).

Hirsch (2007) identifies three routes through which unionization-induced productivity gains can be realized: (1) union-induced wage increases may induce increases in technical efficiency; (2) reductions in turnover costs; and (3) productivity-enhancing personnel policies resulting from increased employee involvement in the production process. Although wage increases have been found to induce factor substitution and reduce quality for nursing homes (Cawley *et al.*, 2006), a number of studies document the beneficial effects of turnover reduction and increased employee involvement in hospital settings in both the USA and the UK (Propper and Van Reenen, 2010; Cebul *et al.*, 2008; Phibbs *et al.*, 2009).

Nurses' Unions and Hospital Output

Older studies analyzing the long-term performance effects of nurses' unions in the hospital industry have generally concluded that unions adversely affect health care production, and that the effects are especially relevant for hospitals with an established union presence. Sloan and Steinwald's (1980) estimates indicate that average hospital costs increase 2–3% in the year collective bargaining is introduced, with larger effects of 4–6% for more established unions. Salkever (1982, 1984) also finds evidence of larger union effects for established unions. His analysis suggests that union impacts on total costs are not positive during the first 2 years of unionization, but that a union presence increases hospital costs by 3.3–9% overall. Though not focused on nurses specifically, Sloan and Adamache's (1984) analysis of a national sample of AHA member hospitals shows an increase in hospital costs per adjusted patient day and adjusted admission of 3.5% and 4.1%, respectively, at unionized hospitals in which there were no recent incidents of labor strife. Their results are larger for unionized hospitals with a recent strike or job action; costs per adjusted admission in these hospitals are 9–10% higher than at nonunionized hospitals.

Of note in this literature are the suggested mechanisms through which these cost increases occur. Sloan and Adamache (1984) attribute their findings entirely to the cost impacts of the union wage effect, implying no union productivity effect. Salkever (1984) and Groshen and Krueger (1990), however, attribute their findings primarily to nonwage factors. Salkever (1984)

estimates that nonwage components account for two-thirds of the cost increase due to unionization, whereas Groshen and Krueger (1990) indicate that unionized hospitals are more limited in their ability to adjust wages and staffing levels than are nonunionized hospitals. Furthermore, both of Salkever's studies attribute hospital cost increases to employee groups other than nurses; when nurse unions are separately analyzed, he finds negative and insignificant effects of RN unions on costs.

Register's (1988) analysis of the union effect on hospital production was the first to examine the productivity effects of unions using data collected after the introduction of Medicare's Prospective Payment System (PPS). Although hospital reimbursement was cost-based before the introduction of PPS in 1983, under PPS, hospitals were reimbursed a fixed amount for each diagnosis-related group regardless of the actual expenses incurred in caring for a patient. Because this created new incentives for efficient operation, it became more important to realize beneficial union productivity effects after the introduction of PPS. His study finds evidence of productivity effects using both a national sample and state-specific data, and indicates that average costs at unionized hospitals are approximately 9% lower than at nonunionized hospitals.

Effects on Quality

Although the aforementioned studies base their conclusions on measures of hospital output such as total discharges and patient days, none of these studies are able to account for the quality of production. Quality is a major concern in health care, and is often thought of as a more important dimension of production than in other industries (Gaynor, 2006). Furthermore, quality is costly; Romley and Goldman (2008) find that an interquartile improvement in hospital quality would increase hospital costs by nearly 50%. Thus, an analysis of the effects of nurses' unions on hospital production that fails to account for the role of nurses' unions on the quality of hospital services could lead to erroneous conclusions regarding the full extent to which unions affect productivity.

Seago and Ash (2002) and Ash and Seago (2004) are the only studies to directly address this concern in their study of 344 California hospitals from 1993. Their studies analyze the effects of nurses' unions on heart attack mortality, using risk-adjusted heart attack mortality rates collected as part of a California Hospital Outcomes Project. Their findings indicate that risk-adjusted heart attack mortality rates for hospitals with unionized nurses are 5.5% lower than at nonunionized hospitals. Though their identification strategy cannot rule out the potential for nonrandom selection of unions into high-quality hospitals, they argue the potential for the selection of unions into low-quality hospitals is likely as well, given the potential for poor employee morale in these facilities. Consequently, though their study addresses the issue of patient selection in a particularly thorough manner given the constraints of their data, their discussion stops just short of arguing for a purely causal interpretation of their results.

Impact of the Labor Relations Environment

The state of the labor relations environment can also greatly impact productivity within a unionized firm. A number of multi-industry studies provide evidence that productivity can deteriorate as a result of strikes and labor unrest (Neumann, 1980; Neumann and Reder, 1984; Becker and Olson, 1986; Kramer and Vasconcellos, 1996; Kleiner *et al.*, 2002). Strikes and poor labor–management relations have also been shown to negatively impact the quality of production. For example, Krueger and Mas (2004) found that tire defect rates were particularly high at a tire plant during periods of labor unrest, and Mas (2008) found that workmanship for construction equipment produced at factories that experienced contract disputes was significantly worse relative to equipment produced at factories without labor unrest, as measured by the resale value of the equipment.

Given the importance of nurses to health care production, coupled with the complex nature of health care delivery where workers exhibit a high degree of interdependence (Cebul *et al.*, 2008), it is perhaps not surprising that health care quality has been shown to be particularly susceptible to labor unrest. Mustard *et al.* (1995) found that the incidence of adverse newborn outcomes increased during a month-long Ontario nurses strike, conjecturing that disruption in the normal standards of care was a contributing factor to the elevated rate of adverse outcomes. Gruber and Kleiner (2012) also find evidence of deterioration in patient outcomes in their study of 50 hospital strikes in New York State. Their results indicate that nurses' strikes increase in-hospital mortality by 18.3% and 30-day readmission by 5.7% for patients admitted during a strike. Furthermore, their results highlight the importance of employee tenure within a firm, as they find that hospitals staffed by replacement workers during strikes perform no better during these strikes than those that do not hire substitute employees.

Conclusion and Areas for Future Research

Although a considerable literature has investigated the effects of nurses' unions in the health care sector, the conclusions of this literature are far from definitive. Although the current body of research could be characterized as suggesting that nurses' unions raise wages and contribute to reductions in firm performance, sufficient evidence exists within this literature to challenge these conclusions. Further research could contribute to our understanding of the role of nurses' unions in five ways. First, the conclusions of the existing literature are based largely on old data that do not reflect the current state of the health care industry. For example, of the studies examining the production effects of health care worker unions, only Register (1988) uses data from a period following the implementation of PPS, and none of these studies examine the impact of unions after the growth of managed care and subsequent restructuring of the health care system. As Norrish and Rundall (2001) indicate, this restructuring affected aspects of nursing such as staffing ratios, and workload, both of which are likely to affect the role of nurses within the institutions for which they work. Thus, our understanding of the effects of nurses' unions would greatly benefit from research utilizing more recent data. Second, with the exception of Hirsch and Schumacher's research on union wage premiums, the aforementioned studies rely on cross-sectional variation to identify

their results. Given the potential for selection of unions into firms with higher wages and poor labor relations (which may contribute to reduced productivity), future work should employ updated econometric techniques to better account for this possibility. Third, more work on the impact of nurses' unions on the quality of hospital production is needed. With the exception of Ash and Seago's demonstration of a positive relationship between union status and patient outcomes, and Gruber and Kleiner (2012), who find a short-run decrease in quality due to strikes, little attempt has been made to quantify these effects, despite the claims of nurses' unions that patient care is a priority in negotiations (Clark and Clark, 2006). Fourth, although nurses' unions are similar in their objectives across countries (Clark and Clark, 2003) and union wage and productivity effects have been documented in a number of developed nations (Blanchflower and Freeman, 1992), little is known about the wage and productivity effects of nurses' unions in countries other than the USA. Finally, the mechanism by which nurses' unions affect hospital performance is not well understood. A better understanding of the means by which these unions affect productivity would assist in managerial and public policy decision making.

See also: Market for Professional Nurses in the US. Monopsony in Health Labor Markets

References

Adamache, K. W. and Sloan, F. A. (1982). Unions and hospitals: Some unresolved issues. *Journal of Health Economics* **1**(1), 81–108.

Aiken, L. H., Blendon, R. J. and Rogers, D. E. (1981). The shortage of hospital nurses: A new perspective. *Annals of Internal Medicine* **95**, 365–371.

American Hospital Association (2010). Trendwatch chartbook 2010: The economic contribution of hospitals, [Online]. Available at: http://www.aha.org/research/reports/tw/chart- book/2010/chapter6.pdf (accessed 27.06.13).

Ash, M. and Seago, J. A. (2004). The effect of Registered Nurses' Unions of heart-attack mortality. *Industrial and Labor Relations Review* **57**, 422–442.

Becker, B. E. and Olson, C. A. (1986). The impact of strikes on shareholder equity. *Industrial and Labor Relations Review* **39**(3), 425–438.

Blanchflower, D. G. and Freeman, R. B. (1992). Unionism in the United States and other advanced OECD countries. *Industrial Relations* **31**(Winter), 56–79.

Bruggink, T. H., Finan, K. C., Gendel, E. B. and Todd, J. S. (1985). Direct and indirect effects of unionization on the wage levels of nurses: A case study of New Jersey hospitals. *Journal of Labor Research* **6**, 407–416.

Bureau of Labor Statistics (2011). Union affiliation of employed wage and salary workers by occupation and industry, [Online]. Available at: http://www.bls.gov/webapps/legacy/cpslutab3.htm (accessed 24.04.11).

California Nurses Association (CNA) (2009). *The ratio solution: CNA/NNOC's RN-to-patient ratios work – Better care, more nurses.* Oakland, CA: California Nurses Association. Available at: http://www.area-c54.it/public/california%20nurses%20association%202.pdf (accessed 24.05.11).

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica* **64**, 957–979.

Cawley, J., Grabowski, D. and Hirth, R. (2006). Factor substitution in nursing homes. *Journal of Health Economics* **25**(2), 234–247.

Cebul, R., Rebitzer, J.B., Taylor, L.J. and Votruba, M. (2008). *Organizational fragmentation and care quality in the U.S. Health care system.* National Bureau of Economic Research, Working Paper 14212.

Clark, P. F. and Clark, D. A. (2003). Challenges facing Nurses' Associations and Unions: A global perspective. *International Labour Review* **142**, 29–48.

Clark, P. F. and Clark, D. A. (2006). Union strategies for improving patient care: The key to nurse unionism. *Labor Studies Journal* **31**(1), 51–70.

Clark, P. F., Delaney, J. T. and Frost, A. C. (eds.) (2002). *Collective bargaining in the private sector.* Ithaca, NY: Cornell University Press.

Currie, J., Mehdi, F. and MacLeod, W. B. (2005). Cut to the bone? Hospital takeovers and nurse employment contracts. *Industrial & Labor Relations Review* **58**(3), 471–493.

Daly, J., Speedy, S. and Jackson, D. (2004). *Nursing leadership.* Marrickville, NSW, Australia: Elsevier.

Danish Nurses' Organization (2009). Nurse in Denmark? A guide on salary, pension and employment. Available at: http://www.dsr.dk/Artikler/Documents/English/Nurse%20in%20Denmark.pdf (accessed 05.06.12).

Doucouliagos, C. and Laroche, P. (2003). What do unions do to productivity? A metaanalysis. *Industrial Relations* **42**, 650–691.

Draper, D. A., Felland, L. E., Liebhaber, A. and Melichar, L. (2008). The role of nurses in hospital quality improvement research brief, [Online]. Available at: http://www.hschange.com/CONTENT/972/972.pdf (Center for Studying Health System Change) (accessed 09.03.10).

Feldman, R. and Scheffler, R. (1982). The union impact on hospital wages and fringe benefits. *Industrial and Labor Relations Review* **35**, 196–206.

Freeman, R. B. (1980). Unionism and the dispersion of wages. *Industrial and Labor Relations Review* **34**, 3–23.

Freeman, R. B. (1982). Union wage practices and wage dispersion within establishments. *Industrial and Labor Relations Review* **36**, 3–21.

Freeman, R. B. and Kleiner, M. M. (1990). The impact of new unionization on wages and working conditions. *Journal of Labor Economics* **8**(1, part 2), S8–S25.

Freeman, R. B. and Medoff, J. L. (1984). *What do unions do?* New York: Basic Books.

Fuchs, V. R., Krueger, A. B. and Poterba, J. M. (1998). Economists' view about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature* **36**, 1387–1425.

Gaynor, M. (2006). *What do we know about competition and quality in health care markets?* NBER Working Paper 12301.

Groshen, E. L. and Krueger, A. B. (1990). The structure of supervision and pay in hospitals. *Industrial and Labor Relations Review* **43**(February), 134–146.

Gruber, J. and Kleiner, S. A. (2012). Do strikes kill? Evidence from New York state. *American Economic Journal: Economic Policy* **4**(1), 127–157.

Health Resources and Services Administration (2010). *The registered nurse population findings from the 2008 National Sample Survey of Registered Nurses.* U.S. Department of Health and Human Services. Available at: http://bhpr.hrsa.gov/healthworkforce/rnsurveys/rnsurveyfinal.pdf

Hirsch, B. T. (2007). What do unions do for economic performance? In Bennett, J. T. and Kaufman, B. E. (eds.) *What do unions do? A twenty year perspective*, pp. 193–237. Piscataway, NJ: Transaction Publishers.

Hirsch, B. T. (2008). Sluggish institutions in a dynamic world: Can unions and industrial competition coexist? *Journal of Economic Perspectives* **22**, 153–176.

Hirsch, B. T. and Macpherson, D. A. (2013a). Union membership, coverage, density, and employment among all wage and salary workers, 1973–2012, [Online]. Available at: Unionstats.com (accessed 31.07.13).

Hirsch, B. T. and Macpherson, D. A. (2013b). Union membership, coverage, density, and employment by occupation, 2010, [Online]. Available at: Unionstats.com (accessed 31.07.13).

Hirsch, B. T. and Schumacher, E. J. (1995). Monopsony power and relative wages in the labor market for nurses. *Journal of Health Economics* **14**(4), 443–476.

Hirsch, B. T. and Schumacher, E. J. (1998). Union wages, rents, and skills in health care labor markets. *Journal of Labor Research* **19**(1), 125–147.

Hirsch, B. T. and Schumacher, E. J. (2001). Private sector union density and the wage premium: Past, present, and future. *Journal of Labor Research* **22**, 487–518.

Hirsch, B. T. and Schumacher, E. J. (2004). Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics* **22**, 689–722.

Hirsch, B. T. and Schumacher, E. J. (2005). Classic or new monopsony? Searching for evidence in nursing labor markets. *Journal of Health Economics* **24**(5), 969–989.

Hurd, R. W. (1973). Equilibrium vacancies in a labor market dominated by non-profit firms: The "shortage" of nurses. *Review of Economics and Statistics* **55**, 234–240.

Huszczo, G. E. and Fried, B. J. (1988). A labor relations research agenda for health care settings. *Employee Responsibilities and Rights Journal* **1**(1), 69–84.

Ichniowski, C., Freeman, R. and Lauer, H. (1989). Collective bargaining laws, threat effects, and the determination of police compensation. *Journal of Labor Economics* **VII**, 191–209.

Informetrica Limited (2011). Quick Facts prepared for the Canadian Federation of Nurses Unions. Available at: http://www.nursesunions.ca/sites/default/files/overtime_and_absenteeism_quick_facts.pdf (accessed 04.06.12).
Jakubson, G. (1991). Estimation and testing of fixed effects models: Estimation of the union wage effect using panel data. *Review of Economic Studies* **58**, 971–991.
Jensen, G. A. and Morrisey, M. A. (1986). The role of physicians in hospital production. *The Review of Economics and Statistics* **68**(3), 432–442.
Kaufman, B. E. (2007). What do unions do? Insights from economic theory. In Bennett, J. T. and Kaufman, B. E. (eds.) *In what do unions do? A twenty-year perspective*, pp. 12–45. New Brunswick: Transaction Publishers.
Keefe, T. and Rakich, J. S. (2004). A profile of hospital union election activity, 1985–1994: NLRB rulemaking and results in right-to-work states. *Hospital Topics* **82**(2), 2–11.
Kleiner, M. M., Leonard, J. S. and Pilarski, A. M. (2002). How industrial relations affect plant performance: The case of commercial aircraft manufacturing. *Industrial and Labor Relations Review* **55**(2), 195–218.
Kramer, J. K. and Vasconcellos, G. M. (1996). The economic effect of strikes on the shareholders of nonstruck competitors. *Industrial and Labor Relations Review* **49**(2), 213–222.
Krueger, A. B. and Mas, A. (2004). Strikes, scabs, and tread separations: Labor strife and the production of defective bridgestone/firestone tires. *Journal of Political Economy* **112**(2), 253–289.
Lewis, H. G. (1986). *Union relative wage effects: A survey*. Chicago, IL: University of Chicago Press.
Link, C. R. and Landon, J. H. (1975). Monopsony and union power in the market for nurses. *Southern Economic Journal* **41**, 649–656.
MaCurdy, T. E. and Pencavel, J. H. (1986). Testing between competing models of wage and employment determination in unionized markets. *Journal of Political Economy* **94**(3), S3–S39.
Martin, A., Lassman, D., Whittle, L. and Catlin, A. (2011). National health expenditure accounts team. Recession contributes to slowest annual rate of increase in health spending in five decades. *Health Affairs (Millwood)* **30**(1), 11–22.
Mas, A. (2008). Labour unrest and the quality of production: Evidence from the construction equipment resale market. *Review of Economic Studies* **75**(1), 229–258.
Matsudaira, J. D. (2010). *Monopsony in the low-wage labor market? Evidence from minimum nurse staffing regulations*. unpublished manuscript, Cornell University.
McCue, M., Mark., B. A. and Harless, D. W. (2003). Nurse staffing, quality, and financial performance. *Journal of Health Care Finance* **29**, 54–76.
Mellow, W. S. (1979). *Unionism and wages: A longitudinal analysis*. Washington, DC: Bureau of Labor Statistics, U.S. Department of Labor.
Metcalf, D. (2005). British Unions: Resurgence or Perdition? The Work Foundation, Provocation Series, vol. 1, no. 1. Available at: http://www.theworkfoundation.com/DownloadPublication/Report/68_68_British%20Unions.pdf (accessed 04.06.12).
Mustard, C., Harmon, C., Hall, P. and Dirksen, S. (1995). Impact of a nurses' strike on the cesarean birth rate. *American Journal of Obstetrics and Gynecology* **172**(2), 631–637.
Neumann, G. R. (1980). The predictability of strikes: Evidence from the stock market. *Industrial and Labor Relations Review* **33**(4), 525–535.
Neumann, G. R. and Reder, M. W. (1984). Output and strike activity in U.S. manufacturing: How large are the losses? *Industrial and Labor Relations Review* **37**(2), 197–211.
Nicholson, S. (2003). *Barriers to entering medical specialties*. Working Paper 9649. Cambridge MA: NBER.
NLRB (2004). *Sixty-Eighth Annual Report of the National Labor Relations Board for the Fiscal Year Ended 30 September 2003 at Table 16*. Available at: http://www.nlrb.gov/sites/default/files/documents/119/nlrb2003.pdf (accessed 27.06.13).
Norrish, B. R. and Rundall, T. G. (2001). Hospital restructuring and the work of registered nurses. *Milbank Quarterly* **79**, 55–79, IV.
OECD (2011a). *Health at a glance 2011: OECD indicators*. OECD Publishing. Available at: http://dx.doi.org/10.1787/health_glance-2011-en
OECD (2011b). *"Nurses," in OECD factbook 2011–2012: Economic, environmental and social statistics*. OECD Publishing. Available at: http://dx.doi.org/10.1787/factbook-2011-111-en
OMalley, B. H. (2012). The labor union and the registered nursing profession, [Online]. Available at: http://www.registered-nurse-canada.com/labor_union.html (accessed 31.07.13).
Pencavel, J. (1984). The tradeoff between wages and employment in trade union objectives. *Quarterly Journal of Economics* **99**(2), 215–231.
Phibbs, C. S., Bartel, A. P., Giovannetti, B., Schmitt, S. K., Stone, P. W.(2009). The type of nurse matters: The effects of nurse staffing levels, non-RNs, and contract nurses on length of stay, [Online]. Available at: http://www0.gsb.columbia.edu/faculty/abartel/Type%20of%20Nurse%20Matters%20-NEJM.pdf (accessed 09.03.10).
Propper, C. and Van Reenen, J. (2010). Can pay regulation kill? Panel data evidence on the effect of labor markets on hospital performance. *Journal of Political Economy* **118**(2), 222–273.
Public Policy Institute of California (1996). Research brief: Are hospitals reducing nursing staff levels?, [Online]. Available at: http://www.ppic.org/con-tent/pubs/rb/RB_1096JSRB.pdf (accessed 27.04.11).
Register, C. A. (1988). Wages, productivity, and costs in union and nonunion hospitals. *Journal of Labor Research* **9**(4), 325–345.
Robinson, J. C. (1988). Market structure, employment, and skill mix in the hospital industry. *Southern Economic Journal* **55**, 315–325.
Romley, J. A. and Goldman, D. (2008). *How costly is hospital quality? A revealed-preference approach*. Working Paper 13730. Cambridge, MA: National Bureau of Economic Research.
Salkever, D. S. (1982). Unionization and the cost of producing hospital services. *Journal of Labor Research* **3**(3), 311–333.
Salkever, D. S. (1984). Cost implications of hospital unionization: A behavioral analysis. *Health Services Research* **19**(5), 639–664.
Schumacher, E. J. and Hirsch, B. T. (1997). Compensating differentials and unmeasured ability in the labor market for nurses: Why do hospitals pay more. *Industrial & Labor Relations Review* **50**, 557–579.
Scott, C. and Simpson, J. (1989). Union election activity in the hospital industry. *Health Care Management Review* **14**(4), 21–28.
Seago, J. A. and Michael, A. (2002). Registered nurse unions and patient outcomes. *Journal of Nursing Administration* **32**(3), 143–151.
Sloan, F. A. and Adamache, K. W. (1984). The role of unions in hospital cost inflation. *Industrial and Labor Relations Review* **37**, 252–262.
Sloan, F. A. and Steinwald, B. (1980). *Hospital labor markets: Analysis of wages and workforce compensation*. Lexington, MA: D.C. Heath.
Spetz, J., Ash, M., Konstantinidis, C. and Herrera, C. (2010). The effect of unions on the distribution of wages of hospital-employed registered nurses in the united states. *Journal of Clinical Nursing* **20**, 60–67.
Staiger, D., Spetz, J. and Phibbs, C. (2010). Is there monopsony in the labor market? Evidence from a natural experiment. *Journal of Labor Economics* **28**(2), 211–236.
Sullivan, D. (1989). Monopsony power in the market for nurses. *Journal of Law and Economics* **32**, S135–S178.
Tomey, A. M. (2004). *Guide to nursing management and leadership*. 7th ed. Missouri: Mosby.
Wunnava, P. V. and Ewing, B. T. (1999). Union–nonunion differentials and establishment size: Evidence from the NSLY. *Journal of Labor Research* **20**, 177–183.
Zacur, S. R. (1983). *Health care labor relations: The nursing perspective*. Ann Arbor: UMI Research Press.

Further Reading

Bennett, J. T. and Kaufman, B. E. (eds.) (2007). *What do unions do? A twenty year perspective*. New Brunswick, NJ: Transaction Publishers.
Buerhaus, P. I., Staiger, D. O. and Auerbach, D. I. (2009). *The future of the nursing workforce in the United States*. Sudbury, MA: Jones and Bartlett.

Nutrition, Economics of

M Bitler, University of California Irvine, Irvine, CA, USA
P Wilde, Tufts University, Boston, MA, USA

Introduction

The close relationship between economics and nutrition runs in two directions.

- First, nutrition influences economic conditions. Economic historian Robert Fogel has argued that improved nutrition was a decisive factor for improved health and successful economic development in Europe and the United States during the nineteenth and twentieth centuries (Fogel, 2004). More recently, rising rates of obesity have affected rates of chronic disease in developed countries, raising health-care costs.
- Second, economic conditions influence nutrition. Prices and incomes are leading determinants of food choices, dietary quality, and household food security.

This article addresses selected aspects of the economics of nutrition in developed countries. The economics of nutrition in developing countries are discussed in other chapters. Developed countries (and some less-developed countries) simultaneously face nutrition problems from both overconsumption and underconsumption. Throughout this article, the United States is taken as a case study for the developed world, in part because of the easy availability of data.

Four leading causes of death in the United States are influenced by dietary choices: heart disease, cancer, stroke, and diabetes. Increased rates of overweight and obesity in recent decades have been associated with each of these diseases. This has happened at the same time as rates of being overweight or obese among adults have increased from 50% of men and 40% of women in the early-1960s to 73% of men and 64% of women in the mid- to late-2000s (National Center for Health Statistics (2011)). In the United States, obesity may be responsible for $147 billion per year in medical costs, approximately 10% of all medical expenditures (Finkelstein *et al.*, 2009), with some economists estimating even larger effects.

Meanwhile, the federal government assesses the extent of food-related hardship in the United States using a survey-based measure of food insecurity, which is defined as not being able to afford sufficient food for an active healthy life for all household members at all times. Based on responses to questions in the Current Population Survey, the US Department of Agriculture (USDA) estimates that 17 million households (or 14.7% of all US households) were food insecure in 2009. In 4.6% of US households in 2009, the survey respondent reported experiencing hunger at some point during that year.

This article focuses on economic principles and research that are useful for understanding policy decisions about nutrition issues. Section The Effect of Economic Conditions on Nutrition describes how prices, income, and other factors may affect nutrition. Section A Framework for Nutrition Policy Options summarizes economic as well as commonly applied noneconomic perspectives on policy options to address nutrition concerns. The remaining sections review real-world applications related to four leading categories of policy responses to nutrition concerns: The next four Sections Food Assistance Programs, The Economics of Information Policy, Direct Interventions: Taxes and Subsidies, and Government Supply Interventions discuss food assistance programs; policies to improve nutrition information; taxes and subsidies to guide consumer food choices; and other government interventions to affect supply. The final Section Behavioral Economics: Nudges discusses recent insights from behavioral economics regarding policies to nudge consumers in the direction of healthier choices.

The Effect of Economic Conditions on Nutrition

As theoretical foundations, economists use both (1) a traditional simple theory of consumer choice and also (2) more elaborate theories that more specifically address health and nutrition issues.

The traditional economic theory of choice suggests that rational consumers seek to purchase a bundle of goods (x_1, x_2,..., x_k) that satisfies their preferences (represented by the utility function U) and is feasible given the budget constraint. Because the budget constraint depends on the consumer's income (Y) and the prices of the goods (p_1, p_2,..., p_k), the theory suggests that consumer choices respond in a systematic way to prices and incomes. Formally, the consumer solves the choice problem:

$$\text{Max } U(x_1, x_2,..., x_k),$$
$$\text{subject to } p_1x_1 + p_2x_2 + ... + p_k\, x_k \leq Y$$

For normal goods, if income increases, then the quantity purchased increases. In standard cases, if the price increases, then the quantity purchased decreases. Goods whose consumption amounts tend to rise and fall together (such as peanut butter and jelly) are complements: if the price of a good increases, then the quantity purchased of a complementary good decreases. Other goods (such as beef and pork) are substitutes: if the price of a good increases, then the quantity purchased of a substitute good increases. The responsiveness of these price and income effects is measured using elasticities, which denote the percentage change in consumption of a food in response to a 1% increase in the explanatory variable.

In studying nutrition issues, economists have in various ways extended the simple traditional theory of choice to take account of the health consequences of food choices. Consider a consumer who must choose goods to purchase and the amount of time to spend on an array of activities (z_1, z_2,..., z_k), including daily hours of sleep, work, watching TV, exercise,

grocery shopping, and cooking. Just as there is a budget constraint for total purchases, there is also a time constraint for the total number of hours available in a day. Such extended theories of choice offer insight into nutritionally relevant decisions. For example:

- Weight change from one month to the next depends on the amount of food energy consumed and the amount of energy expended on daily activities;
- Dietary quality depends in part on how the consumer combines purchased food with time spent in activities (as when a consumer chooses between a fast-food meal and a home-cooked meal);
- The time costs of preparing food may influence consumer choices among restaurant meals, convenience food, and home-cooked meals;
- The quality and convenience of healthy food options in the local food retail environment may influence consumer food choices;
- Consumer choices about food purchases and activities may reflect preferences in direct ways (as when one eats the foods one likes or does the activities one enjoys) and also indirect ways (as when one spends time in difficult physical exercise or eats less-preferred foods because of their healthiness). Recent work using household-level data from the US, France, and the UK, by Dubois *et al.* (in press) suggests preferences may play a large role in food choice across countries in addition to the role of prices and nutrient characteristics.

The traditional theory of choice points our attention toward trends in food prices and incomes. With regard to prices, one important pattern in the United States is that the price of food has fallen relative to the price of other goods and services. A second pattern is that prices evolve differently for some food groups than for others. Using 1981 as the base time period, Figure 1 shows in subsequent months how each food group's price increased relative to the average increase in the level of prices for all foods and beverages combined. Thus, by 2008, the price of fruits and vegetables had increased most rapidly (index ≈ 130%), the price of food away from home had increased at the same rate as overall food and beverage average prices (index ≈ 100%), and the price of soda and other nonalcoholic beverages had increased much more slowly than the overall average (index ≈ 75%). These trends have raised the concern that comparatively more healthy choices have become relatively more expensive. For example, a French study found that energy dense foods like fats and oils, sugar, refined grants, and others were the cheapest (Drewnowski and Darmon, 2005).

With regard to incomes, the United States is sufficiently prosperous that food uses up less than 10% of income for most households. Yet, there remains a significant population of low-income Americans for whom the income constraint is more influential. The overall US poverty rate (share of persons in families with incomes below the official poverty line) recently increased from 11.3% in 2000 to 15.1% in 2010, which nearly equals the peak levels of 15.2% during the recession of the early-1980s. Real median household income rose only slowly from 1980 to 2000 and did not increase further after 2000.

Beyond prices and incomes, more elaborate nutrition-oriented theories point our attention toward additional broad trends in recent years. Time spent preparing food has fallen as technologies have changed. At the same time, work has

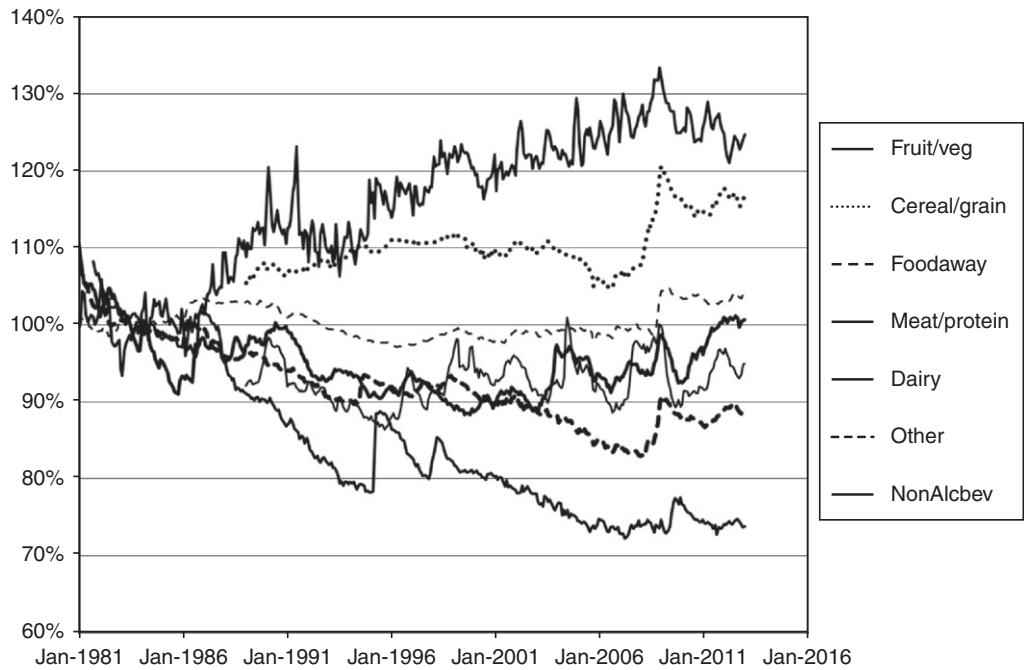


Figure 1 Relative prices of various foods, seasonally adjusted US city averages, 1981–2010 (BLS Consumer Price Index).

changed, perhaps reducing calories expended thereby. Women are increasingly in the workforce and spending less time in home production. Restaurant food and convenience food have become more important food sources.

A Framework for Nutrition Policy Options

Market Outcomes and Market Failures

Economic theory suggests a rationale for society to rely predominantly on private-sector markets to produce and distribute food. At their best, markets can overcome major challenges in motivating economic actors to make socially beneficial production and consumption decisions, while also providing them with the information they need to do so. The traditional economic theory suggests that governments should defer to market outcomes except in specific circumstances of 'market failure.' Several market failures that have been suggested as motivations for government will be discussed. In addition, other noneconomic perspectives with influence in food and nutrition policy will be discussed.

Economists focused on nutrition concerns, particularly obesity, have suggested three market failures that could motivate government action.

- Food choices of children. Most economic theory assumes that consumers are rational adults who can make their own spending decisions. There is a strong justification for government institutions, such as public schools, to promote nutrition for children. (Public health motivations for policies that also affect adults are mentioned below.)
- Imperfect information. Market outcomes are efficient only if consumers (and other actors) have the information they need to make purchases that satisfy their preferences. Consumer preferences include nutritional qualities and food safety as well as the more obvious taste qualities of food. Private-sector markets offer plenty of profit incentive for providing food that tastes good, but they may offer an insufficient incentive for providing information about nutrition qualities. Asymmetric imperfect information (discussed in Section Imperfect Information) is a commonly cited justification for government activities that affect the food system.
- Negative externalities. Market outcomes may not be optimal if there are externalities, where one person's decisions affect other people's well-being through nonmarket interactions. One type of interaction comes from the operation of insurance and health programs. For a person who receives health care from the government, the financial costs of illness are paid largely by taxpayers. Even for people who have private insurance, through their employer or by paying their own insurance premiums, financial costs of illness are shared with other people in the same insurance risk pool. It seems possible that insurance markets could reduce the incentive to maintain a healthy weight. However, some economists suggest that labor markets may already partly compensate for variations in insurance costs, by paying some workers more than others, in which case the negative externalities through health insurance markets would be smaller (Bhattacharya and Bundorf, 2009).

There are several additional reasons why food and agricultural markets may not satisfy the traditional assumptions of perfect competition. In some industries, there may be few competing firms. In other industries, such as the seeds for genetically modified food crops, patent rights may give a small number of firms significant market power. In yet other industries, such as food manufacturing and quick service restaurants, the predominant industry structure may be monopolistic competition, where each firm supplies differentiated branded food products and yet must compete to a certain extent with other firms that provide similar products. The authors return to the topic of differentiated products in discussing food advertising in the Section The Economics of Food Advertising.

Imperfect Information

A situation of imperfect information is described as 'asymmetric' if the producer knows more about a food product's attributes than the consumer does. Relevant attributes may include taste, wholesomeness, safety, and nutrition qualities. Under asymmetric information, food product attributes may be classified based on how the consumer learns about the product:

- For 'search attributes,' the consumer can perceive a product's quality even before purchase.
- For 'experience attributes,' the consumer can discover a product's quality after purchasing it. Even though it may be impossible to reverse a particular purchase if the product quality was unsatisfactory, the consumer can learn lessons that are useful for future purchases.
- For 'credence attributes,' the consumer relies to a greater extent on information provided by others, including the seller or by a third party.

A single good may have several attributes. For example, a bag of organic carrots may be seen as a search good for a shopper trying to follow the Dietary Guidelines recommendation to consume orange vegetables; the same bag may be seen as a credence good for a shopper seeking organic produce. Common government responses to asymmetric information range from process regulations (such as specifying what additives are safe to use in food) to food-grading systems to labeling rules (discussed in Section Policy Response).

In addition to having imperfect information, consumers may not satisfy the traditional economic theory's assumptions about rationality. A lively and rapidly growing body of research addresses behavioral economics, including strategies for 'nudging' economic actors in the direction of more optimal food choices, without taking away their freedom to make their own decisions (see Section Behavioral Economics: Nudges).

Policy Response

Economists commonly favor government policies that narrowly target a market failure that has been identified. In situations where there is no market failure, such as when well-informed adults freely choose and accept the consequences of

unhealthy eating patterns, many economists say there is no need for a government policy response.

In addition to the economic perspective on diagnosing market failures, there are other motivations that strongly influence government policies regarding nutrition and thus are given attention here to help researchers understand ongoing policy debates.

- A 'public health' perspective gives less deference to market outcomes and favors use of a broad range of policies to affect food choices and dietary quality.
- A 'consumer activist' perspective argues that food and beverage companies create a toxic food environment and should be more strongly regulated.
- A wide variety of 'producer' perspectives favor government nutrition policies that promote the interests of particular sectors of the agricultural, food manufacturing, and food retail industries.
- An 'egalitarian' perspective focuses on income or resource inequality as a motivation for government intervention in food and agricultural markets. For example, a motivation for federal food assistance programs is not just nutrition promotion but also poverty alleviation.

Responding to both economic and noneconomic motivations, leading policies and policy proposals fall into four broad categories, discussed in the next four sections: food assistance programs, information policies, direct price interventions such as taxes and subsidies, or government restrictions or subsidies to supply. The authors conclude with a discussion of behavioral economics and nudges.

Food Assistance Programs

One way for governments to address nutrition concerns is directly, through food assistance and nutrition programs. This section provides some background about food assistance programs, explains how such programs affect household budgets, and describes research that seeks to measure program effects.

Background on Food Assistance Programs

Food assistance programs may provide food through several mechanisms, including the following: (1) broadly targeted food benefits that low-income consumers may use to purchase food through normal retail channels, (2) more narrowly targeted food vouchers for purchase of specific foods and beverages with particular nutritional qualities, and (3) direct provision of free meals. In addition to food assistance programs, more general income support programs may have benefits for nutrition or food security. This section principally uses US food assistance programs as examples of each type of program, because food assistance plays a bigger role in the social safety net in the United States than in other developed countries.

1. Broadly targeted food benefits. In the United States, the largest food assistance program is the Supplemental Nutrition Assistance Program (SNAP), formerly known as food stamps. SNAP provides targeted benefits for food and nonalcoholic beverages from authorized grocery retailers through Electronic Benefit Transfer (EBT) cards similar to debit cards. It served 40.3 million people per month on average during fiscal year 2010 at a total cost of $68.2 billion. Program eligibility historically has depended largely on having income less than 130% of the federal poverty standard, so the program is counter-cyclical, and caseloads have recently risen to record levels during the recent recession. The primary purpose of the program is to prevent hunger and promote food security.
2. Narrowly targeted food vouchers. The Special Supplemental Food Program for Women, Infants, and Children (WIC) provided nutrition counseling, services, and a package of particular high-nutrient foods and infant formula to approximately 9.2 million people per month, at a cost of $6.8 billion in fiscal year 2010. Only pregnant and postpartum women, infants, and children through the age of 4 years are eligible. Eligibility also requires household income less than 185% of the federal poverty standard or participation in one of several other safety net programs, plus evidence of nutrition risk broadly defined.
3. Direct provision of meals. The National School Lunch Program served 31.6 million lunches, and the smaller and newer School Breakfast served 11.6 million breakfasts, on average, each school day in fiscal year 2010, at a cost of $13.3 billion. A free meal requires income less than 130% of the federal poverty standard, though all school provided meals in schools participating in the federal school meals programs are subsidized to some extent. The Child and Adult Care Food Program served meals in centers and home day care settings, costing $2.6 billion. These programs have primary nutrition goals, but antihunger effects are acknowledged as important secondary purposes.
4. Cash assistance, cash-based social insurance, and tax credits. Finally, governments provide cash welfare, which can be used for food as well as other products. In 2009, 1.8 million families with children obtained Temporary Assistance for Needy Families cash benefits, with total cash benefit payments amounting to $9.3 Billion. Supplemental Security Income (SSI) provided $41 billion in cash assistance for 6.4 million low-income individuals who were disabled, blind, or elderly. Unemployment insurance is typically viewed as a social insurance program but is also another important part of the safety net, with payments of approximately $131 billion in 2009. Tax credits such as the Earned Income Tax Credit may also play a role.

How Food Assistance Affects Household Budgets

The effect of food assistance on food choices depends on the role of program benefits in the household budget. To understand this role in a rigorous way, it helps to compare the effects of: (1) providing a targeted food assistance benefit or (2) providing a hypothetical equivalent cash subsidy. For example, consider a monthly voucher that provides a family with $50 for use in purchasing qualifying food products. The hypothetical comparison program provides $50 in cash.

- If a family with the hypothetical cash benefit would have spent less than $50 per month on qualifying foods, the family is called extramarginal or constrained by the form of the benefits, because the voucher program causes more food spending than would otherwise have occurred.
- In contrast, if a family with the hypothetical cash benefit would have spent more than $50, the family is called inframarginal or unconstrained by the form of the benefits, because it is able to purchase its desired amount of qualifying foods under either the targeted voucher or the hypothetical cash program.

Economic theory predicts that a marginal increase in targeted food assistance benefit will strongly affect food spending for extramarginal participants, and the increase will only weakly affect food spending for inframarginal participants. Empirical research commonly finds bigger program effects than expected for inframarginal participants in targeted food assistance benefits programs (Meyerhoefer and Yang, 2011).

Measuring the Effects of Food Assistance Programs

A challenge discussed elsewhere in this volume is obtaining causal estimates of the effects of food assistance programs. Due to selection bias in who takes up these programs, one cannot merely compare the outcomes of recipients and others. Given that participation is tied to low income and asset holdings or poor nutritional status or both, it is clear that comparisons of outcomes for program recipients to those of the general population are likely to be biased estimates of the effects of these programs. Even among those eligible for the programs, recipients may be positively or negatively selected compared with eligible nonparticipants due to the fact that participation is a choice variable. If recipients are healthier, more motivated, or more knowledgeable about the programs than nonparticipants, comparisons may suggest the program has a more positive effect than it actually does. Alternatively, if participants are more disadvantaged than eligible nonparticipants, comparisons of these two groups could lead to underestimates of the effects of the program. For example, skeptics often attribute many of the positive effects of the WIC program to positive selection among women participating in the program, although Bitler and Currie (2005) find little evidence than WIC participants who also participate in Medicaid are positively selected.

There are several approaches that researchers have taken to avoid selection bias (Gundersen *et al.*, 2011). Just four approaches are mentioned here. One approach is to compare outcomes among individuals in geographic areas with different program rules. This approach is comparatively less useful for SNAP policies that are national in scope and comparatively more useful for SNAP policies that have substantial state-to-state variation (Ratcliffe *et al.*, 2011).

A second approach, when program rules do not vary across areas is to look at the effects of the introduction of programs; comparing otherwise similar individuals in places before and after programs are introduced. The introduction of the food stamp program led to increases in food consumption (Hoynes and Schanzenbach, 2009), whereas the introduction of the WIC program led to an increase in average birth weight (Hoynes *et al.*, 2011). A limitation with this approach is that program rules may change over time, potentially raising questions about the ongoing validity of historical estimates for evaluating programs today.

A third approach to studying the effects of programs is to use random assignment. If program administrators are able to randomly assign an offer of program participation to otherwise identical eligible individuals, then comparisons of those assigned to be eligible for the program with those denied the option to participate can yield unbiased estimates of the effects of the program. In the case of food stamps, there have been several demonstration projects funded by USDA, which have yielded evidence about the effects of cashing out food stamp benefits on food spending.

A fourth approach to studying the effects of such programs is to use variation in program rules that use exogenous thresholds in income, age, or other characteristics to assign program eligibility. These regression discontinuity approaches compare otherwise similar individuals who because of small differences in a characteristic such as age face different program rules, while controlling for age. These 'regression discontinuity' approaches typically bring both considerable internal validity and limited external validity given the local nature of the estimates they yield.

Much progress has been made on the effects of food assistance programs, yet there are outstanding questions. For example, how does a program like WIC obtain positive results with relatively low benefit levels? What channels do effects of these programs work through? What is the role of information in the effects of these programs?

The Economics of Information Policy

A second type of public policy intervention to address nutrition concerns is to seek to influence the information environment in which consumers make food-related decisions. For example, governments may seek to promote nutrition through dietary guidance, regulation of food labeling, and regulation of advertising. Private mechanisms that affect quality include brand names and reputation as well as standard setting.

Information policies may be classified on a spectrum, running from mandatory information, to voluntary information provision chosen freely by the producer, to voluntary information with restrictions imposed by the government, to outright prohibitions against a particular type of information provision. This spectrum is shown in **Figure 2** (adapted from Wilde (2013)).

At each point along this spectrum, the policy debate depends in part on what one believes are the real facts about the relationship between food decisions and health. Some information sources recommend diets that are low in carbohydrates; some recommend diets that are low in fat and high in plant foods; some say humans as omnivores can thrive on either of these diets so long as we avoid highly processed manufactured foods. In the United States, the federal government's 'Dietary Guidelines for Americans,' issued every 5 years, recommend a diet with plenty of fruits, vegetables, whole grains, and low-fat dairy products, within the context of

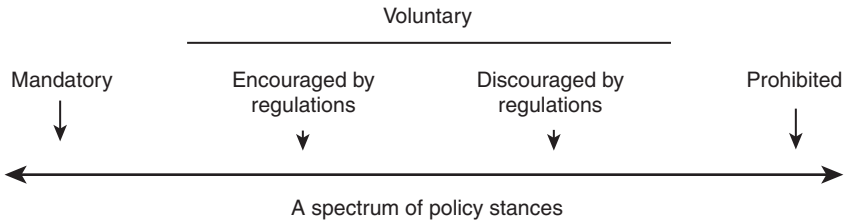


Figure 2 A spectrum of policy stances on information provision.

an overall diet that maintains a healthy weight by balancing total energy intake with energy expenditure needs. In promulgating dietary guidelines, the government seeks to remedy some of the confusion in private-sector information markets by identifying dietary principles that are supported by the balance of the scientific evidence.

There is comparatively little need for government regulation of food labeling and advertising for food attributes such as taste, which can be confirmed by search and experience, because correct information about such attributes is readily available to the consumer. The government has a more substantial role in regulating claims about nutrition and health qualities, which are credence attributes.

Yet, even for credence attributes, markets sometimes function well on their own. Even before nutrition facts panels became mandatory in the early-1990s, many food products carried nutrition information voluntarily. One might at first suspect that only the healthiest products would provide nutrition information, but competitive pressure may force a wider range of products to disclose information. Once consumers catch on that nondisclosure indicates that a product lacks a healthful characteristic that competitors have, there is an incentive for all but the least healthy products to disclose information voluntarily. This theory of competitive information disclosure works better in some situations than in others. For example, if all products in a food category share a certain characteristic, such as the dietary cholesterol in eggs, there is no incentive for companies to advertise the shortcomings of their competitors.

Regulation of Food Labeling

US food labeling policies include rules that mandate some kinds of information provision and rules that prohibit other kinds. In the first half of the twentieth century, the US federal government established standards of identity for many food products and began requiring disclosure of net weights and ingredients for manufactured food and beverages. Nutrition facts panels were introduced on a voluntary basis in the 1970s and became more widespread during the 1980s. Since the passage of the 1990 Food Labeling and Education Act (NLEA), nutrition facts panels have been mandatory on most packaged food in the United States. More recent legislation has required mandatory country-of-original labeling (COOL) on a wider variety of food products. The 2008 Farm Bill for the first time required disclosure of calorie information in chain restaurants. Bollinger *et al.* (2011) look at the effects of a policy in New York City requiring chains to post calories for food on purchases at Starbucks, finding that average calories per transaction fell, and this was through changes in purchases of food, not beverages.

Until the 1980s, the Food and Drug Administration (FDA) generally refused to permit health claims on food labels, fearing that consumers would be misled. The 1990 NLEA allowed health claims on food labels if there is 'significant scientific agreement' about the merit of the claims. For example, many low-sodium products may use a health claim that reducing sodium can lower high blood pressure.

Subsequent legislation and court cases have forced the FDA to allow a broader range of claims. The 1994 Dietary Supplement and Health Education Act (DSHEA) led to a weaker standard of evidence for what are called structure–function claims, which do not mention a specific disease as health claims do. For example, 'calcium builds strong bones' is a structure–function claim. More recently, in response to a successful lawsuit about health claims on dietary supplements, the FDA has begun to allow health claims for which the evidence is not strong enough to quality as significant scientific agreement, so long as the food package label includes a disclaimer describing the level of scientific evidence for the claim.

Remedies for the problem of misleading claims may come from the private sector, the government, or both in combination. The for-profit media frequently cover food issues, describing recent research in nutrition science and food safety. Private not-for-profit organizations also can serve as independent watchdogs. In some cases, government agencies can take steps to share information more widely, thereby allowing private sector market incentives to function better. Jin and Leslie (2009) study the question of reputational incentives through the example of restaurant hygiene. Using data from Los Angeles County, they find that reputational incentives indeed provide a market-based mechanism for quality, but that quality increased further when the government-issued hygiene report cards for restaurants.

In one business model, not-for-profit organizations and for-profit businesses may offer third-party certification of nutrition labeling claims made by food companies. For example, in return for a fee paid by food companies, the American Heart Association (AHA) allows food manufacturers to label qualifying products with an AHA heart-check symbol certifying that they are low in saturated fat and cholesterol or high in whole grains. In still other cases, the government helps set standards for a food label claim, which then may be used voluntarily by private-sector businesses that meet the standard. For example, the 'certified organic' label may only be used on foods that meet a checklist of process standards, which exclude the use of certain chemicals and practices in

agricultural production and food processing. Qualifying food producers choose voluntarily to produce food organically. Food retail chains have considerable influence over the adoption of voluntary food labeling strategies. If a retailer with a large market share increases its marketing of certified organic products, or develops a new front-of-pack nutrition label, these decisions influence decisions by food manufacturers and other suppliers throughout the food marketing chain.

The Economics of Food Advertising

Under perfect competition, producers of a standardized commodity have less incentive to advertise, because each firm's investment in advertising would reap gains in consumer demand that are shared with all of the firm's competitors. Each perfect competitor would have an incentive to be a 'free rider,' allowing competitors to bear the cost of advertising. Clearly, this model of industry structure does not provide a compelling description of the food and beverage manufacturing industries or the chain restaurant industry, where heavy advertising is widespread.

Instead, a model of industry structure that better describes these industries is monopolistic competition. In this model, each firm is the monopoly producer of its own branded product, and it competes to a certain degree with competitors who are similar. Competitors may be similar because they are physically nearby, as in the case of local food retailers who compete most strongly with other firms that are geographically close. Competitors also may be similar in a psychological sense, as in the case of quick-service restaurants that serve a similar clientele and occupy a similar marketing space. Under monopolistic competition, firms have a strong incentive to advertise.

Direct Interventions: Taxes and Subsidies

Some public-health advocates, researchers, and policy-makers recommend that the government go even further, guiding consumers toward healthier diets by altering prices, taxing less healthy foods and beverages and subsidizing their more healthy counterparts. In 2010, the Director of the Centers for Disease Control and Prevention, Thomas Frieden, and two CDC colleagues wrote: "A tax of 1 cent an ounce on sugar-sweetened beverages – about a 10 percent price increase on a twelve-ounce can – would be likely to be the single most effective measure to reverse the obesity epidemic" (Frieden *et al.*, 2010).

The success of such a proposal depends on the size of the consumer response to a change in the price of food. Recall from Section 2 that an own-price elasticity shows the percentage change in a product's quantity consumed, in response to a 1% increase in the product's price. Economists who study taxation policy recommend that taxes be placed on goods for which demand is inelastic (not responsive to a change in price), because such taxes raise more revenue and do not distort the market equilibrium as much as taxes on goods with elastic demand, causing smaller dead-weight losses from the policy. In contrast, health policy advocates tend to prefer that taxes be placed on unhealthy foods whose demand is elastic (responsive to a change in prices), because these taxes have the biggest impact on food choices. Some clever proposals seek the best of both worlds, by taxing an unhealthy food and earmarking the resulting revenue for health promotion programs.

Table 1 Selected own-price, cross-price, and expenditure elasticities for food demand

Description	*Elasticity*
Mean value of the own-price elasticity for selected foods (literature review by Andreyeva *et al.*, 2010):	
Food away from home	−0.81
Soft drinks	−0.79
Juice	−0.76
Beef	−0.75
Pork	−0.72
Fruit	−0.70
Poultry	−0.68
Dairy	−0.65
Sweets/sugars	−0.34
Eggs	−0.27
Own-price elasticities for salty snacks (Kuchler *et al.*, 2004):	
Potato chips	−0.45
All chips	−0.22
US beverage demand elasticities, 1998–2007 (Smith *et al.*, 2010):	
Caloric sweetened beverages (own-price elasticity)	−1.264
Caloric sweetened beverages (elasticity of response to price of juice)	0.233
Caloric sweetened beverages (expenditure elasticity)	1.054
Juice (own-price elasticity)	−1.012
Juice (elasticity of response to price of caloric sweetened beverages)	0.557
Juice (expenditure elasticity)	0.878

There is a large research literature that seeks to estimate consumer demand elasticities. A sampling of results from many such studies is presented in Table 1. To interpret such estimates correctly, the reader must keep in mind that food groups are defined differently in different studies. For example, own-price elasticity for a narrowly defined food group with many substitutes (such as potato chips) will be larger in absolute value than the own-price elasticity for a more broadly defined food group with fewer substitutes (such as all types of chips combined).

Also, to correctly anticipate the nutrition consequences of a proposed tax, the reader must consider cross-price effects. For example, Smith *et al.* (2010) combine data on consumer grocery purchases of beverages from the Neilsen Homescan panel with data from the National Health and Nutrition Examination Survey on individual consumption of beverages and height and weight. They first use the variation across places in beverage purchases to estimate a demand system for the effects of price changes for caloric sweetened beverages on consumption of drinks. The corresponding price elasticities are used to predict the effect of a proposed tax on caloric sweetened beverages, not only on consumption of the targeted beverages but also on potential substitutes, including juice. Using their estimates, Table 1 shows that each 1% increase

in the price of caloric sweetened beverages might reduce intake of these targeted beverages by 1.264%, whereas simultaneously increasing the intake of juice by 0.557%. After considering both direct effects and such substitutions toward other beverages, the authors estimated that a tax-induced 20% price increase on caloric sweetened beverages would change overall food energy intake by enough to reduce adult overweight prevalence from 66.9% to 62.4% (Smith *et al.*, 2010). The Smith *et al*. study addresses issues missed in some other demand studies; including careful attention to cross-good price elasticities and differentiating caloric sweetened beverages from diet beverages; and is also clear about what cannot be estimate (the effect of food purchased away from home is naturally absent from data on grocery store purchases).

However, there is still a challenge in applying these or other estimates to predict effects of a large tax change on sugar-sweetened beverages. Many such demand studies make use of time series variation in prices and aggregate data on expenditures, which may or may not be applicable to changes in taxes and individual-level consumption. It is also challenging to find appropriate data to estimate effects fully on complements and substitutes for specific goods, and often the analyst has data only on purchases at grocery stores, or alternatively, consumption data with no corresponding price and location of purchase. One study that comprehensively analyses purchases across categories of goods using household data for 2002–07 while also accounting for local access to particular store types, demographics, prices, and nutrients is Harding and Lovenheim (2013), who simulate the effects of both product specific and nutrient taxes. Yet, ideally data on food at home and away from home would be available along with prices and quantities.

An alternative approach uses individual-level data and tax changes. Fletcher *et al.* (2010) find that for children and adolescents, tax-induced reductions in soft drink consumption are offset with increases in consumption of milk and other beverages, raising doubt about the value of such taxes in reducing obesity. Unfortunately, one limitation to this approach is that samples in datasets like NHANES are relatively small, and the combined state-year panel are limited by sample design.

Nutrition taxes face several sources of opposition. First, they generally are regressive, with a higher relative budget impact for low-income populations than for higher-income populations. Second, as noted throughout the article, there may be disputes over the nutrition science on which they are based, although one exception with somewhat broader – but far from universal – scientific support is a tax on caloric sweetened beverages. Quantifying the causal effects of prices and taxes on food consumption and health is an ongoing area of research where new approaches and data would be useful. New causal evidence awaits more policy interventions, better data on purchase prices and quantities at and away from home, or both.

Government Supply Interventions

The final set of policies we consider is related to increasing or restricting access to food. One subject which has received much attention is whether some areas have sufficient access to appropriate food. In a relatively large public-health literature, evidence is presented about the correlations between areas with a high concentration of low income residents and a dearth of large retail food stores selling healthy foods such as food and vegetables. Congress mandated in the 2008 Farm Bill that USDA study this so-called problem of food deserts (areas with limited access to affordable and nutrition food) and suggest policy responses, resulting in a report (USDA Economic Research Service, 2009), a food desert locator, as well as other action. Although there is ample evidence that some local areas have limited food access, little or no research has established the causes of such limited access, and such information is a key input to designing appropriate policies (Bitler and Haider, 2011). For example, limited access could be the result of supply or demand factors, and if it is the result of demand factors, supply interventions are not likely to ameliorate deficiencies. Yet further policy intervention seems likely, and may provide useful variation for new research.

Another policy lever that is widely discussed is zoning or other regulations limiting the types of food establishments or types of foods available in various locations. This is in part based on findings about locations of fast food outlets affecting calorie intake and obesity (Currie *et al.*, 2010).

Policy-makers also may limit sales of competitive foods in schools (competitive foods are all foods offered for sale at schools besides those provided by USDA school meals programs). Should more localities enact policies banning such sales, it may provide variation to understand how access to such foods affects school health. Schools facing financial pressures are more likely to allow competitive food sales and have students with larger BMI (Anderson and Butcher, 2006). However, some research finds that such sales do not necessarily lead to more consumption of junk food, suggesting substitution across in school and out of school locations (Datar and Nicosia, 2012).

Behavioral Economics: Nudges

The economic understanding of consumer responses to prices and income and the policy proposals for new subsidies or taxes and supply interventions all rely on an economic theory of consumer choice. A lively body of current economic research investigates situations where consumers do not behave rationally, perhaps leading to opportunities for 'nudging' consumers toward more healthful choices (Thaler and Sunstein, 2008).

Neoclassical theory predicts that consumers will eat less when the marginal cost of an additional unit – the price to the consumer – is higher. They will tend to overeat at an all-you-can-eat restaurant, because the marginal cost of additional food is zero, no matter what the entry price of the meal. Yet, surprisingly, recent research found that consumers of an all-you-can-eat pizza meal actually consumed more pizza if the price of the meal was higher (Just and Wansink, 2011).

These differences between actual consumer behavior and traditional economic assumptions about rational behavior do not mean consumers are irrational or foolish in the everyday sense of the term. Instead, these behaviors may show that consumers need to simplify the cognitive burden of difficult

decisions by following predefined heuristics or 'rules of thumb.' Some of these heuristics are the subject of considerable research:

- Default offerings may affect consumer choices. For example, if a quick service restaurant chain includes milk by default in children's meals, customers may agree to purchase the milk with the meal. Yet, if the chain includes soda by default, the customers may more frequently keep the sugar-sweetened beverage rather than make a special effort to request milk.
- Distractions also may affect consumer choices. For example, it has been found that consumers who were required to make other decisions at the same time were more likely to choose cake over fruit salad, whereas consumers who were not distracted were more likely to choose the healthier offering (Shiv and Fedorikhin, 1999). Hunger or time stress also may affect people's decisions.

This new approach to behavioral economics has raised some hopes for inexpensive nutrition improvements, by making subtle changes to the setting or environment in which choices are made. For example, some suggest that students in school meals programs might make better decisions if the location of the salad bar were altered, or if a different tender (cash or school meals program card) were required for different products. This approach also has generated renewed scrutiny of the empirical evidence for other health policy proposals, such as taxes on less healthy food or new labeling rules for restaurants (Loewenstein, 2011). Of course, many of the same lessons can also be used by food marketing professionals to promote food options with any health profile. Future research will determine whether these new tools of behavioral economics make a small or big difference for consumer choices. And, if the effect is big, future developments in both social and commer cial marketing will determine whether the changes are helpful for dietary quality. In either case, the willingness to scrutinize assumptions and follow the empirical evidence in new directions is entirely good news for future research on the economics of nutrition.

See also: Health Econometrics: Overview. Instrumental Variables: Informing Policy. Instrumental Variables: Methods. Intergenerational Effects on Health – *In Utero* and Early Life. Macroeconomic Causes and Effects of Noncommunicable Disease: The Case of Diet and Obesity. Nonparametric Matching and Propensity Scores. Nutrition, Health, And Economic Performance. Panel Data and Difference-in-Differences Estimation

References

Anderson, P. and Butcher, K. (2006). Reading, writing, and refreshments: Are school finances contributing to children's obesity? *Journal of Human Resources* **61**(3), 467–494.

Andreyeva, T., Long, M. W. and Brownell, K. D. (2010). The impact of food prices on consumption: A systematic review of research on the price elasticity of demand for food. *American Journal of Public Health* **100**(2), 216–222.

Bhattacharya, J. and Bundorf, M. K. (2009). The incidence of the healthcare costs of obesity. *Journal of Health Economics* **28**, 649–658.

Bitler, M. P. and Currie, J. (2005). Does WIC work? The effects of WIC on pregnancy and birth outcomes. *Journal of Policy Analysis and Management* **24**(1), 73–91.

Bitler, M. P. and Haider, S. J. (2011). An economic view of food deserts in the United States. *Policy Retrospectives. Journal of Policy Analysis and Management* **30**(1), 153–176.

Bollinger, B., Leslie, P. and Sorensen, A. (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy* **3**(1), 91–128.

Currie, J., Della Vigna, S., Moretti, E. and Pathania, V. (2010). The effect of fast food restaurants on obesity and weight gain. *American Economic Journal* **2**, 32–63.

Datar, A. and Nicosia, N. (2012). Junk food in schools and childhood obesity: Much ado about nothing? *Journal of Policy Analysis and Management* **31**(2), 312–337.

Drewnowski, A. and Darmon, N. (2005). Food choices and diet costs: An economic analysis. *Journal of Nutrition* **135**(4), 900–904.

Dubois, P., Griffith, R. and Nevo, A. (in press). Do prices and attributes explain international differences in food purchases. *American Economic Review*.

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W. and Dietz, W. (2009). Annual medical spending attributable to obesity: Payer-and service-specific estimates. *Health Affairs* **28**(5), w822–W831.

Fletcher, J., Frisvold, D. and Tefft, N. (2010). The effects of soft drink taxation on soft drink consumption and weight for children and adolescents. *Journal of Public Economics* **94**(11–12), 967–974.

Fogel, R. W. (2004). *The Escape from hunger and premature death, 1700–2100 Europe, America, and the Third World*, vol. 38. New York, NY: Cambridge University Press.

Frieden, T. R., Dietz, W. and Collins, J. (2010). Reducing childhood obesity through policy change: Acting now to prevent obesity. *Health Affairs* **29**(3), 357–363.

Gundersen, C., Kreider, B. and Pepper, J. (2011). The economics of food insecurity in the United States. *Applied Economic Perspectives and Policy* **33**(3), 281–303.

Harding, M. and Lovenheim, M. (2013). The effect of product and nutrient specific taxes on shopping behavior and nutrition: Evidence from scanner data. Cornell, NY: Cornell University Working Paper.

Hoynes, H. and Schanzenbach, D. (2009). Consumption responses to in-kind transfers: Evidence from the introduction of the Food Stamp Program. *American Economic Journal: Applied Economics* **1**(4), 109–139.

Hoynes, H. W., Page, M. and Stevens, A. (2011). Can targeted transfers improve birth outcomes? Evidence from the introduction of the WIC program. *Journal of Public Economics* **95**(7–8), 813–827.

Jin, G. and Leslie, P. (2009). Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics* **1**(1), 236–267.

Loewenstein, G. (2011). Confronting reality: Pitfalls of calorie posting. *American Journal of Clinical Nutrition* **93**, 679–680.

Meyerhoefer, C. D. and Yang, M. (2011). The relationship between food assistance and health: a review of the literature and empirical strategies for identifying program effects. *Applied Economic Perspectives and Policy* **33**(3), 304–344.

National Center for Health Statistics (2011). *Health, United States, 2010: With special feature on death and dying*. MD: Hyattsville.

Ratcliffe, C., McKernan, S. M. and Zhang, S. (2011). How much does the Supplemental Nutrition Assistance Program reduce food insecurity? *American Journal of Agricultural Economics* **93**, 1082–1098.

Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research* **26**(3), 273–292.

Smith, T., Lin, B.-H. and Lee, J.-Y. (2010). Taxing caloric sweetened beverages: Potential effects on beverage consumption, calorie intake, and obesity. *ERR-100*. Washington, DC: U.S. Department of Agriculture, Economic Research Service.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven Connecticut: Yale University Press.

USDA Economic Research Service (2009). Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences, *Report to Congress, June*. Available at: http://www.ers.usda.gov/publications/ap-administrative-publication/ap-036.aspx#.Ub-RKvYjqn9 (accessed 16.06.13).

Wilde, P. E. (2013). Food Policy in the United States: An Introduction. New York, NY: Routledge/Earthscan.

Nutrition, Health, and Economic Performance

DE Sahn, Cornell University, Ithaca, NY, USA

Glossary

Demographic dividend The period of time during the process of economic development where there is a dramatic drop in the dependency ratio owing to the lag between the drop in fertility that follows from reductions in child and infant mortality rates that can contribute to more rapid economic growth.

Health inequality Differences in the incidence of health care spending, access to health care, incidence of disease, and/or health outcomes between different populations groups.

HIV/AIDS Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome.

Life course An approach that examines how early events in a person's life influences future decisions and outcomes, and thus examines the sequencing of events in the life of an individual.

Worker productivity Measures the efficiency of production of a unit of labor, and is formally the total output per unit of input of labor.

Introduction

Health and nutrition outcomes are critical to the well-being of households and individuals and their economic productivity and prosperity. Although it seems evident that a debilitated worker will be less productive, there are numerous indirect, subtle, and complex pathways that link poor health and nutrition to economic output, such as sick children, or children of sick adults, being less likely to accumulate other forms of human capital, and disease-ridden societies with high infant and child mortality rates having higher fertility, with the consequent economic burdens associated with the risks of child bearing and a higher population growth rate. Compounding the challenges in understanding the impact of health on economic outcomes is the complexity of the temporal dimensions, as the productivity consequences of poor health extend far beyond the short term and affect outcomes across the life course and from one generation to the next.

The relationship between health and economic outcomes is particularly important in developing countries. First, health problems are most severe in these countries, and the ability to perform hard physical labor most important for employment. Second, self-employment and self-provisioning are of particular importance, and under such circumstances, reduced levels of output, from temporary ailments and disease, for example, can contribute to large consumption shortfalls – an outcome less likely to occur in more market-oriented economies. Third, the propensity for market failures, such as those of credit markets, will also simultaneously contribute to economic inefficiencies as mediated by the underinvestment in health and agricultural capital. This raises the prospect of the poor being caught in a low-level equilibrium with binding constraints in terms of time available to devote to the production of health, home production (e.g., care of children), and farm production.

It is also notable that the link between health and productivity is of special importance for women, who often assume a predominant role in the production of food crops. The greater vulnerability of women also results from the extraordinarily high maternal mortality and morbidity, with one of nine women dying during childbirth in some regions of the world. Additionally, women suffer the acute burden associated with social norms and behaviors that have resulted in them bearing the brunt of the ravages of human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS), especially in Africa. Finally, women have unique responsibilities in the home, particularly in terms of care of children. Health and nutrition shocks that adversely affect women not only adversely affect their productive role in labor markets but also impact their joint production role as caregivers for their children, and thus they induce a recurring and intergenerational cycle of crisis and deprivation.

Before turning to a discussion of the evidence on the relationship among health, nutrition, and economic outcomes, the author emphasizes that the imperative of, or justification for, improving health and nutritional status goes beyond their importance in promoting economic growth. Rather, well-being is multidimensional, comprising factors such as good health and adequate nutrition. These capabilities are intrinsically important, and merit recognition above and beyond their impact on productivity, output, and money metric measures of poverty. Although this article focuses on the contribution of health and nutritional status to productivity and economic outcomes, this may be of secondary importance relative to the intrinsic value of health.

Impact of Health on Economic Development

Macro and Cross-Country Evidence

Economic historians have argued persuasively that nutrition and health have contributed in an important way to increases in productivity and economic growth. Among the seminal work in this area are papers by Robert Fogel who showed that inadequacies in diet contributed to disease and early mortality, greatly limiting the possibility for productive work in eighteenth century England and France. His estimates indicate that

 doi:10.1016/B978-0-12-375678-7.00107-3

50% of Britain's growth since 1800 was attributable to increases in dietary energy available for work and improvements in the efficiency in the transformation of nutrients, particularly calories, into work.

Numerous other authors have examined cross-country associations among health, economic growth, and poverty. Some work employs general measures of health, such as life expectancy, whereas other studies focus on specific diseases, such as the impact of malaria, tuberculosis, and HIV/AIDS on economic output and growth. Among the numerous studies to directly estimate the impact of aggregate measures of health, particularly life expectancy, on economic output, it has been estimated that an increase in life expectancy of 10% will lead to an increase in economic growth of 0.4% per year. These estimates are consistent with similar research, including work that indicates that an increase in life expectancy of 1 year raises gross domestic product (GDP) per capita by 4%. Using adult survival rates to measure health, it has been reported that if health status were equalized across countries, the variance of log GDP per worker would be reduced by 9.9%. The results also suggest that eliminating health gaps would reduce disparities in country level mean incomes. But overall, the results show relatively small effects of poor health on economic development compared to studies that rely on cross-country regressions. Another article suggests an even more limited impact of health on economic outcomes, using an instrumental variable (IV) approach to tell a surprising story of how improvements in life expectancy led to lower GDP per capita. This unexpected result is explained by the fact that increases in worker productivity were offset by rapid population growth in the face of fixed land and a base of physical capital that was slow to adjust, contributing to declines in income. Despite the rigor and compelling nature of this article, it has been criticized for many underlying assumptions, particularly that lagged health has no effect on economic outcomes, and likewise, that it does not address possibilities such as whether reductions in fertility will offset population increases accompanying lower mortality.

In disease-specific literature, a malaria ecology index has been used as an instrument in estimating cross-country regressions of GDP per capita. The results show a dramatic impact of malaria on growth. A concern with this work, as with other such articles, is that of omitted variables contributing to the malaria index having a greater negative effect than it would otherwise have in a more fully specified model.

More recently, much of the attention on macroeconomic impacts of disease has focused on HIV/AIDS. Early work observed little impact on economic growth. This optimism was in fact based on a Solow-type growth framework where the impact of disease on growth was mitigated by a drop in the supply of labor relative to that of capital, which in turn increases the productivity of labor. It has further been suggested that there is a low impact of HIV/AIDS on growth through a process of the epidemic contributing to reduced fertility and a decline in the dependency ratio that subsequently leads to increases in per capita consumption as well as savings. It is assumed that such changes will not only increase investment but also provide resources for health and related support for those suffering from AIDS. Another article also shows little impact of HIV/AIDS. Using an IV technique that relies on the rate of male circumcision as an instrument, it argues that the differences are attributable to exogenous cultural factors. Although the exclusion restrictions are certainly open to debate, the authors show that the circumcision rate is a strong predictor of HIV prevalence, and that it is uncorrelated with other determinants of growth.

These optimistic assessments stand in contrast with other more sobering findings. One report, for example, finds that a 1% increase in HIV prevalence will contribute to a marginal impact on income per capita of negative 0.59%. It argues that the excess labor arguments that have mitigated the macroeconomic impacts of AIDS are not being realized. Another article estimated that GDP was reduced by 17% and per capita incomes by 8% between 1997 and 2010 as a result of the AIDS epidemic in South Africa. Another group of researchers have published an article discussing the possibly devastating effects of HIV/AIDS if the epidemic in Southern Africa continues unchecked.

Although such estimates are informative, the challenges of arriving at actual details of the impact of these communicable diseases on economic growth are clearly daunting. They depend on the economic structure of each country, the relative importance of agriculture, whether land or labor are greater constraints to growth, and the existence of economic and social infrastructure. Thus, there is a need to better understand the intricacies of how HIV/AIDS (and other diseases) are impacting economic relationships and performance and the role that mediating factors, such as the effects of large number of orphans on education capital, play.

Likewise, the extent to which interventions such as the provision of antiretroviral therapy (ART) are available will have an enormous impact on such estimates, both through mitigating the productivity consequences of the disease and the fiscal costs associated with governments contributing to the treatment costs. Most estimates of the economic costs of HIV/AIDS were made before treatment with antiretroviral drugs was widely available. Recent studies have shown a dramatic reversal in the physical well-being of those being treated, clearly reducing the costs of disease in terms of productivity losses. At the same time, the costs of treatment will be staggering. Although these may be largely born by foreign donors, the ability or willingness of the international community to sustain the financial support for ART is questionable and will likely result in more of the burden falling on patients and local health systems.

As the costs of disease in developing countries is considered, there is an epidemiological transition underway as infectious diseases become less prominently a cause of death and disability; instead, there is an emerging epidemic of chronic disease. For example, in 2000 there were more than 7 million cases of diabetes in Africa alone, and it is estimated that direct treatment costs would exceed purchasing power parity US$1000 per person. Treating diabetes and other noninfectious lifestyle diseases in the future will be a formidable challenge for households and governments, with both incurring large financial costs.

Finally, another channel through which improved health will impact economic outcomes is the so-called 'demographic dividend.' The pathway is quite simple: improvements in health services and availability of modern technology will

bring about a decline in mortality, and after a considerable lag, fertility will fall in response to the expectation for longer life spans and higher probabilities of survival into adulthood. This will lead to a bulge (which in the case of East Asia has been estimated to last nearly 50 years) in the working age population relative to the rest of the population. This demographic transition will in turn contribute to a large economic dividend. However, it has been suggested that the demographic dividend will not necessarily materialize in sub-Saharan African countries for a range of reasons, including the slow rate of fertility decline and HIV/AIDS. A recent article revisits this issue and concludes that the demographic dividend can be expected to materialize in Africa. However, the article also points out the importance of institutional reform and a transparent political and economic environment as a prerequisite for the bulging number of working individuals to be productively engaged.

In sum, caution is necessary in interpreting literature on the macroeconomics of health, even that which makes efforts to deal with problems of omitted variables and endogeneity. Specifications are often ad hoc, data are often unreliable, and most importantly, even the best attempts to deal with problems of omitted variables and unobservables that may jointly affect health status and income are open to serious criticisms. Perhaps a greater lesson is that there is a need to better understand the (primarily microeconomic) pathways, such as the impact on worker productivity and schooling, through which health impacts economic outcomes. Similarly, to the extent that certain factors have reduced the potential impact of health improvements on economic outcomes, such as population growth, this suggests that policymakers consider emphasizing programs to control fertility, and similarly, consider promoting economic opportunities for the burgeoning labor force through, for example, encouraging foreign investment.

Microeconomic Evidence

Pioneering work on the efficiency wage theory links health and nutrition to labor market outcomes, with the basic idea being: output is a concave function of labor inputs, including the number and level of effort among workers. Higher wages will thus improve nutritional intake of workers, and subsequently effort. From the producer's perspective, therefore, the optimal wage will minimize the wage bill in terms of the wage rate divided by the effort level.

The test of this theory involves determining whether wages respond to nutritional intake of workers. The correlation between health and wages of individuals has been well established with household survey data, but making a causal argument is far more challenging. For example, healthier workers may also be better educated, and likewise, healthier workers may have parents who make choices that not only contribute to their better health but also instill a greater work ethic.

Several microeconomic studies have made serious attempts to overcome the econometric problems inherent in examining such as relationship. Much of the evidence has been comprehensively reviewed.

Among the research that relies on nonexperimental methods, height, which largely reflects health conditions and investments both *in utero* and during early childhood, is often employed as an indicator of general healthiness. There is compelling evidence of the productivity effects associated with greater stature in numerous studies from developing countries. There is evidence from the historical literature that height affected the price of slaves, presumably reflecting the expected probability gains associated with greater stature.

Another anthropometric indicator that is widely used is the body mass index (BMI), and results indicate a loss of productivity associated with leanness. Similarly, estimates of a farm production function for Sierra Leone finds that calories per adult equivalent have significant positive effects on the marginal product of agricultural labor. One study from Sri Lanka instruments per capita household calories using prices, and its results indicate that there is a positive effect on market wages for rural men but not women. Many academics, however, emphasize the limitations of relying on household calorie intake to measure the effect on productivity. Research on workers in Ghana and Côte d'Ivoire indicates that wage returns to height and BMI in Ghana were also quite large, with a centimeter increase being associated with an 8–10% increase in wages.

Other studies of the impact of nutrient consumption rely on individual 24 h recalls and the measuring of food prepared and/or consumed in the household. Here, the evidence is more mixed. One article studied the impact of individual calorie consumption on agricultural production functions and wage equations. Employing fixed effects to control for individual heterogeneity, results show no impact of calories on either the marginal product of agricultural labor or agricultural wage rates. Interestingly, an impact of weight-for-height, a measure of leanness like BMI, on these outcomes is found to be consistent with the findings of other work. A further study on rural India finds some interesting seasonal effects: calories have a greater impact on productivity in the peak season for men, but weight-for-height is more important in the preharvest season when work is less demanding. Another study on agricultural workers from the Philippines finds that individual calorie intake from 24 h recall has no significant impact on productivity, unlike BMI where the effects on earning are significant. Noteworthy is that all these findings ignore that there are likely additional labor productivity effects that operate through occupation choice.

The impact of days ill on productivity has also been examined in a number of articles. Researchers found that each extra day of illness in Peru contributed to a 1% decline in hourly earnings among male wage workers and a 3% decline among the self-employed. For females, the comparable figure is a 2% decline. Overall, however, the general picture emerges of reduced labor supply in response to illness, although the impacts on productivity are more mixed. This perhaps reflects that such studies are examining agricultural productivity, and as mentioned previously, there is considerable latitude for substitution of labor, either with other family members or hired labor.

Recently, a great deal of attention has been accorded to examining the micro impact of HIV/AIDS on productivity. A comprehensive review of this issue notes numerous studies that focus on the impact of AIDS illness and death on household incomes and expenditures, largely mediated through declines

in labor supply, a fall in farm production, and the burdens associated with spending on health care and funerals. Likewise, there is troubling evidence that these economic stresses often lead to household dissolution, and, of course, a dramatic increase in orphanage, which is shown to have significantly deleterious economic and social consequences.

There is also evidence that declines in labor availability due to illness lead households to change cropping patterns and cultivation practices. One study shows that although Kenyan households afflicted by AIDS protect land under food cultivation, land devoted to cash crop production declines. A similar finding has been reported for Uganda. Other studies, however, have not found such changes in labor supply. An interesting study that focuses on the impact of the provision of ART for AIDS patients in Kenya reports a 20% increase in the probability of being in the labor force 6 months after treatment and that the hours worked increases by 35% among the treated. Ethical consideration naturally precluded randomization of treatment, and instead the authors needed to rely on other survey data collected during the same time on households without treatment to control for time varying factors that could bias the estimate.

Even more compelling evidence on the links between health and productivity comes from experiments designed to isolate the causal impact of health on productivity and labor market outcomes. There has been a considerable amount of experimental research on the impacts of micronutrients on labor market outcomes, with perhaps the greatest attention given to examining the impact of iron deficiency. Two biological pathways have been identified. First, aerobic capacity declines with decreasing levels of hemoglobin. Depletion of iron stores also contributes to reductions in the amount of oxygen available to muscles. As a consequence, endurance suffers, and there are greater demands on the heart in order to achieve the same activity. Iron deficiency also raises susceptibility to disease and is associated with fatigue and impaired cognitive development. Noteworthy among the many studies that examine causal effects of iron supplementation are the impacts on the output of rubber workers in Indonesia, cotton mill workers in China, and tea plantation workers in Sri Lanka. Additionally, several studies have demonstrated how the cognitive development of children is impaired by iron deficiency.

A particularly interesting field experiment is the Work and Iron Status Evaluation study that provides iron supplements to older adults in Central Java, Indonesia. Approximately half the male workers in the study are self-employed (primarily as rice farmers), and the other half are paid a time-wage. There is no evidence that hours of work corresponded to the treatment for time-wage workers, although those receiving treatment reduced the amount of time spent sleeping, and there is evidence that after a year they took on more work in self-employment. Among males who earned a time-wage, there is no evidence of changes in productivity as indicated by their hourly earnings; of course, if their wages are set by an employer, it is not obvious the worker will reap the benefits of greater productivity. This is not true for the self-employed. Males who were self-employed and iron deficient at baseline reported approximately 20% higher hourly earnings after 6 months of supplementation relative to similar controls. Although the study demonstrates that iron deficiency has a causal impact on time allocation and economic productivity, it also highlights the importance of including behavioral responses to the experiment itself in assessing the impact of treatment.

Experimental evidence of other forms of nutrition interventions is less compelling. One study that randomized food supplementation of sugarcane cutters in Guatemala indicated that those living in treatment villages were not more productive than the control villages. Another study in Kenya found a limited impact of food supplementation on the productivity of road workers. An experiment in Indonesia exploited the application of user fees at randomly selected 'treatment' districts while prices were held constant (in real terms) in neighboring 'control' districts. Two years after the intervention, relative to control areas, health care utilization and labor force participation had declined in treatment areas (where prices had increased). Reductions in employment were particularly large (and significant) for men and women at the bottom of the education distribution, those whom we would expect to be the most vulnerable. The most plausible interpretation is that the average treatment effects on labor supply indicate a causal role of improved health on the allocation of time to the labor market.

Beyond the issue of worker productivity and labor market outcomes, the impact of health on schooling and cognition has also been widely studied. One study reports that in a randomized control trial, treatment of helminthic infections in schools contributes to a reduction in absentee rates by one-quarter, although it does not find an improvement in test score outcomes. Another study uses IVs and a fixed effects estimator and finds that stunted growth among young children will lead to delayed enrollment, but not eventual attainment. A further study from Zimbabwe that employs a quasi-experimental approach indicates a large impact of heights on school attainment. Finally, there is strong evidence from the Philippines that children's performance in school is enhanced by better nutritional status.

Similar evidence indicates that specific diseases contribute to worse school outcomes. One report finds that in Paraguay and Sri Lanka, reducing the prevalence of malaria by 10 percentage points would increase years of schooling by 0.1 years and raise the probability of being literate by 1–2%. Corroborating results were reported elsewhere in Latin America.

Finally, beyond these impacts of the health of the child on schooling and cognition, there is also evidence that the health of the parents may impact a child's human capital accumulation, particularly through illness of mothers and fathers contributing to early withdrawal from school. For example, one study reports that the death of a 15-year-old child's mother raises the probability of the child dropping out of school within 3 years by 15.8% points. Similarly, the death of a 15-year-old's father raises the probability of dropout by 18.7% points. Illness among parents also has a large impact on the likelihood of dropout. Among 15-year-olds, for example, a child is 13.1 percentage points more likely to drop-out if her father has a prolonged illness that interferes with work and other normal activities. The comparable number for the mother is 14.8 percentage points.

Life Course and Intergenerational Issues

Poor health and nutrition will not only limit a worker's productivity and earnings, but, as discussed in the Section Microeconomic Evidence, will also contribute to a cycle of poverty, poor health, and poor human capital outcomes across generations. Of particular concern is the evidence that traumas *in utero* or in early childhood, such as exposure to toxins (including alcohol and tobacco), or nutrient deficiencies of folate or iodine will contribute to permanent dysfunction over the entire life course. The Barker hypothesis (Barker *et al.*, 2005) argues that nutritional and other stresses to the fetus contribute to imprinting on the genes and metabolic changes, which in turn contribute to heightened risks of obesity, diabetes, heart disease, and other chronic disease later in life.

There is a growing body of evidence in support of the fetal origins of disease theory. One study suggests that reduced infections during childhood contributed dramatically to adult height and longevity due to lower levels of inflammation. Another observes the importance of the year in which children were born in the business cycle in the Netherlands in the nineteenth- and early twentieth-centuries on mortality rates. Likewise, work using data from the US observes longstanding and major consequences for the schooling, productivity, and health of the offspring of mothers afflicted during the flu pandemic. Finally, there exists compelling evidence that raising birth weights will contribute to better labor market outcomes, particularly among low birth weight babies.

One particularly informative set of studies was conducted in Guatemala, where a randomized experiment of children who had been enrolled in an early childhood nutrition supplementation program were followed as young adults. Adult men, who had received the protein-rich nutrition supplement as children, were found to have hourly earnings that were US$0.67 greater than the control group that had received a drink containing no protein. This represented a 46% higher average wage rate. One important finding, however, is that the largest impact of the supplementation resulted from treatment during the first 2 years of life; there was no impact of receiving the supplement from ages 36 to 72 months. These strong positive effects were not observed for women. These results were consistent with other follow-up studies of this cohort that found increased schooling and cognition among the treated.

Another study, using panel data from Brazil, Guatemala, India, the Philippines, and South Africa, addressed the question of the relative importance of low birth weight, and weight gain in the first 2 years of life and between the ages of 2 and 4 years on schooling outcomes. Although not able to examine causality, the researchers report that from 0 to 24 months of age, weight gain had a particularly important impact on schooling, whereas there was no significant effect between 2 and 4 years of life. To get a sense of the magnitude of the effects, their comparative statistics suggested that children whose growth was stunted at 2 years or age were likely to have completed nearly 1 year less of schooling and suffer from a 16% increased risk of failing at least one grade. The authors' calculations, based on returns to schooling in the population, indicate that child stunting during the first 2 years of life would be associated with a reduction in lifetime income of approximately 10% in the countries studied.

Another avenue through which poor health has implications over the life course arises from the expectations for a short life span, which will in turn reduce savings and thus investment in physical capital. Related to the accumulation of physical capital is the fact that disease and early mortality among the children themselves have adverse intertemporal effects. Illness and malnutrition among children reduces the incentives for parents to invest in their education. The difficulties of identifying the impact of health on investing in human capital, and specifically distinguishing those effects from how health may directly effect human capital through other channels, such as the impact on school attendance and ability of children to learn, has resulted in few studies that causally show this relationship. One notable exception is a study from Sri Lanka that shows that the decline in maternal mortality risk by 4.1% resulted in a 2.5% increase in female literacy. The elasticity of human capital with respect to life expectancy was thus calculated as between 0.6 and 1.0.

Finally, one manifestation of poor maternal health and inadequacy of health care for women is the prospect of unwanted pregnancy. A woman's lack of control over her fertility has long-term impacts on members of her household and the accumulation of human capital of her children.

Inequalities in Health

There is a potential relationship between inequalities in health and various socioeconomic outcomes. It has been proposed that inequality in health contributes to a lack of social cohesion. There is evidence that relative deprivation contributes to stress, as well as other outcomes such as loss of dignity, shame, and stigmatization, which may have effects on labor market opportunities and incomes. High inequality also lowers the likelihood that social networks and mutual assistance relationships will mitigate the deleterious effects of health shocks that compromise health status directly. Inequalities in health (and other dimensions) may also contribute to differences in preferences and thus reduce political support for investments in public goods. Health inequality, therefore, may partially explain why public institutions are both inefficient and fail to protect and promote the needs of those in greatest need. Furthermore, where inequality reduces trust and increases crime and violence, or where low social status makes people feel disrespected, it may generate violence or, at the very least, add to the political tensions contributing to disproportionate shares of budgets and state and private resources being allocated to political repression, internal security, and other nonproductive spending.

Yet another aspect by which health inequalities can slow economic growth is that there are decreasing returns in terms of productivity to health: populations with more health inequality will have lower productivity on average.

In regressing GDP, both in levels and growth rates, on health inequality and a range of other covariates, it has been found that the reduction in health inequality caused by a reduction in the number of children who die before the age of 5 of approximately 4.25 per 1000 children per year born to

mothers with a low education level leads to an almost 8% increase in GDP per capita after a period of 10 years. Although this study is relatively unique and plagued by several serious methodological and data limitations, it does add to the limited empirical evidence on the relationship between health inequality and growth, and will hopefully motivate further research in this area.

Conclusions

There is considerable theoretical and empirical evidence that points to large productivity increases and economic gains from improved health and nutrition over the life course and across generations in developing countries. These are mediated by a wide range of pathways, including increases in strength and stamina, impacts on schooling (age of entry, duration, and attendance) and cognitive abilities, reduced fertility, and increased savings associated with reduced expenditures on health and greater incentives to invest in children who are expected to live longer and be more economically productive over their life course.

Information asymmetries, as well as market failures, such as for credit and insurance, will likely contribute to underinvestment in health-related human capital. This adds further justification for government policies to address these market failures. Additionally, there are likely to be large economic externalities associated with improving health that even further support government investments in the health sector. Thus, the large efficiency gains from investment in health find further justification in the likelihood that social rates exceed private rates of return.

Reference

Barker, D. J. P., Osmond, C., Forsén, T. J., Kajantie, E. and Eriksson, J. G. (2005). Trajectories of growth among children who later develop coronary heart disease or its risk factors. *New England Journal of Medicine* **353**(17), 1802–1809.

Further Reading

Acemoglu, D. and Johnson, S. (2007). Disease and development: The effect of life expectancy on economic growth. *Journal of Political Economy* **115**(6), 925–985.

Ahuja, A., Wendell, B. and Werker, E. (2009). Male circumcision and AIDS: The macroeconomic impact of a health crisis. *Harvard Business School Working Paper 07-025*. Boston: Harvard Business School

Alderman, H., Behrman, J. R., Lavy, V. and Menon, R. (2001). Child health and school enrollment: A longitudinal analysis. *Journal of Human Resources* **36**(1), 185–205.

Alderman, H., Hoddinott, J. and Kinsey, B. H. (2006). Long-term consequences of early childhood malnutrition. *Oxford Economic Papers* **58**(3), 450–474.

Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of *in utero* influenza exposure in the post-1940 U.S. Population. *Journal of Political Economy* **114**(4), 672–712.

Arndt, C. and Lewis, J. D. (2000). The macro-implications of HIV/AIDS in South Africa: A preliminary assessment. *South African Journal of Economics* **68**(5), 380–384.

Ashraf, Q. H., Ashley, L. and Weil, D. N. (2008). When does improving health raise GDP? In Acemoglu, D., Rogoff, K. and Woodford, M. (eds.) *NBER macroeconomics annual 2008*, vol. 23, pp. 157–204. Cambridge, MA: National Bureau of Economic Research, Inc.

Banerjee, A., Iyer, L. and Somanathan, R. (2008). Public action for public goods. In Schultz, T. P. and Strauss, J. A. (eds.) *Handbook of development economics*, vol. 4, pp. 3117–3154. Amsterdam: Elsevier.

Barker, D. J. P. (1994). *Mothers, babies, and disease in later life*. London: BMJ Publishing.

Barnett, T., Tumushabe, J., Bantebya, G. B., et al. (1995). The social and economic impact of HIV/AIDS on farming systems and livelihoods in rural Africa: Some experience and lessons from Uganda, Tanzania, and Zambia. *Journal of International Development* **7**(1), 163–176.

Barro, R. (1997). *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Basta, S., Soekirman, K. and Scrimshaw, N. (1979). Iron deficiency anemia and productivity of adult males in Indonesia. *American Journal of Clinical Nutrition* **32**(4), 916–925.

Beegle, K. (2003). Labor effects of adult mortality in Tanzanian households. *World Bank Policy Research Working Paper 3062*. Washington, DC: World Bank.

Beegle, K., De Weerdt, J. and Dercon, S. (2005). *Orphanhood and the long-run impact on children*. Washington, DC, and Oxford: World Bank Economic Development Institute and Oxford University. Mimeo.

Beegle, K., Goldstein, M. and Thirumurthy, H. (2010). Microeconomic perspectives on the impacts of HIV/AIDS. In Sahn, D. E. (ed.) *The socioeconomic dimensions of HIV/AIDS in Africa: Challenges, opportunities and misconceptions*, pp. 57–73. Ithaca: Cornell University Press.

Behrman, J. and Deolalikar, A. (1988). Health and nutrition. In Chenery, H. and Srinivasan, T. N. (eds.) *Handbook of development economics*, vol. 1, pp. 631–712. Amsterdam: Elsevier.

Behrman, J. R. and Rosenzweig, M. R. (2004). Returns to birthweight. *Review of Economics and Statistics* **86**(2), 586–601.

Bell, C., Devarajan, S. and Gersbach, H. (2006). The long-run economic costs of AIDS: Theory and an application to South Africa. *World Bank Economic Review* **20**(1), 55–89.

Bleakley, H. (2010). Malaria in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics* **2**(2), 1–45.

Bliss, C. and Stern, N. (1978a). Production, wages, nutrition: Part I: The theory. *Journal of Development Economics* **5**(4), 331–362.

Bloom, D. E., Canning, D., Mansfield, R. K. and Moore, M. (2007). Demographic change, social security systems, and savings. *Journal of Monetary Economics* **54**(1), 92–114.

Bloom, D. E., Canning, D. and Sevilla, J. (2004). The effect of health on economic growth: A production function approach. *World Development* **32**(1), 1–13.

Bongaarts, J. and Bulatao, R. A. (1999). Completing the demographic transition. *Population and Development Review* **25**(3), 515–529.

Bruner, E. and Marmot, M. (1999). Social organization, stress, and health. In Marmot, M. and Wilkinson, R. G. (eds.) *Social determinants of health*, pp. 17–43. Oxford: Oxford University Press.

Case, A. and Ardington, C. (2005). The impact of paternal death on school enrollment and achievement: Longitudinal evidence from South Africa. Paper presented at the International Union of the Scientific Study of Population *Seminar on Interactions between Poverty and HIV/AIDS*, Cape Town, South Africa. Rondebosch: Centre for Social Science Research, University of Cape Town.

Crimmins, E. M. and Finch, C. E. (2006). Infection, inflammation, height, and longevity. *Proceedings of the National Academy of Sciences of the USA* **103**(2), 498–503.

Deolalikar, A. B. (1988). Nutrition and labour productivity in agriculture: Estimates for rural South India. *Review of Economics and Statistics* **70**(3), 406–413.

Dow, W., Gertler, P., Schoeni, R., Strauss, J. and Thomas, D. (1997). Health care prices, health, and labor outcomes: Experimental evidence. *Labor and Population Working Paper 97-01*. Santa Monica, CA: RAND.

Edgerton, V. R., Gardne, G., Ohira, Y., Gunawardena, K. A. and Senewiratne, B. (1979). Iron-deficiency anemia and its effect on worker productivity and activity patterns. *British Medical Journal* **2**(6204), 1546–1549.

Evans, D. K. and Miguel, E. (2007). Orphans and schooling in Africa: A longitudinal analysis. *Demography* **44**(1), 35–57.

Fogel, R. W. (1994). Economic growth, population theory, and physiology: The bearing of the long-term processes on making of economic policy. *American Economic Review* **84**(3), 369–395.

Fogel, R. W. (2004). Health, nutrition, and economic growth. *Economic Development and Cultural Change* **52**(3), 643–658.

Galor, O. and Weil, D. N. (1999). From Malthusian stagnation to modern growth. *American Economic Review* **89**, 150–154.

Glewwe, P. and Jacoby, H. G. (1995). An economic analysis of delayed primary school enrollment in a low-income country: The role of early childhood nutrition. *Review of Economics and Statistics* **77**(1), 156–169.

Glewwe, P., Jacoby, H. G. and King, E. (2001). Early childhood nutrition and academic achievement: A longitudinal analysis. *Journal of Public Economics* **81**(3), 345–368.

Glick, P. (2007). Reproductive health and behavior, HIV/AIDS, and poverty in Africa. *Cornell Food and Nutrition Policy Program Working Paper No. 219*. Ithaca, NY: Cornell University.

Glick, P. and Sahn, D. E. (1997). Gender and education impacts on employment and earnings in West Africa: Evidence from Guinea. *Economic Development and Cultural Change* **45**(4), 793–823.

Glick, P., Sahn, D. E. and Walker, T. F. (2011). Household shocks and education investment in Madagascar. *Cornell Food and Nutrition Policy Program Working Paper #240*. Ithaca, NY: Cornell University.

Godfrey, K. M. and Barker, D. J. P. (2000). Fetal nutrition and adult disease. *American Journal of Clinical Nutrition* **71**(5), 1344S–1352S.

Grimm, M. (2011). Does inequality in health impede economic growth? *Oxford Economic Papers* **63**(3), 448–474.

Haas, J. D. and Brownlie, IV, T. (2001). Iron deficiency and reduced work capacity: A critical review of the research to determine a causal relationship. *Journal of Nutrition* **131**(supplement), 676S–690S.

Haddad, L. J. and Bouis, H. E. (1991). The impact of nutritional status on agricultural productivity: Wage evidence from the Philippines. *Oxford Bulletin of Economics and Statistics* **53**(1), 45–68.

Hoddinott, J., Maluccio, J. A., Behrman, J. R., Flores, R. and Martorell, R. (2008). Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan Adults. *Lancet* **371**(9610), 411–416.

Hosegood, V., McGrath, N., Herbst, K. and Timæus, I. M. (2004). The impact of adult mortality on household dissolution and migration in rural South Africa. *AIDS* **18**(11), 1585–1590.

Kirigia, J. M., Sambo, H. B., Sambo, L. G. and Barry, S. P. (2009). Economic burden of diabetes mellitus in the WHO African region. *BMC International Health and Human Rights* **9**, 6.

Lee, R. (2003). The demographic transition: Three centuries of fundamental change. *Journal of Economic Perspectives* **17**(4), 167–190.

Leigh, A., Jencks, C. and Smeeding, T. M. (2009). Health and economic inequality. In Salverda, W., Nolan, B. and Smeeding, T. M. (eds.) *The oxford handbook of economic inequality*, pp. 384–405. Oxford: Oxford University Press.

Li, R., Chen, X., Yan, H., et al. (1994). Functional consequences of iron supplementation in iron-deficient female cotton workers in Beijing, China. *American Journal of Clinical Nutrition* **59**(4), 908–913.

Lucas, A. M. (2010). Malaria eradication and educational attainment: Evidence from Paraguay and Sri Lanka. *American Economic Journal: Applied Economics* **2**, 46–71.

Margo, R. A. and Steckel, R. H. (1982). The height of American slaves: New evidence on slave nutrition and health. *Social Science History* **6**(4), 516–538.

Martorell, R., Horta, B. L., Adair, L. S., et al. (2010). Weight gain in the first two years of life is an important predictor of schooling outcomes in pooled analyses from five birth cohorts from low- and middle-income countries. *Journal of Nutrition* **140**(2), 348–354.

McDonald, S. and Roberts, J. (2006). AIDS and economic growth: A human capital approach. *Journal of Development Economics* **80**(1), 228–250.

Meyerhoefer, C. and Sahn, D. E. (2010). The relationship between poverty and maternal morbidity and mortality in Sub-Saharan Africa. In Ajakaiye, O. and Mwabu, G. (eds.) *Reproductive health, economic growth and poverty reduction in Africa: Frameworks of analysis*. Nairobi, Kenya: University of Nairobi Press.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**(1), 159–217.

Murrugarra, E. and Valdivia, M. (2000). The returns to health for Peruvian urban adults by gender, age, and across the wage distribution. In Savedoff, W. D. and Schultz, T. P. (eds.) *Wealth from health: Linking social investments to earnings in Latin America*, pp. 151–188. Washington, DC: Inter-American Development Bank.

Mushati, P., Gregson, S., Mlilo, M., Lewis, J. and Zvidzai, C. (2003). Adult mortality and the economic sustainability of households in towns, estates, and villages in AIDS-affected eastern Zimbabwe. Paper presented at the Scientific Meeting on Empirical Evidence for the Demographic and Socio-Economic Impact of AIDS, Durban, South Africa. Washington, DC: International Food Policy Research Institute.

Over, M. (2010). Prevention failure: The ballooning entitlement burden of U.S. global AIDS treatment spending and what to do about it. In Sahn, D. E. (ed.) *The Socioeconomic dimensions of HIV/AIDS in Africa: Challenges, opportunities and misconceptions*, pp. 186–230. Ithaca: Cornell University Press.

Pitt, M. and Rosenzweig, M. R. (1986). Agricultural prices, food consumption, and the health and productivity of Indonesian farmers. In Singh, I., Squire, L. and Strauss, J. (eds.) *Agricultural household models: extensions, applications, and policy*, pp. 335. Baltimore, MD: Johns Hopkins University Press.

Pollitt, E. (2001). The developmental and probabilistic nature of the functional consequences of iron-deficiency anemia in children. *Journal of Nutrition* **131**(2), 669S–675S.

Sachs, J. D. (2003). Institutions don't rule: Direct effects of geography on per capita income. *NBER Working Paper 9490*. Cambridge, MA: National Bureau of Economic Research.

Sahn, D. E. (ed.) (2010). *The socioeconomic dimensions of HIV/AIDS in Africa: Challenges, opportunities, and misconceptions*. Ithaca, NY: Cornell University.

Sahn, D. E. and Alderman, H. (1988). The effects of human capital on wages, and the determinants of labor supply in a developing country. *Journal of Development Economics* **29**(2), 157–183.

Schultz, T. P. (2008). Population policies, fertility, women's human capital and child quality. In Schultz, T. P. and Strauss, J. (eds.) *Handbook of development economics*, vol. 4, pp. 3249–3304. Amsterdam: Elsevier.

Schultz, T. P. and Tansel, A. (1997). Wage and labor supply effects of illness in Côte d'Ivoire and Ghana: Instrumental variable estimates for days disabled. *Journal of Development Economics* **53**(2), 251–286.

Sen, A. (1985). *Commodities and capabilities*. Amsterdam: North-Holland.

Strauss, J. and Thomas, D. (1998). Health, nutrition, and economic development. *Journal of Economic Literature* **36**(2), 766–817.

Strauss, J. and Thomas, D. (2008). Health over the life course. In Schultz, T. P. and Strauss, J. (eds.) *Handbook of development economics*, vol. 4, pp. 3375–3474. Amsterdam: Elsevier.

Thirumurthy, H., Graff Zivin, J. and Goldstein, M. (2008). The economic impact of AIDS treatment: Labor supply in western Kenya. *Journal of Human Resources* **43**(3), 511–552.

Thomas, D., Frankenberg, E., Friedman, J., et al. (2006). Causal effect of health on labor market outcomes: Experimental evidence. *California Center for Population Research On-Line Working Paper Series CCPR-070-06*. Los Angeles: University of California.

Thomas, D. and Strauss, J. (1997). Health and wages: Evidence on men and women in urban Brazil. *Journal of Econometrics* **77**(1), 159–185.

Van den Berg, G. J., Lindeboom, M. and Portrait, F. (2006). Economic conditions early in life and individual mortality. *American Economic Review* **96**(1), 290–302.

Weil, D. N. (2007). Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* **122**(3), 1265–1306.

Wilkinson, R. G. (1996). *Unhealthy societies: The afflictions of inequality*. London: Routledge.

Wilkinson, R. G. (2000). The need for an interdisciplinary perspective on the social determinants of health. *Health Economics* **9**(7), 581–583.

Wolgemuth, J. C., Latham, M. C., Hall, A., Chesher, A. and Crompton, D. W. (1982). Worker productivity and the nutritional status of Kenyan road construction laborers. *American Journal of Clinical Nutrition* **36**, 68–78.

Observational Studies in Economic Evaluation

D Polsky, University of Pennsylvania, Philadelphia, PA, USA
M Baiocchi, Stanford University, Stanford, CA, USA

Glossary

Average treatment effect (ATE) The average of the individual level treatment effect over all individuals i in the population of interest. This is in contrast to other parameters of interest such as the average treatment on the treated and the complier average causal effect. Mathematically:

$$ATE = E[\Delta] = E[Y(1) - Y(0)]$$

where the expectation is taken over the population of interest.

Average treatment on the treated (ATT) The average of the individual level treatment effect over all individuals i in the population of interest who were observed in the dataset to have actually received the treatment. This is in contrast to other parameters of interest such as the average treatment effect and the complier average causal effect. Mathematically:

$$ATT = E\left[\Delta \middle| T = 1\right] = E\left[Y(1) - Y(0) \middle| T = 1\right]$$

where the expectation is taken over the population of interest, and the conditional statement constrains the effect to those in the dataset who actually received the treatment.

Complier average causal effect (CACE) The average of the individual level treatment effect over all individuals i in the population of interest who had their treatment level changed by the instrumental variable. This is in contrast to other parameters of interest such as the average treatment effect and the average treatment effect on the treated. When the treatment has more than two levels, this effect is generally known as the local average treatment effect (LATE).

Hidden bias Variables which do not appear in the analyst's dataset (unobserved), are not the treatment variable or outcome variable, are not posttreatment, and (unless appropriately controlled for) bias the effect estimate of the treatment. Methods which exploit randomization within the treatment assignment are often used to mitigate hidden bias, such methods include instrumental variables or difference-in-differences. When no randomization in treatment can be used, sensitivity analysis may be performed in order to put bounds on the impact of hidden bias on the estimation of treatment effects.

Individual level treatment effect The difference, for individual i, between the outcome under different levels of the treatment (e.g., difference in outcome for an individual under treatment vs. control). Except for contrived situations, this quantity is never directly observable because taking the treatment precludes taking the control and vice versa. Mathematically:

$$\Delta_i = Y_i(1) - Y_i(0)$$

Neonatal intensive care unit (NICU) An intensive care unit staffed with specially trained medical providers and equipped with technology meant to provide care for premature or medically complicated newborns.

Overt selection bias Variables which appear in the analyst's dataset (observed), are not the treatment variable or outcome variable, are not posttreatment, and (unless appropriately controlled for) bias the effect estimate due to sorting of individuals into different levels of the treatment. The notion of overt selection bias is often invoked by assumptions of 'strongly ignorable treatment assignment' or 'conditional independence.' Overt selection bias is often believed to be controlled for by common statistical approaches such as regression, matching, and inverse weighting; this is in contrast to hidden bias.

Strongly ignorable treatment assignment assumption Loosely, one can think of this assumption as saying: selection into the treatment is occurring only on variables which are observed. Formally, this assumption is often written as

$$(Y(0), Y(1)) \perp T \middle| X, \quad 0 < pr\left(T = 1 \middle| X\right) < 1$$

where $\perp$ denotes the conditional independence between the treatment and the joint distribution of the counterfactual outcomes. Two random variables are conditionally independent given a third variable if and only if they are independent in their conditional distribution given the conditioning variable.

Nomenclature

$\boldsymbol{\Delta_i}$ The individual level treatment effect. This is the difference, for individual i, between the outcome under different levels of the treatment (e.g., change in outcome for an individual under treatment vs. control).

$\boldsymbol{A \perp B}$ Denotes the independence random variables A and B.

$\boldsymbol{Y(T=t)}$ Denotes the distribution of the outcome, Y, when treatment is set to level $T=t$. This is from 'Potential Outcomes Framework.'

$A \perp B \middle| C$ Denotes the conditional independence random variables A and B, when B is conditioned on random variable C.

Introduction

The goal of an economic evaluation of medical interventions is to provide actionable information for policy makers. Modern policy decision makers are driven by data-backed arguments regarding what might change as a result of an intervention. As analysts, this requires specific attention to determining the causal impact between a given intervention and future outcomes. To justify a change in the way medicine is practiced, correlation is not sufficient; detecting and quantifying causal connections is necessary.

Medicine has relied on randomized controlled studies as the gold standard for detecting and quantifying causal connections between an intervention and future outcomes. Randomization offers a clear mechanism for limiting the number of alternate possible explanations for what generates the differences between the treated and control groups. The demand for causal evidence in medicine far exceeds the ability to practically control, finance, and/or conduct randomized studies. Observational data offer a sensible alternative source of data for developing evidence about the implications of different medical interventions. However, for studies using observational data to be considered as a reliable source for evidence of causal effects, great care is needed to design studies in a way that limits the number of alternative explanations for observed differences in outcomes between intervention and control. This article highlights a number of the techniques and tools used in high-quality observational studies. A few of the common pitfalls to be aware of are also discussed.

Example

The development of medical care for premature infants (preemies) has been a spectacular success for modern medicine. This care is offered within neonatal intensive care units (NICUs) of varying intensity of care. Higher intensity NICUs (those classified as various grades of level 3 by the American Academy of Pediatrics) have more sophisticated medical machinery and highly skilled doctors who specialize in the treatment of tiny preemies.

Although establishing value requires addressing questions of both costs and outcomes, the example will focus on estimating the difference in rates of death between the higher-level NICUs and the lower-level NICUs. Using data from Pennsylvania from the years 1995–2005, the authors start with a simple comparison of the 1.25% rate of death at low-level facilities to the 2.26% rate of death at high-level facilities. This higher death rate at high-level facilities is surprising only if one assumes preemies were randomly assigned to either a high or low-level NICU, regardless of how sick they were. In fact, as in most health applications, the sickest patients were routed to the highest level of intensity. As a result, one cannot necessarily attribute the variation in the outcome to variation in the treatment intensity. Fortunately, the data provide a detailed assessment of baseline severity with 45 covariates including variables such as gestational age, birth weight, congenital disorder indicators, parity, and information about the mother's socioeconomic status. Yet even with this level of detail, the data cannot characterize the full set of clinical factors that a physician or family considers when deciding whether to route a preemie to a high-intensity care unit. As can be seen, these missing attributes will cause considerable problems later on.

What is wanted is not the naïve comparison of rates of death – that is the percentage of preemies who died at the different types of NICUs – but one would like to have the difference in probabilities of death for each preemie given whether the preemie was to be delivered at a low-level facility or a high-level facility. This will be called the causal effect of treatment. This concept is formalized in Section Parameters of Interest.

The Fundamentals

The Potential Outcome Framework

The literature has made great use of the potential outcomes framework (as described in Neyman, 1990; Rubin, 1974; Holland, 1986) as a systematic, mathematical description of the cause-and-effect relationship between variables. Suppose for the moment, assume there are three variables of interest: the outcome of interest Y, the treatment variable T, and $\boldsymbol{X}$ as a vector of covariates. For most of this article, it will be assumed that there are only two treatment levels (e.g., the new intervention under consideration vs. the old intervention), though this assumption is only for simplicity's sake and treatments with more than two levels are permissible. These two levels shall be referred to using the generic terms 'treatment' and 'control,' without much discussion of what those two words mean aside from saying that they serve as contrasting interventions to one another. In the potential outcomes framework, the notion is that each individual has two possible outcomes – one which is observed if the person were to take the treatment and one if the person were to take the control. In practice one is only able to observe one of these outcomes because taking the treatment often precludes taking the control and vice versa. The notation used for subject i taking the treatment is $T_i=1$ and for patient i taking the control is $T_i=0$. To formally denote the outcome subject i would experience under the treatment and control the authors write $Y(T_i=1)$ and $Y(T_i=0)$, respectively. Informally the notation is simplified to $Y_i(1)$ and $Y_i(0)$ for the potential outcome under treatment and the potential outcome under control. This article will think of Y being a scalar, though it is possible to develop a framework where Y is a vector of outcomes.

Excellent resources exist for reading up on the potential outcomes framework: Rosenbaum (2002), Pearl (2009), and Hernan and Robins (2013).

Now there is enough mathematical language to describe the ultimate, often unattainable, quantity of interest – namely, 'the individual level treatment effect:'

$$\Delta_i = Y_i(1) - Y_i(0)$$

Thus Δ_i will tell us the difference in outcome, for subject i, between taking the treatment and control. If one could observe this quantity then the benefit from intervention would be known explicitly. But, in practice only one is observed or the other of the potential outcomes. To see this, one may write

the observed outcome, denoted Y_i^{obs} for the ith individual, as a function of the potential outcomes (Neyman, 1990; Rubin, 1974):

$$Y_i^{obs} = T_i * Y_i(1) - (1 - T_i) * Y_i(0)$$

Observing one of the potential outcomes precludes observing the other. In all but the most contrived settings, this problem is intractable. One will not be able to observe both the treatment and control outcomes. So one must turn to other parameters of interest.

Parameters of Interest

Suppose we, as the analysts, have collected characteristics of the subjects in our study. It is important to stress that these baseline characteristics should be based on the state of the subject before the intervention to avoid the potential to bias the treatment effect (see Cox, 1958, section 4.2 and Rosenbaum, 2002, pp. 73–74). For example, say a new drug is being tested for its ability to lower the risk of heart attack. High blood pressure is known to correlate with higher risk of heart attack, so it is tempting to control for this covariate. Controlling for blood pressure is likely to improve the precision of the estimate if a pretreatment blood pressure measure is used. It would be a mistake to use a posttreatment measurement of blood pressure as a control because this measurement may be affected by the drug and would thus result in an attenuated estimated causal effect. Intuitively, this is because the estimation procedure is limiting comparison in outcome not just between people who took the drug and who didn't but between people who took the drug and then had a certain level of blood pressure to people didn't take the drug and had the same level of blood pressure. The impact from the drug may have already happened via the lowering of the blood pressure.

Let us denote these measured pretreatment characteristics as X_i for the ith subject. Furthermore, the subjects are likely to have characteristics which were not recorded. Let us denote these unobserved characteristics as U_i for the ith subject. There is a not unreasonable belief that the observable outcomes can be thought of as a function of these covariates (in this notation you can think of the treatment level as being an observed covariate). That is $Y_i^{obs} = f(X_i, U_i)$. To keep things simple it will be assumed that the covariates are linearly related to the outcomes like so

$$Y_i(1) = X_i\beta(1) + U_i\alpha(1)$$

$$Y_i(0) = X_i\beta(0) + U_i\alpha(0)$$

Note that one needs to index the coefficients by the treatment level in order to account for interactions between the treatment level and the covariates. Also, it may appear strange putting coefficients on the unobserved variables, but this is required at the bare minimum to make the dimensions agree. In practice one gets a bit sloppy and write $\varepsilon_i(T)$ in place of the clunkier $U_i\alpha(T)$, but this is a move of convenience rather than discipline. It is known that there is not just one scalar, unobservable covariate that has been omitted from the dataset, so it is more realistic to write $U_i\alpha(T)$. Note that this means something a bit magical is happening when an author proposes a functional form for $\varepsilon_i(T)$.

Combining the equations for the observed outcome and the linear models, one gets a decomposition of the observed outcome in terms of covariates, both observed and unobserved, as well as the treatment.

$$Y_i^{obs} = X_i\beta(0) + T_i[(X_i\beta(1) - X_i\beta(0)) + (U_i\alpha(1) - U_i\alpha(0))] + U_i\alpha(0)$$

It is standard in econometrics to think of the above model as a regression, where the coefficient on the treatment variable comes from two sources of variation – the first source is the variation due to the observed covariates $(X_i\beta(1) - X_i\beta(0))$ and the second is the variation due to the unobserved covariates, $(U_i\alpha(1) - U_i\alpha(0))$. It is common to interpret the first source of variation as the gains for the average person with covariate levels X_i, and the second source of variation to be referred to as idiosyncratic gains for subject i. The idiosyncratic gains are the part of this model which allows persons i and j to differ in outcomes even when $X_i = X_j$.

For reasons laid out in the last subsection, one moves from estimating Δ_i for an individual and instead consider population level parameters. Historically, the quantity of interest for many studies has been the average treatment effect (ATE):

$$ATE = E[\Delta] = E[Y(1) - Y(0)]$$

Note that it is more common to use the conditional ATE:

$$ATE(X) = E[\Delta|X] = E[Y(1) - Y(0)|X]$$

This quantity $ATE\ (X)$ is interpreted as being the expected change in outcome for everyone with characteristics X if they were to go from taking the control to taking the treatment. This is a useful quantity if the researchers are considering a total replacement of a standard treatment ('control') with a new treatment ('treatment').

Another quantity of interest is the average treatment on the treated (ATT):

$$ATT = E[\Delta|T = 1] = E[Y(1) - Y(0)|T = 1]$$

Which is also more usefully thought of as a conditional quantity:

$$ATT(X) = E[\Delta|T = 1, X] = E[Y(1) - Y(0)|T = 1, X]$$

This quantity limits itself to considering the change in outcomes for only those people who actually received the treatment. The ATT is the pertinent quantity to estimate if people who received the treatment in the dataset are similar to the people who are anticipated to take the treatment in the future. Differences between the ATT and ATE often arise when a new treatment is introduced into a population. As an example, say that a new, more invasive, surgical procedure is introduced as a replacement to a technique which is less invasive, though also believed to be less efficacious. You might imagine that the new technology, because of the acute stress of the procedure, would be used on a relatively healthy subset of the population until the relative efficacy and burdens of the two procedures are well known. Sometimes the difference between the group which receives the treatment is observed and recorded in the covariates. As a simple example of this, let

us say the new treatment group has only patients who are less than 40 years old. If this is the case, to estimate the ATE one may have to extrapolate into parts of the overall population for which there are no people who were treated. The problem here is that the efficacy of the treatment may vary for different parts of the population. But if one is not careful to present the estimate as ATT (i.e., appropriate for just a subset of the population) it is quite possible that it will be interpreted as an ATE which may lead to incorrect estimates of the benefit on the entire population. This problem is made even more difficult to address when the treated group is different from the entire population in some unobserved way. See the Sections Methods to Address Selection Bias and Methods for Overt Bias and Bias Due to Omitted Variables on methods for overt bias and bias due to omitted variables for more discussion on this.

Perhaps more is made out of the ATE and ATT distinction than is really necessary. Both of them are discussed in order to make the researcher aware of potential issues of the population of interest and generalizability. Careful thought about the kinds of people who will be impacted by the proposed intervention will typically guide the researcher to the correct choice of either ATE or ATT. For a more exhaustive discussion see Imbens (2004).

Note that ATEs are discussed over populations. These are valuable quantities and are often quite useful for designing policy interventions. But one should point out that it is often plausible that there are subpopulations within the larger group that experience either bigger or smaller treatment effects. This variation in the individual unit's treatment effect is an important problem and deserves attention, but this issue will not be addressed (except for briefly in the IV and RD sections). The study of 'treatment heterogeneity' is gaining attention in the literature and this will add to the usefulness for policy interventions. The quest for 'personalized medicine' is in large part an acknowledgment that treatments often vary across subgroups within the population.

Selection Bias

One of the biggest problems with observational studies is that there is selection bias. Loosely speaking, selection bias arises from how the subjects are sorted (or sort themselves) into the treatment or control groups. The intuition here is: the treatment group was different from the control group even before the intervention, and the two groups would probably have had different outcomes even if there had been no intervention at all. Selection bias can occur in a couple of different ways, but one way to write it is

$$f(X,U|T=1) \neq f(X,U|T=0)$$

that is, the joint distribution of the covariates for those who received the treatment is different than for those who received the control. (A bit of a warning: Some confusion may arise when using a conditional statement. This confusion occurs because different academic traditions tend to think of the conditional statement in slightly different ways. For example, a statistician may read a statement like $f(X,U|T=1)$ to mean roughly 'the joint distribution of X and U when the researcher intervenes and sets T level 1.' This is in contrast to the way an econometrician may read the statement as, roughly, 'the joint distribution of X and U limited to those units of observation which were observed to have $T=1$.' For a more detailed discussion of what is being asserted in the conditional statement we suggest reading about the 'do operator' introduced by Judea Pearl. It is a quite enlightening discussion. See Pearl, 2009 for details.) If this is true, that there is selection bias, then

$$E[Y(1)-Y(0)|X] \neq E[Y(1)|X,T=1] - E[Y(0)|X,T=0]$$

This is problematic because the left-hand side of this equation is the unobservable quantity of interest but the right-hand side is made up of directly observable quantities. But it seems like the above equation is used in other settings, namely, experimentation. Why is that acceptable?

In an experiment, because of randomization it is known that

$$(X,U) \perp T$$

And it follows that

$$E[Y(1)-Y(0)|X] = E[Y(1)|X,T=1] - E[Y(0)|X,T=0]$$

Though it is often a dubious claim, many of the standard techniques require an assumption which essentially says that the only selection between treated and control groups is on levels of the observed covariates (X). This is sometimes referred to as overt selection bias. Typically, if overt selection bias is the only form of bias then either conditioning on observed covariates (e.g., by using a regression) or matching is enough to address overt bias. One particular assumption that is invoked quite often in the current health literature is the absence of omitted variables (i.e., only overt bias) and this is used to justify recovering ATT and ATE in some settings. Overt bias will be covered in the propensity score section.

Hidden bias exists when there are imbalances in the unobserved covariates. Let us use the observed outcome formula again, rewriting it like so:

$$Y_i^{obs} = X_i\beta(0) + T_iE[\Delta|X] + U_i\alpha(0) + T_i(U_i\alpha(1) - U_i\alpha(0))$$

A least-squares model of Y on T based on the model above will tend to produce biased estimates for $E[\Delta|X]$ when T is correlated with either $U_i\alpha(0)$ or $(U_i\alpha(1) - U_i\alpha(0))$. This can arise from unobserved covariates which influence both outcome under both treatment and control and selection into treatment. The resulting bias, referred to as hidden bias, is given by

$$E[U_i\alpha(1)|X,T=1] - E[U_i\alpha(0)|X,T=0] \neq 0$$

Methods to Address Selection Bias

In a randomized experiment setting, inference on the causal effect of treatment on the outcome requires no further assumption than the method for randomizing subjects into the treatment or control (Fisher, 1949). The randomization guarantees independence of assigned treatment from the covariates. And one should pause to stress the point that this independence is for all covariates, both observed and unobserved. By observed covariates it is meant that those

covariates which appear in the analyst's dataset and unobserved all of those that do not. If the sample is large enough then this independence means that the treatment group will have quite similar covariate distribution as the control group. Therefore, any variation noted in the outcome is more readily attributed to the variation in the treatment level rather than variation in the covariates.

The primary challenge to observational studies is that selection into treatment is not randomly assigned. Usually there are covariates, both observed and unobserved, which determine who receives treatment and who receives control. In such a case variation in the outcome is not easily attributable to treatment levels because covariates are different between the different levels as well. There are study designs which were created to address this selection bias and we introduce these methods here. These methods will be classified as (roughly) into two groups: (1) those methods which address only the observed selection bias and (2) those methods which attempt to address selection bias on both the observed as well as unobserved covariates.

Methods Which Address Only Overt Bias

Methods which address selection bias based only on observed covariates tend to be easily implemented, but they also tend to leave the analyst open to major criticism. The assumptions required for the methods in this section are hefty. The authors hope that as the electronic medical records come in to common use; the quality and detail of the covariates available to the health policy researcher will begin to make it more believable that one has access to all of the important covariates. Better quality data is always much appreciated. Better methods can only help so much.

Model-based adjustment (e.g., linear regression)

Across much of the applied econometric literature, model-based adjustment is the most common method for addressing selection on observed covariates. The most common form of model-based adjustment is linear regression. More complex methods can be developed from maximum-likelihood models, Bayesian hierarchical methods, or other more complex methods. These methods are often designed to estimate the ATE under a powerful assumption. Loosely, one can think of this assumption as saying: selection into the treatment is occurring systematically only on variables which are observed. Formally, this assumption is often written as

$$(Y(0),Y(1)) \perp T|X, \quad 0<pr(T=1|X)<1$$

where $\perp$ denotes the conditional independence between the treatment and the joint distribution of the counterfactual outcomes. Two random variables are conditionally independent given a third variable if and only if they are independent in their conditional distribution given the conditioning variable. The above assumption, essentially saying one has all the covariates one needs, has a few different names: strongly ignorable treatment assignment (Rosenbaum and Rubin, 1983), selection on observables (Heckman and Robb, 1985), conditional independence, no confounders (in the epidemiology literature), or 'overt bias' and 'the absence of omitted variable bias' (in the econometrics literature). This article tries to use the 'overt bias'/'absence of omitted variable bias' labels consistently, but please feel free to mentally replace those terms with your favorite.

Propensity score methods

Propensity score methods, propensity score matching in particular, have been of particular interest in the health policy literature. It is speculated that the reason for this interest is related to the fact that with propensity score methods it is possible to mimic the feel of the familiar and salient randomized controlled experiment. The propensity score is defined as:

$$e(x)=pr(T=1|X)$$

Rosenbaum and Rubin (1983) showed that assuming there is only overt bias, conditioning on the propensity score will lead to independence of the treatment and the potential outcomes. That is,

$$(Y(0),Y(1)) \perp T|e(x)$$

This conditioning statement is quite useful; it justifies matching techniques and inverse weighting techniques (see Sections Propensity score matching and Inverse probability weighting). But what is the difference between conditioning on the propensity score and the assumption that there is only overt bias? The assumption, as laid out in the formula above, seems to require that there are two people with exactly the same values of all of their covariates before treatment selection is independent of the potential outcomes. That would mean that if even one covariate value was different between two subjects then one would be concerned that treatment was still being confounded by the covariates, and the treatment estimate would be biased. Finding two identical units of observation is quite difficult (but not impossible, see difference-in-differences and before-and-after study designs discussed in Section Before-and-after (difference-in-differences)). The usefulness of the propensity score is that, assuming only overt bias, it shows that one does not need to find identical units merely needed to find units with the same probability of treatment assignment. This makes the analyst's job easier, because one only requires agreement on the propensity score in order to get a valid estimate of the causal effect. And matching on a one-dimensional feature is more easily accomplished than a high-dimensional vector. Said another way, the propensity score (a scalar) contains all of the requisite information contained in the covariates (often quite high dimensional).

It should be noted that exact matching on the propensity score (i.e., matching individuals i and j such that $e(\boldsymbol{x}_i)=e(\boldsymbol{x}_j)$) does not lead to exact matching on the covariates even in infinite samples. This means that the covariates within a pair match are likely to never be exactly the same (i.e., $x_i \neq x_j$). But what matching on the propensity score does guarantee (asymptotically) is that $f(X|T=1)=f(X|T=0)$. Is this a problem? Having two units of observation which are identical on all of their observed covariates, $\boldsymbol{x}_i=\boldsymbol{x}_j$, would be ideal because one could give unit i the treatment and unit j the control and any variation observed in the outcome would be plausibly

(we only say 'plausibly' because it is still quite likely that there are differences on the unobserved covariates, $\boldsymbol{u}_i \neq \boldsymbol{u}_j$) attributed to the variation in treatment. But that is not even what happens in a clinical trial. In a clinical trial one may match on important variables, but it is the randomization on which the inference is built. In a randomized trial the joint distribution of the covariates is similarly guaranteed to be equivalent, but the inference requires only that the randomization to be understood. For a discussion of both the relative benefits of exact matching (a.k.a. reducing heterogeneity) and propensity score match (i.e., randomization-based inference), as well as the intellectual history behind these two drives in research, refer to Rosenbaum (2005).

Typically the propensity score is estimated using a logistic regression, though this is not required. Conceivably, other techniques which estimate the probability of treatment given a unit's covariates would be valid.

Once the propensity score has been estimated there are several ways that it can be used to adjust for selection bias. A very simple technique is to enter the propensity score in the regression estimating the relationship between the treatment and the outcome. In this simple example, the propensity score acts very much like the model-based adjustment methods described in Section Model-based adjustment (e.g., linear regression). (The analyst should be aware that there are a few hefty assumptions required to use the propensity score in the regression framework.) There are other techniques for implementing propensity scores such as propensity score matching and inverse probability weighting. These more sophisticated techniques have distinct advantages when the treatment effect is not equivalent across the entire distribution of the X's because these methods focus the estimation on the part of the distribution where there is a substantial probability that either treatments might be selected.

Propensity score matching

Once the propensity score is estimated, $\hat{e}(x)$, then units are ideally matched to each other so that treated units with a particular value of $\hat{e}(x)$ are matched to control units with the same value of the propensity score. It is most common to do pair matching (one treated unit matched to one control unit) though it is also possible to match more than one treated unit to one control, or more than one control unit to one treated unit. (see Hansen, 2004, for more on full matching and K to 1 matching.)

Propensity score matching is often thought of as attempting to replicate a randomized controlled experiment. As such, once matching has taken place it is common to assess the covariate balance between the control and treated units. This is often done using the means of the covariates (citation). Once a properly balanced study design has been achieved, something as simple as a paired t-test is often run to estimate the treatment effect.

Matching also has the benefit of forcing the analyst to be aware of covariate overlap, or lack thereof. In many applications the treatment group and the control groups have different values of the covariates. For example, the control group may have people who are younger than the treatment group – say the youngest person in the treatment group is 50 but half of the control group is less than 30. This is important because part of the assumption of only overt bias, $0 < pr(T = 1|X) < 1$, requires there are units at all covariate levels which take on treatment and control. Model-based approaches leave the unaware-analyst at a disadvantage because it is not routine to check for covariate overlap between the treated and control groups. Nonoverlap is a significant violation of a fundamental assumption.

One of the more famous applications of propensity score matching was a study of right heart catheterization – see Connors *et al.* (1996).

Inverse probability weighting

Inverse probability weighting (IPW) takes each unit and weights it by the inverse of the propensity score. That is a weight, suppose we choose $w = \frac{1}{e(x)}$, is assigned to each unit of observation. If a unit has a particularly low probability of treatment then the weight will take on a very large value. Once all the observations have been weighted as such, the treatment is independent of the potential outcomes and inference on the treatment effect is trivial. When there are extremely low values of the propensity score (i.e., certain covariate values are strongly associated with selection into the control), the standard errors associated with IPW inference can get quite large. This problem is not unique to IPW methods, matching and regression face similar challenges. There are a number of different estimators based on IPW.

Although uncommon in some health policy settings, inverse probability weighting has been applied to great advantage in epidemiology. Many of the statistical techniques of analysis with IPW can be ported over from the survey sampling literature where sampling weights are heavily used.

Combining propensity score methods and covariate adjustment

Rubin (1973) used simulation studies to examine the tradeoffs between model-based approaches and matching-based approaches. Models tended to be more statistically efficient than matching-based techniques, with the significant caveat that this was true with the model was correctly specified (i.e., that the proposed model was exactly the right model for the process which actually generated the data). In fact, if the proposed model is incorrect, then model-based methods may actually exacerbate the bias. Matching-based methods were shown to be fairly consistent in reducing overt bias. The study concluded that a combination of the two methods produced estimates which were both robust and efficient. A diligent analyst, with strong justification for a specific model, may first match to ensure covariate overlap between the treated and control and then run the model-based inference on the subjects which were part of the matching.

Methods for Overt Bias and Bias Due to Omitted Variables

Regression, propensity score matching and any methods predicated on only overt bias do not address selection on unobserved covariates. It is important to be aware of this because a well-informed researcher needs to judge if available covariates are enough to make a compelling argument for the absence of omitted variables. This is often a dubious claim

because (1) a clever reviewer will usually find several variables missing from your dataset and/or (2) there are 'intangible' variables that are difficult, or perhaps inconceivable, to measure. The following study designs are presented below in order to help you address these situations.

It is important to note that none of the designs below come 'for free,' that is without some hefty assumptions. It is important to consider these assumptions carefully before proceeding.

Instrumental variables

An instrumental variable (IV) design takes advantage of some randomness which is occurring in the treatment assignment to help address imbalances in the unobserved variables. IV methods go beyond simple methods (like propensity score or multivariate regression) which are only designed to address imbalances in observed covariates.

An instrument is a haphazard nudge toward acceptance of a treatment that affects outcomes only to the extent that it affects acceptance of the treatment. In settings in which treatment assignment is mostly deliberate and not random, there may nevertheless exist some essentially random nudges to accept treatment, so that use of an instrument might extract bits of random treatment assignment from a setting that is otherwise quite biased in its treatment assignments. Holland (1986) offers an intuitive introduction to how an ideal IV would work. Angrist *et al.* (1996) used the potential outcomes framework to bring greater clarity to the math of IV.

This intuition for IV discussed above enhances the classic econometric presentation of IVs where the focus is on correlation with the error term. To introduce this more formally, the authors will introduce the 'complier terminology' from Angrist *et al.* (1996).

Notation first: Z is used to denote the instrument. If these random variables have subscripts one is referring to an individual's values. If the random variables are in bold then one is referring to the vector of values for all observations in our dataset. This section will assume that the treatment is binary (i.e., $T_i=1$ if the ith unit takes the treatment and $T_i=0$ otherwise) and that the instrument is binary (i.e., $Z_i=1$ if the ith unit is encouraged to take the treatment and $Z_i=0$ otherwise). The notation $T_i(z=1)$ is used to denote the treatment that the ith unit actually receives if encouraged, $z=1$, to take the treatment.

The story goes, the instrument either encourages the unit to receive the treatment ($Z_i=1$) or not ($Z_i=0$). The unit is then allowed to either comply with that encouragement or not. Because both the treatment and the instrument are assumed to be binary, it follows that there are four compliance classes. Using counterfactuals, the authors label these compliance classes like so:

1. Always takers: $T_i(Z_i=1)=T_i(Z_i=0)=1$
2. Compliers: $T_i(Z_i=1)=1, T_i(Z_i=0)=0$
3. Never takers: $T_i(Z_i=1)=T_i(Z_i=0)=0$
4. Defiers: $T_i(Z_i=1)=0, T_i(Z_i=0)=1$

Under any possible random assignment of the instrument one will never be able to discern the treatment effect for the always-takers nor the never-takers because no matter what one will never be able to observe the counterfactual treatment assignment. Assumption 4 (monotonicity) says that the defiers do not exist. Thus one is only able to estimate the treatment effect for the compliers, those who are randomly assigned by the instrument. This estimand is often referred to as the local average treatment effect (LATE) because it is only true for a subpopulation (a 'local' group). It has also been referred to as the complier average causal effect (CACE). CACE is a special case of LATE; CACE is often used when the treatment and instrument are binary. LATE is more broadly defined. A more fundamental estimand, the local instrumental variable (LIV), can be derived from the use of an instrument. The LIV is capable, assuming the proper specification of the model and proper weighting of population covariates, of estimating the ATE and TT (Heckman and Vytlacil, 1999, 2000). Thus the LIV is a useful tool which allows the analyst to shift between estimating effects on different parts of a populations.

An instrument is weak if the random nudges barely influence treatment assignment or strong if the nudges are often decisive in influencing treatment assignment. Another way to think of a 'strong' versus 'weak' instrument is to think of the percentage of compliers. A strong instrument will induce higher rates of compliance. A study with a weaker instrument will have a lower percentage of compliers. Although ideally an ostensibly random instrument is perfectly random and not biased, it is not possible to be certain of this; thus a typical concern is that even the instrument might be biased to some degree. It is known from theoretical arguments that weak instruments are invariably sensitive to extremely small biases – Bound *et al.* (1995); for this reason, strong instruments are preferred.

The most common method for implementing IV is two-stage least squares (2SLS). 2SLS is valid when the outcome of interest is continuous, and all of the typical model requirements for least squares are met. If the instrument is binary and the outcome is linear, then the 2SLS estimate is the Wald estimator (Angrist, 1991). If the outcome of interest is something other than nonlinear then there are a couple of other methods available. The two methods the authors cite are both rather new to the literature and are only beginning to work their way into use. Two-stage residual inclusion (2SRI) method is a parametric method for dealing with nonlinear outcomes (Terza *et al.*, 2008). Near/far matching is a nonparametric method that attempts to replicate the structure of a randomized controlled experiment (Baiocchi *et al.*, 2010). Near/far matching may feel a bit similar to propensity score matching, with the addition feature of taking into account the randomness from the instrument.

As laid out in Angrist *et al.* (1996), there are five assumptions for IV when you have a binary instrument and a binary treatment. This is a bit surprising to some folks because we typically only discuss two assumptions. The two assumptions from econometrics are broken apart into assumptions 1, 2, and 3 below. Assumptions 1 and 2 are often combined. Assumptions 4 and 5 are often overlooked in the literature. All of these assumptions are important and thus need to be justified before an IV analysis is to be taken seriously.

(1) Uniform Random Assignment

$$\Pr(Z=z)=\Pr(Z=z')$$

for all possible treatment assignments z and z' such that $1^T z = 1^T z'$, where 1 is the N-dimensional column vector with all elements equal to one.

This assumption guarantees that the instrument (Z) is randomly assigned, it says nothing directly about the treatment actually received. This assumption can be restated such that the probabilities are conditional on the observed covariates. See the section on 'Instrumental Variables – Complier Terminology.'

(2) No direct effect of the instrument on the outcome (Angrist *et al.* (1996) refers to this as this assumption as the 'exclusion restriction,' which may be a bit confusing given the use of this term in the econometrics literature to refer to both assumptions 1 and 2 in the Angrist *et al.* (1996) framework.)

$$Y(Z,T) = Y(Z',T) \quad \text{for all } Z,Z' \text{ and for all } T$$

Intuitively, this assumption says that the instrument has no impact on the outcome except through the instrument's influence on which treatment the unit actually receives. There are a number of ways to violate this assumption. One way this would be violated is in the study of a treatment if there is reason to believe in a 'placebo effect' – whereby merely believing in the treatment has an effect – where the unit will have a different outcome based merely on whether being assigned to take the treatment or not, rather than through the actual treatment taken. This assumption is quite important and is often a source of difficulty in justifying the validity of an IV method.

(3) Nonzero Average Causal Effect of Z on T.

The average causal effect of Z on T, $E[T_i(Z_i = 1) - T_i(Z_i = 0)]$ is not equal to 0.

This assumption ensures that the instrument actually has an impact on the treatment. If the instrument does not change the probability of the treatment assignment, then the instrument is useless because one cannot harness any of the randomization from the instrument to examine the effect of the treatment on the outcome. Note that the average causal effect is estimating the percentage of compliers. If there are more compliers in the study, then one has a stronger IV. If this connection between the instrument and the treatment received is weak then serious problems can arise – see Bound *et al.* (1995) for a discussion on weak instruments.

(4) Monotonicity

$$T_i(Z_i = 1) \geq T_i(Z_i = 0) \quad \text{for all } i = 1, \ldots, N$$

The monotonicity assumption means that the instrument must either encourage units to take the treatment or discourage units from taking the treatment, it cannot have both effects. The monotonicity assumption says that the defiers – those who do exactly the opposite of what they are encouraged to do – are not present in our study. This is an interesting addition to the literature.

(5) Stable Unit Treatment Value Assumption (SUTVA):

$$\text{If } Z_i = Z_i', \quad \text{then } T_i(Z) = T_i(Z')$$

$$\text{If } Z_i = Z_i' \text{ and } T_i = T_i', \quad \text{then } Y_i(Z,T) = T_i(Z',T')$$

SUTVA implies that the potential outcomes for each person i are unrelated to the treatment status of other individuals. This assumption means that settings in which one unit's treatment assignment impacts another unit's outcome are outside of our investigative range. Some examples of tricky situations: immunizations because the probability of unit i being infected depends on how many immunized people there are in the community (i.e., 'herd immunization') and the effect of academic ability from a teacher on a student is tricky to identify because students will learn from peers who potentially receive instruction from other teachers.

Some informal thoughts about the IV assumptions: Assumption 1 (Uniform Random Assignment) is challenging to defend because the assumption is about unobserved quantities. One method for reassuring the reviewer that Assumption 1 is at least plausible is by checking to see if the observed covariates look reasonably random across the different values of the instrument. This is not a guarantee of the randomness of the instrument (nor is it technically a disproval), but it is perhaps reassuring. Assumption 2 is sometimes dubious and will be a point of contention if the reviewer is clever. Assumption 3 is testable from the data because the association is observable in the data. Assumption 4 is often feasible; see the 'complier' terminology below to see why. Assumption 5 (SUTVA) is most often violated in studies involving infectious disease and settings where there is a 'spillover' effect from one subject to another.

Heterogeneity and compliance classes

If the treatment is believed to affect people differently, then it is possible that the compliance classes can be thought of as arising from heterogeneity. It is likely that the always-takers know they will benefit from the treatment, possibly more than others. The never-takers possibly have lower expected benefit. And the compliers would have an unknown level of benefit. This is not necessarily how things work in an example, but is a plausible enough scenario to show that estimating LATE is likely to be different than estimating ATE. The analyst needs to be aware of this issue. The estimate that we get from an IV analysis is only on a subset of the population, and perhaps this subset of the population is not representative of the overall population.

A few examples of an instrument in the medical literature: travel time to treatment facility (McClellan *et al.*, 1994), regional variation in treatment practices (Hadley *et al.*, 2003), and for drug utilization the instrument of prior patient's drug prescription (Brookhart *et al.*, 2006).

Regression Discontinuity

Regression discontinuity (RD) designs take advantage of an abrupt difference in treatment assignments. An example: say we are interested in the effects of a new blood pressure drug. An RD design might be available if there were protocols for treatment selection based on weight. Let us say that there was a policy requiring that anyone lesser than 70 kg is ineligible for the new drug. It might be possible, if physicians and patients strictly adhere to this policy, that the patients who weigh 69.5 kg and the patients who weigh 70.5 kg are actually quite similar in terms of their important covariates but face quite

different prospects for receiving the new blood pressure medicine.

RD designs can be thought of as a special case of IVs, where the analyst has a dichotomized instrument (i.e., whether the subject is above or below the discontinuity). Like an IV design, the assumption of random assignment to the levels of the treatment needs to be discussed. Continuing with the blood pressure medicine example, one might justify the 70 kg cut-off as being similar to random assignment because (1) the scales are likely to have measurement error, (2) patients' weights can vary throughout the day, and (3) the method for weighing (e.g., with clothing or without) will impact the patient's estimated weight. These arguments help with Assumption (1) for the IV assumptions, but the other assumptions need to be similarly addressed.

Often an RD estimate is valid for only those people who are 'near' the discontinuity. In the example, it is likely that one is looking at similar patients if one is considering people who are 69–69.9 kg versus 70.1–70.9 kg. However, it seems likely that the groups defined by 50–69.5 kg and 70.5–90 kg are quite different. This is a case of LATE, where the compliers are additionally restricted to some neighborhood around the discontinuity point. The authors have heard this estimand referred to (in jest) as 'very LATE.'

Before-and-after (difference-in-differences)

The before-and-after and the difference-in-differences (DiD) methods are common techniques to address the possibility that there are unobserved covariates which are causing confounding. Both techniques take advantage of multiple measurements taken at different periods in time. These techniques have been used to great benefit – see Card and Krueger (1994). Both are important techniques in the field, but the authors will do little more than mention them here. For a more detailed discussion the authors recommend looking in standard econometric textbooks for 'panel data' techniques.

Sensitivity analysis

In one sense sensitivity analysis is a 'meta' method because it functions as an analysis of the results of an already existent analysis. The researcher must first select a method, possibly one of the methods described in Sections Methods to Address Selection Bias and Methods for Overt Bias and Bias Due to Omitted Variables, and then perform a sensitivity analysis on that. Acknowledging that the assumptions are just that, merely assumptions, a sensitivity analysis will reanalyze the analysis considering violations of the assumptions occur. A general formulation of a sensitivity analysis is difficult, because it is dependent on the underlying method of analysis. But sensitivity analysis offers a powerful tool for observational studies to explore the effect of the assumptions necessary to make a causal interpretation of the data analyzed.

In an experiment, the randomization to treatment or control allows the researcher to address unobserved variation. In observational studies, the analyst is forced to rely on assumptions to address unobserved variation. Again, the clever reviewer knows how to come up with plausible scenarios and variables which will invalidate the assumptions required to use the method you are employing. (This is not unique to observational studies, in experimental settings it is often called into question whether or not the experimenters did the proper kind of adjustments and randomization in order to truly randomize the experimental subjects.) The good thing about a sensitivity analysis is that it switches the burden of defending your analysis from a case-by-case defense against each possible scenario and instead moves the argument to an order of magnitude (e.g., yeah, each of these arguments are interesting, but they would need to increase selection into the treatment group by a factor of 5 and at the same time increase the rate of death by four times).

A detailed description of sensitivity analyses can be found in Rosenbaum (2002, Chapter 4). Note that a sensitivity analyses will only indicate the magnitude of hidden biases that would alter a study's conclusions but does not address how to overcome these biases.

Example Revisited

NICUs have been established to deliver high-intensity care for premature infants (those infants born before 37 weeks of gestation). If one looks at all of the preemies that were delivered in Pennsylvania between 1995 and 2005, it is seen that 2.26% of the preemies delivered at high-level NICUs died whereas only 1.25% of the preemies who were delivered at low-level NICUs died. No one believes the difference in outcomes reported above is solely attributable to the difference in level of intensity of treatment. People believe it is due to difference in covariates. Based on the observable covariates, this is plausible because it is seen that preemies delivered at high-level NICUs weighed approximately 250 g less than the preemies which were delivered at low-level NICUs (2454 at high-level NICUs vs. 2693 at low-level NICUs). Similarly preemies delivered at high-level NICUs were born a week earlier than their counter parts at low-level NICUs on average (34.5 vs. 35.5 weeks). If you perform a propensity score matching using the observed covariates then the analysis will give you an estimate saying that there is a reduction of 0.05% of deaths if the preemies were to be delivered at high-level NICUs. Inverting a paired *t*-test, the confidence interval for this goes from $(-0.05\%, 0.15\%)$, and is thus an insignificant result. This is meaningful result for policy if the assumption of overt bias only, which is a necessary assumption in propensity score matching, holds in this example.

But one does not have access to medical records. One only has access to health claims data. It is quite likely one does not have all necessary covariate in our dataset, so assuming only overt bias is likely to lead to biased estimates. To attempt to deal with this problem Baiocchi *et al.* (2010) used an IV approach. They used distance to treatment facility as an instrument, because travel time largely determines the likelihood that mother will deliver at a given facility but appears to be largely uncorrelated with the level of severity a preemie experiences. Using this approach Baiocchi *et al.* (2010) estimated a CACE of 0.9% with a confidence interval of (0.57%, 1.23%). Be aware that the authors are estimating a different parameter. It is only appropriately thought of as estimating for a subset of the population, so one cannot readily compare the two estimates. But it is suggestive to note the larger estimated effect, as well as the significance of the result.

You should not walk away from this example thinking that IV methods are always preferred over propensity score matching methods. That is most definitely not the point here. But you should be aware that there are several different methods out there, and you should be comfortable thinking about what is the appropriate method to use in a given situation. In the NICU example, because one only has access to medical claims data – instead of medical charts – it is likely one is missing covariates that would inform us about what we believe to be important selection bias. Given that, one needs to use some method to address the unobserved selection bias, above and beyond the approaches one used to deal with the observed selection bias.

Acknowledgments

The grant is from the Department of Health and Human Services and has designation: 1RC4CA155809-01.

See also: Comparative Performance Evaluation: Quality. Economic Evaluation of Public Health Interventions: Methodological Challenges. Evaluating Efficiency of a Health Care System in the Developed World. Health Econometrics: Overview. Heterogeneity of Hospitals. Sample Selection Bias in Health Econometric Models

References

Angrist, J. (1991). Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics* **47**, 243–266.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with Discussion). *Journal of the American Statistical Association* **91**, 444–455.

Baiocchi, M., Small, D., Lorch, S. and Rosenbaum, P. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* **105**, 1285–1296.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–450.

Brookhart, M. A., Wang, P. S., Solomon, D. H. and Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17**(3), 268–270.

Card, D. and Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* **84**(4), 772–793.

Connors, A., Speroff, T., Dawson, N., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* **276**, 889–897.

Cox, D. R. (1958). *Planning of experiments*. New York: John Wiley.

Fisher, R. A. (1949). *Design of experiments*. Edinburgh: Oliver and Boyd.

Hadley, J., Polsky, D., Mandelblatt, J., et al. (2003). An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a Medicare population. *Health Economy* **12**, 171–186.

Hansen, B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* **99**(467), 609–618.

Heckman, J. and Robb, R. (1985) Alternative methods for evaluating the impacts of interventions. In Heckman, J. J. and Singer, B. (eds.) *Longitudinal analysis of labor market data*. New York: Cambridge University Press.

Heckman, J. and Vytlacil, E. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* **96**, 4730–4734.

Heckman, J. and Vytlacil, E. (2000). The relationship between treatment parameters within a latent variable framework. *Economic Letters* **66**, 33–39.

Hernan, M. A. and Robins, J. M. (2013). *Causal inference*. Chapman & Hall. Available at: http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/ (accessed 06.05.13).

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**(396), 968–970.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86**(1), 4–29.

McClellan, M., McNeil, B. J. and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction reduce mortality? *Journal of the American Medical Association* **272**(11), 859–866.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. *Statistical Science* **5**, 463–480.

Pearl, J. (2009). *Causality: models, reasoning, and inference*, 2nd ed. Cambridge: Cambridge University Press.

Rosenbaum, P. (2002). *Observational studies*, 2nd ed. New York: Springer.

Rosenbaum, P. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician* **59**(2), 147–152.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Occupational Licensing in Health Care

MM Kleiner, University of Minnesota and NBER, Minneapolis, MN, USA

Introduction

Many developed countries require occupational licenses for everyone from surgeons to interior decorators. Licensing in effect creates a regulatory barrier to entry into licensed occupations, and thus results in higher income for those with licenses. However, licensing is assumed to protect the public interest by keeping incompetent and unscrupulous individuals from working with the public. According to data collected in a 2008 national survey of the US workforce, 76% of all nonphysician medical occupations were licensed, the highest among all occupations in the survey conducted by the Krueger (2008). The goal of this article is to outline the major tensions between the monopoly face of licensing and the consumer protection face of occupational regulation in the health care industry. To do this, a theory of licensing is presented, which includes how it is used and some of the controversies surrounding its implementation, and the limited empirical results examining its effectiveness in enhancing quality or restricting competition.

To distinguish various forms of regulation, licensing, certification, and registration is defined below.

- Licensing: Licensing refers to situations in which it is unlawful to carry out a specified range of activities for pay without first having obtained a license. This confirms that the license holder meets prescribed standards of competence. Workers who require such licenses to practice include doctors, lawyers, and nurses.
- Certification: Certification refers to situations in which there are no restrictions on the right to practice in an occupation, but job holders may voluntarily apply to be certified as competent by a state-appointed regulatory body. Two examples of certification would be a certified financial analyst or a certified respiratory therapist.
- Registration: Registration refers to situations in which one can register one's name and address and qualifications with the appropriate regulatory body. Registration provides a standard for being on the list, but complaints from consumers or improper listing of credentials can result in removal from the list.

Licensing has been among the fastest growing labor market institutions in the US. Figure 1 shows the growth of occupational licensing relative to the decline of union membership since the 1950s. By 2008, occupational licensing in the US had grown to 29% of the workforce, up from below 5% in the 1950s. In contrast, unions represented as much as 33% of the US workforce in the 1950s, but declined to less than 12% of the US workforce by 2008. Much of this change was because of the shift from manufacturing employment to service sector employment such as medical services (e.g., nurses), where unions have continued to grow.

A similar trend exists for the UK with declining unionization trends, but growth in occupational licensing. Figure 2 shows that these trends in UK for the period 1978–2008 are consistent with the US trend line. In contrast to the US, approximately 14% of the UK workforce are licensed, but only approximately 22% belong to unions (Hummphris *et al.*, 2011). The data were compiled from the British Labour Force Survey for both unions and licensed occupations. With the growth in the service industries, the percentage of the workforce in licensed occupations appears to be rising in the UK.

The Theories of Occupational Licensing

Here the evolution of theories of occupational licensing is reviewed, ranging from the mechanistic ones to those that utilize human capital theory. It begins by outlining the simplest theory of occupational licensing, which draws more heavily on administrative procedures than on economics. Insights from more complex theoretical models is then incorporated that challenge some of the straightforward assumptions of the simple theory and which thereby provide richer insights into the operation and effects of regulation.

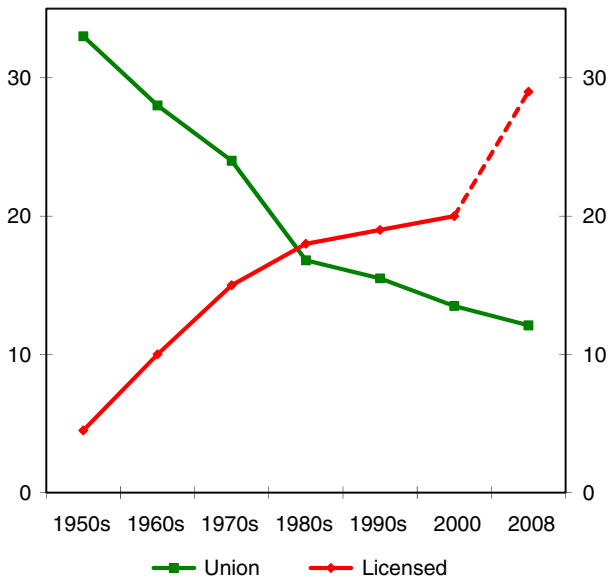


Figure 1 Comparisons in the time-trends of two labor market institutions in the US: Licensing and unionization. The dashed line shows the value from state estimates of licensing based on the Gallup survey and PDII survey results. The union membership estimates are from the current population survey.

 doi:10.1016/B978-0-12-375678-7.01101-9

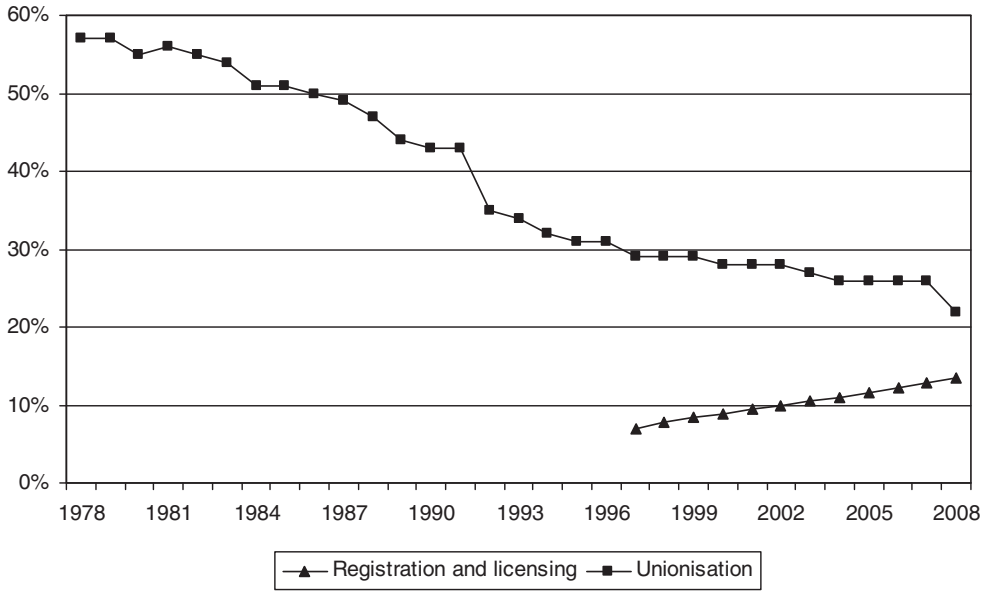


Figure 2 Comparisons of two labor market institutions: Licensing and unionization in the UK. Labour Force Survey (2008).

An Administrative Theory of Licensing

A simple theory of occupational licensing envisions a costless supply of unbiased, capable gatekeepers and enforcers. The gatekeepers screen entrants to the occupation, barring those whose skills or character suggest a tendency toward low-quality output. The enforcers monitor incumbents and discipline those whose performance is below standard with punishments that may include revocation of the license needed to practice. Assuming that entry and performance are controlled in these ways, the quality of service in the profession will almost automatically be maintained at or kept above standards that are set by the gatekeeper to the profession. Within this approach only those who have the funds to invest in training and the ability to do the work are able to enter the occupation.

Introducing economics to this otherwise mechanical model by noting that a key discipline on incumbents – the threat of revoking one's license – may not mean much if incumbents can easily re-enter the profession, such as by moving to a new firm, or by shifting to an alternative occupation with little loss of income. Because grandfathering (i.e., allowing current workers to bypass the new requirements) is the norm when occupations seek to become licensed, incumbent workers are usually supportive of the regulation process. In the absence of grandfathering, lower skilled workers in the occupation may have to seek alternative employment. For example, if sales skills are the key to both providing licensed sales of heart monitors and the nonlicensed selling of shoes or cars, then individuals may shift between these lines of work with little loss of income.

Under these circumstances, meaningful discipline for license holders may require deliberate steps to ensure that loss of license entails significant financial loss. Such additional steps could include imposition of fines, improved screening to prevent expelled practitioners from re-entering the occupation, or requiring all incumbents to put up capital that would be forfeited upon loss of the license. To offset the possibility that incumbents could shift to other occupations with little loss of income, entry requirements could be tightened to limit supply and create monopoly rents within the licensed occupation. The threat of losing these monopoly rents could, in principle, give incentives to incumbents to maintain quality standards. This may also result in some increases in human capital investments in order to attain the additional requirements. The rents could also motivate potential entrants to invest in high levels of training in order to gain admittance. This suggests that licensing can raise quality within an industry by restricting supply, raising labor wages and output prices. Increasing prices may signal either enhanced quality because of perceived or actual skill enhancements or restrictions on the supply of regulated workers.

State-regulated occupations can use political institutions to restrict supply and raise the wages of licensed practitioners. This is assumed to be a once-and-for-all income gain that accrues to current members of the occupation who are 'grandfathered' in, and do not have to meet the newly established standards (Perloff, 1980). Generally, workers who are 'grandfathered' are not required to ever meet the standards of the new entrants. Individuals who attempt to enter the occupation in the future will need to balance the economic rents of the field's increased monopoly power against the greater difficulty of meeting the entrance requirements.

Once an occupation is regulated, members of that occupation in a geographic or political jurisdiction can implement tougher statutes or examination pass rates and may gain relative to those who have easier requirements by further restricting the supply of labor and obtaining economic rents for incumbents. Restrictions would include lowering the pass rate on licensing exams, imposing higher general and specific requirements, and implementing tougher residency requirements that limit new arrivals in the area from qualifying for a

license. Moreover, individuals who have finished schooling in the occupation may decide not to go to a particular political jurisdiction where the pass rate is low because both the economic and shame costs may be high.

One additional effect of licensing is that individuals who are not allowed to practice at all in an occupation as a consequence of regulation may then enter a nonlicensed occupation, shifting the supply curve outward and driving down wages in these unregulated occupations. If licensing requirements contain elements of required general human capital, then it is possible that these workers may raise the average skill level in their new occupation.

Applications to Health Care

Standard economic theory of the effects of occupational licensing regulations on prices and quantities in the health care industry begins with the analysis of Friedman and Kuznets (1945) and Friedman (1962). In this line of reasoning, licenses act as a barrier to entry that can restrict supply and increase wages and other prices relative to a counterfactual competitive market. By contrast, paternalistic arguments and the existence of asymmetric information favor regulating health service providers. The issue is that because providers (e.g., physicians, nurses and physical therapists) may know more about a patient's health condition and the available treatment options, consumers may unwittingly receive low-quality care and possibly that this low-quality care will have larger and sometimes irreversible consequences. Governments might fear that by allowing 'lower skill' providers – such as nurses relative to doctors – to provide health services they may be exposing consumers to increased risk. In some situations the risks could also impose externalities: for instance, if low-skill health providers increase the transmission of an infectious disease then there might be a case for regulation. This raises two issues. One is that a paternalistic regulator might want to increase the quality of care received in the market. Another is that a regulator might want to ensure that providers have a minimum level of competency to minimize the negative consequences of asymmetric information. Both of these issues are supported by evidence and analysis.

A major argument for the licensing of medical occupations is that it eliminates or reduces the patient's health risk of seeking the services from an occupation. If testing and background checks 'eliminate charlatans, incompetents, or frauds' (Council of State Governments, 1952), then consumers may be willing to pay a higher price for the services offered by the regulated occupation. A review of the body of theory from experimental economics and psychology shows that consumers value the reduction in downside risk more than they value the benefits of a positive outcome (Kahneman *et al.*, 1991). This preference by consumers for the status quo or reducing the risk of a highly negative outcome has been called 'loss aversion,' which is an element of prospect theory developed by Kahneman and Tversky (1979). For example, as discussed earlier, social welfare may be increased substantially by minimizing the likelihood of a poor diagnosis as a consequence of going to an incompetent physician, because the incompetent physicians have been weeded out as a result of licensing. Consequently, licensing may also reduce patients' perceived benefits of receiving nonstandard but potentially highly effective treatment from an unlicensed practitioner of traditional medicine. Using the power of the state to both limit the downside risk of poor quality care and reduce the possibility of an upside benefit may be a trade-off that maximizes consumer utility or welfare. Evidence of the acceptance of this trade-off can be found in the growth of occupational licensing during the past century across virtually all nations that have been studied.

The gains from an unregulated service can be potential benefits from free market competition of lower prices and greater innovation without the constraints of a regulatory body, such as a licensing board. This upside potential gain can be achieved through both the use of nonstandard methods or new research that has not been approved by the licensing agency as appropriate for the service (Rottenberg, 1980). Deviations from the prescribed methods of providing a service are discouraged by licensing boards, and may even be found to be illegal. For example, not having a dentist on site is illegal in the US when providing a service such as teeth cleaning. Dental hygienists generally are not allowed to 'practice' without a dentist on site, with the 'site' being defined by statute or the dental board. In addition, dental hygienists are not allowed to open offices to compete with dentists. Although this policy reduces the chance that a dental hygienist will fail to find a major disease that may require immediate attention, it also reduces the ability of the hygienist to provide the limited services that particular patients require. Moreover, there is little leeway for the dental service industry to provide new or innovative services without the risk being found in violation of the licensing laws. The licensing laws give rise to the labor relations concept of 'featherbedding,' whereby dentists are on the premises – but do little work.

Consequently, regulation through licensing medical services can be the equivalent of a closed shop in unionized markets. Theoretically, higher wages are likely to result from restricted labor supply. Because closed shops in unionized markets are illegal in both the US and the UK, it is interesting that, with respect to organized labor markets, the equivalents of closed shops are nevertheless permitted in licensed occupations.

Illustrations in Medical Markets

One illustration of the potential outcome of licensing in medical markets is presented by the Nobel laureate economist Milton Friedman. Consistent with his work cited earlier in this section, he finds licensing to have an overall negative influence for consumers. The argument can be found on YouTube as follows (Milton Friedman – Health Care in a Free Market, YouTube video, 9:03, from a question-and-answer session with medical professionals at the Mayo Clinic in 1978, posted by 'LibertyPen,' 25 September 2009):

http://www.youtube.com/watchv=-6t-R3pWrRw&feature=related – Milton Friedman

Further illustrations of the influence of licensing on markets can be found in commercial media as well. For example, libertarian commentator John Stossel poses additional questions regarding the value of licensing in an excerpt from his

television show that featured several episodes on occupational regulation ('Stossel Show – Licensing! (Part 1/6),' YouTube video, 9:43 (part 1), posted by 'TheChannelOfLiberty,' 17 March 2010). The excerpt on YouTube, at the link below, serves as an overarching illustration of the influence of licensing on labor and consumer markets:

http://www.youtube.com/watch?v=f0JGu4tlmmk – The Stossel Show – Licensing

On a more practical basis, however, the entry requirements for many health care occupations are often presented as including the requirement for being licensed, as illustrated in another YouTube video ('How to Get Medical Jobs: How to Become an Occupational Therapist,' YouTube video, 0:50, posted by 'expertvillage,' 26 September 2008):

http://www.youtube.com/watchv=GM9ohB2qyQM

As with many occupations in the health care field, the requirements for being licensed are becoming more stringent. For example, the requirement for becoming a physical therapist in the US has increased within the past 5 years from 2 years of post-high school training to requiring a doctor of physical therapy in order to conduct many of the required tasks.

Empirical Evidence in Licensed Health Care Markets

The initial empirical work in medical markets was based on data from the 1930s. The classic and often cited study completed through the National Bureau of Economic Research by Nobel Prize-winning economists Milton Friedman and Simon Kuznets estimated that the 33% earnings premium of physicians relative to dentists could be attributed to more than just 1 year's difference between the requirements to become a doctor versus a dentist (Friedman and Kuznets, 1945). They estimate that the difference in earnings between doctors and dentists should be approximately 17% based on human capital and other observable factors, but that the additional 16% residual gap is in large part a consequence of physicians' greater ability to restrict labor supply. Milton Friedman's book, Capitalism and Freedom, argues that physicians were able to obtain substantial earnings gains over dentists during the 1920s and 1930s because they were able to limit new student enrollment in medical school (Friedman, 1962). More recently, however, a reversal of these trends has taken place.

During a relatively recent period of time (1990–2000), the number of new physicians increased by almost 22% (Public Use Sample, US Census, 2000). In contrast, the total number of dentists during the same period of time decreased. Dental school enrollment increased by only 1% each year during the 1990s, and the number of dentists in the US declined by almost 2% over the decade as a result of both retirements and individuals leaving the occupation (Public Use Sample, US Census, 2000).

A more general review of empirical research on licensing (in the US?) found that licensing is associated with consumer prices that are 4–35% higher than those found among unlicensed occupations, depending on the type of commercial practice and location (Kleiner, 2006). Kleiner and Kudrle (2000), for example, found that tougher state-level restrictions and more rigorous pass rates for dentists were associated with hourly wage rates that were 15% higher than in states with few restrictions, with no measurable increase in observable health benefits.

Occupational licensing appears to increase wages in several nations in the European Union (EU), but the estimates usually are lower than in the US. In the EU nations with greater overall wage disparities, such as in the UK, wages in the licensed occupations of medical practitioners, pharmacists, pharmacologists, and dental practitioners were an estimated 6–65% higher than otherwise similar workers in unlicensed occupations (Kleiner, 2006). In contrast, physicians and dentists in France earn an estimated 8–21% more than their unlicensed colleagues, whereas workers in those professions in Germany, which has lower overall wage disparities, have similar wages relative to unlicensed occupations.

In Europe, according to Dubois *et al*. (2006), a recent trend in the case of medicine and allied professions has been a shift from a system of self-governance that has traditionally granted professional associations' disproportionate power in setting and monitoring standard toward one that grants the state more influence in the process.

Given the high level of licensing within the health care occupations, it is not surprising to find that one of the evolving issues is the question of who is responsible for tasks, and how the government determines the market. For example, dentists control the market for dental care and in most US states dental hygienists must work for a dentist and cannot open their own establishments (Kleiner and Park, 2010). The result is that in those few US states that allow hygienists to open their own offices they make approximately 10% more and have faster employment growth relative to hygienists in more restrictive states.

A complementary study by Wanchek (2010) found that using a detailed dental hygiene professional practice index and a simultaneous equation approach to reduce the potential influence of endogeneity of wages and employment, entry requirements are negatively correlated with dental hygienists' employment and that practice restrictions that limit hygienists' ability to do tasks within the dental office are negatively correlated with their wages. Higher wages and lower employment of hygienists both reduce access to care, as observed in the prevalence of dental office visits. Finally, the author finds that the results are consistent with a state's entry and practice regulations jointly affecting access to oral healthcare.

Similarly, for nurses and doctors, in those US states that allow nurses to do simple procedures, such as 'well-baby' exams without the supervision of a physician, health insurance spending is approximately 10% lower (Kleiner *et al*., 2012) than in more restrictive states with no apparent influence on the quality of health care. Licensing not only influences wages and prices relative to unregulated situations, but also influences wages and prices across regulated occupations.

Kyoung-Hee Yu and Frank Levy (2010) examined the reasons why one might expect it to be more difficult to do offshore licensed professional work than manufacturing work in a globalized world. The authors conduct numerous interviews and provide data on a specific case: the offshoring of diagnostic radiology from the US, the UK, and Singapore, and find that regulation of the occupations matters. As far as professional services in healthcare are concerned then, institutional barriers are real and useful for the professions in

terms of restricting entry. To the extent that institutional frameworks differ across nations, globally integrated markets have yet to emerge for professional services in healthcare.

Conclusions

Licensing can in effect create a regulatory barrier to entry into licensed occupations, and thus results in higher income for those with licenses. Although more research is needed for a definitive answer, preliminary evidence points to licensing raising wages and prices in health care, but with no clear influence on the quality of care or with a clear impact on downside outcomes such as hospital readmissions, repeat visits to a health care professional, or deaths due to incompetent or unscrupulous purveyors of health care services. More detailed analysis using experimental data and field experiments with elements of random assignment would enhance the ability to make policy recommendations regarding the licensing of health care occupations.

See also: Nurses' Unions

References

Council of State Governments (1952). *Occupational Licensing Legislation in the States*. Chicago: Council of State Governments.

Dubois, C. A., Dixon, A. and McKee, M. (2006). Reshaping the regulation of the workforce in european health care systems. In Dubois, C. A. (ed.) *Human resources for health in Europe*, pp. 173–192. Berkshire: WHO – Open University Press.

Friedman, M. (1962). *Capitalism and freedom*. Chicago: University of Chicago Press.

Friedman, M. and Kuznets, S. (1945). *Income from independent professional practice*. New York: National Bureau of Economic Research.

Hummphris, A., Kleiner, M. and Koumenta, M. (2011). How does government regulate occupations in the UK and US? Issues and policy implications. In Marsden, D. (ed.) *Employment in the lean years. Policy and prospects for the next decade*, pp. 87–101. Oxford, UK: Oxford University Press.

Kahneman, D., Knetsch, J. and Thaler, R. (1991). The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* **5**(1), 193–206.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**(2), 263–292.

Kleiner, M. (2006). *Licensing occupations: ensuring quality or restricting competition?* Michigan: W.E. Upjohn Institute.

Kleiner, M. and Kudrle, R. (2000). Does regulation affect economic outcomes? The case of dentistry. *Journal of Law and Economics* **43**(2), 547–582.

Kleiner, M., Mairer, A., Park, K. W. and Wing, C. (2012). 'Relaxing Occupational Licensing Requirements: Analyzing Wages and Prices for a Medical Service,' Working Paper, Minneapolis, Minnesota: University of Minnesota.

Kleiner, M., and Park K. (2010). 'Battles Among Licensed Occupations: Analyzing Government Regulations on Labor Market Outcomes for Dentists and Hygienists.' NBER Working Paper 16560. Cambridge, MA: National Bureau of Economic Research.

Krueger, A. B. (2008). Princeton Data Improvement Initiative (PDII), Conference Report. Princeton, New Jersey: Trustees of Princeton University.

Perloff, J. (1980). The impact of licensing laws on wage changes in the construction industry. *Journal of Law and Economics* **23**(2), 409–428.

Princeton Data Improvement Initiative (2008). Princeton University.

Rottenberg, S. (1980). Introduction. *Occupational Licensure and Regulation*, pp. 1–13. Washington, DC: American Enterprise Institute.

US Census 5-Percent Public Use Microdata Sample (PUMS) Files (2000).

Wanchek, T. (2010). Dental hygiene regulation and access to oral healthcare: Assessing the variation across the US states. *British Journal of Industrial Relations* **48**(4), 706–725.

Yu, K. H. and Levy, F. (2010). Offshoring professional services: Institutions and professional control. *British Journal of Industrial Relations* **48**(4), 758–783.

Further Reading

Kleiner, M. and Krueger, A. (2010). The prevalence and effects of occupational licensing. *British Journal of Industrial Relations* **48**(4), 676–687.

Kleiner, M. M. and Krueger, A. B. (2013). Analyzing the extent and influence of occupational licensing on the labor market. *Journal of Labor Economics* S173–S202.

Organizational Economics and Physician Practices

JB Rebitzer, Boston University, Boston, MA, USA; National Bureau of Economic Research, Cambridge, MA, USA; Case Western Reserve School of Medicine, Cleveland, OH, USA; Center for the Institute of the Study of Labor (IZA), Bonn, Germany, and The Levy Institute, Hudson, NY, USA
ME Votruba, Case Western Reserve School of Medicine, Cleveland, OH, USA

Introduction

It is a commonplace to observe that the healthcare delivery system in the US is in crisis: costs are high and rising rapidly, the quality of care is inadequate along important dimensions, and the delivery system is rife with inefficiencies and waste. What is less commonly acknowledged is that many of the prominent strategies for reforming healthcare delivery systems are based on strong, largely untested beliefs about how organizations can best coordinate and motivate the physicians involved in patient care.

The crisis in the US system is especially severe and there is a large literature that discusses the US experience. For these reasons, this article focuses heavily on the US healthcare system. But many other countries are experiencing similar difficulties with their healthcare delivery systems and many of the issues and ideas discussed will have application beyond the confines of the US.

Health services researchers have long argued that a central problem with healthcare delivery in the US is fragmentation. Individual patients are frequently treated by numerous care providers who have only weak organizational ties with one another, resulting in poor information flows and inadequate care coordination. This fragmentation especially inhibits the close coordination between diverse providers that is required to manage costly chronic diseases and to prevent errors and missteps. The obvious fix, according to this view, is to induce physicians to join or construct more integrated care delivery systems. As a step in this direction, the Patient Protection Affordable Care Act of 2010 directs the Centers for Medicare and Medicaid Services to create a voluntary program for Accountable Care Organizations (ACOs). This program is designed to nudge the US healthcare system toward more integrated care delivery.

In contrast to the health services research community, many economists have argued that the fundamental problem with the delivery system is poor incentives, primarily physician incentives. From this perspective, the continuing dominance of fee-for-service payment systems and flawed payment rates within these systems create strong incentives for physicians to deliver high-cost services. Incentives to improve care quality are largely indirect and weak – patients presumably seek out higher quality physicians but have little ability to evaluate physician quality. Expensive and inefficient practice styles are further supported by overly generous insurance coverage, a consequence of tax breaks for employer-based coverage and the widespread use of supplemental Medicare coverage.

The result is a bloated system where neither physicians nor patients are held directly accountable for the financial consequences of their care decisions. The obvious fix, according to this view, is to get incentives right by reforming payment systems and tax policy. Thus ACOs are encouraged to adopt new payment systems that, in principal, reward physicians for adopting practice styles that reduce costs and improve quality. At the heart of these payment reform proposals is an economic theory of how best to motivate physicians using material incentives.

This article presents an overview of organizational economics as it applies to physician practices. The first section, A Descriptive Overview, offers a brief descriptive overview of the structure of physician practices. Our point will be that the long-anticipated triumph of integrated care delivery has largely gone unrealized. Although fewer physicians are working in solo practices than in the past, self-employed physicians in small group practices remain the industry norm. Integration of physicians' business functions has increased through a variety of cross-group and group-hospital affiliations, but integration of clinical activities has lagged. In the next section, Agency and Pay for Performance, the problem of pay for performance from the perspective of principal agent (PA) models is discussed. Building on these models, the next two sections take up the organizational economics of integrated care. The section Professional Autonomy versus Integration examines the professional norms of autonomy and other factors that complicate effective integration between hospitals and physicians. The section Coordination, Specialization and Innovation, considers how medical specialization affects coordination across providers. Finally the section Prospects for Accountable Care Organizations considers how the problems of motivation and coordination play out in ACOs.

A Descriptive Overview

The physician workforce has evolved overtime in both size and composition. In the mid-twentieth century, there were approximately 14 physicians per 10 000 persons in the US, almost half of whom practiced general primary care (Table 1). In response to concerns about an impending physician shortage, the Kennedy and Johnson administrations successfully championed legislation subsidizing medical schools, doubling the number of medical school graduates from 1965 to 1980, and eventually doubling the physician-to-population ratio from 1970 to 2000. Despite a subsequent stabilization of medical school graduation rates, entry into the medical profession continued to modestly outpace population growth largely through an increase in international medical graduates, who now comprise a quarter of active physicians.

As medical knowledge and technology grew, the share of general practitioners in the medical profession declined

 doi:10.1016/B978-0-12-375678-7.01112-3

Table 1 Physicians' characteristics over time

	Year						
	1949	*1960*	*1970*	*1980*	*1990*	*2000*	*2008*
MDs per 10 000 population[a]	14.1	14.5	16.4	20.6	24.7	28.9	31.4
Active MDs per 10 000 population	12.8	13.8	15.3	18.3	22.0	24.6	25.8
Pct general practice/family medicine	50.1	35.6	18.6	14.5	12.9	12.5	12.0
Pct other primary care	15.0	21.1	24.6	26.7	26.1	27.2	27.0
Percent female[b]			7.1	11.4	17.2	25.2	31.2
Mean hours worked per week[c]				53.5	53.4	51.4	49.6 (2007)
Male				54.8	54.8	53.4	51.7
Female				41.3	46.0	44.8	44.4
Self-employed				53.5	54.3	52.1	50.9
Employed				53.4	51.8	51.0	49.0

[a]Reproduced from National Center for Health Statistics (2011). *Health United States 2010: With special feature on death and dying*. Washington, DC: Centers for Disease Control and Prevention. Before 1970, MDs with unknown address or unclassified primary specialty were counted among active MDs. Starting in 1970, MDs with unknown address or unclassified primary specialty were not counted as active. Percent 'other primary care' includes primary-care specialties of internal medicine, obstetrics/gynecology and pediatrics. Count of primary-care physicians in obstetrics/gynecology is unavailable before 1970. Table assumes 6.0% of active physicians were in obstetrics/gynecology in 1949 and 1960, consistent with the fraction in 1970 (6.0%) and in 1980 (5.9%).

[b]AMA (various years). Reflects percent female among active MDs.

[c]Current Population Survey, as reported by Staiger, D. O., Aurebach, D. I. and Buerhaus, P. I. (2010). Trends in the work hours of physicians in the United States. *Journal of the American Medical Association* **303**(8), 747–753; for nonresident physicians.

Table 2 Changes in care delivery settings

	Year							
	1975	*1980*	*1985*	*1990*	*1995*	*2000*	*2005*	*2008*
Hospital-based care								
Admissions (per 100 population)	16.7	17.1	15.3	13.5	12.7	12.4	12.5	12.3
Mean length-of-stay	11.4	10.0	9.1	9.1	7.8	6.8	6.5	6.3
Outpatient visits (per capita)	1.18	1.16	1.19	1.48	1.84	2.10	2.28	2.33
Office-based care								
Visits (per capita)	2.69	2.63	2.67	2.82	2.65	2.92	3.26	3.14
Primary-care visits (per capita)	1.85	1.74	1.66	1.79	1.63	1.72	1.92	1.87
Specialist visits (per capita)	0.84	0.89	1.01	1.03	1.02	1.20	1.34	1.27

Note: Hospital outpatient visits include visits to the emergency room, hospital outpatient departments, referred visits (pharmacy, EKG, radiology), and outpatient surgeries.
Source: Reproduced from National Center for Health Statistics (2011). *Health United States 2010: With special feature on death and dying*. Washington, DC: Centers for Disease Control and Prevention.

more than 60% between 1949 and 1970, at which time fewer than 20% of active physicians practiced general primary care. The share of general practitioners continued its steep decline through the 1980s, though the decline was partially offset by growing shares of physicians practicing in specialties that typically involve primary-care services, i.e., pediatrics, obstetrics/gynecology, and internal medicine. Since 1990, the share of primary-care physicians (general practitioners plus primary-care specialties) has largely stabilized at 39% of all active physicians. Although the physician–population ratio in the US is similar to that in other developed countries, its share of primary-care physicians is substantially lower and this has been cited by numerous health policy experts as an important deficiency in the US healthcare delivery system.

The trend toward an increasing concentration of medical specialists has been accompanied by an increasing 'feminization' of the physician workforce and, more recently, by a decline in the labor supply of individual physicians. Historically, the medical profession was overwhelmingly dominated by male physicians. In 1970, approximately 7% of active physicians were women but the share of female physicians has grown consistently since then, reaching 31% by 2008. Coinciding with this change has been a decline in the number of hours physicians work each week, especially since the mid-1990s. Thus, although active physicians per capita increased 17% since 1990, physician work hours per capita increased less than 7%.

Table 2 documents trends in care delivery settings. Chief among these is a dramatic decrease in the amount of inpatient care provided. Since 1975, hospital admission rates have declined approximately 25%, whereas the length of hospital stays fell by almost one-half, representing a substantial decline in the amount of care physicians provide on an inpatient basis. In contrast, hospital-based outpatient care increased dramatically over this time, driven in part by a steep rise in

outpatient surgeries. Meanwhile, the per capita rate of office-based visits increased 17% since 1975, but this aggregate figure combines two very different trends. The rate of primary-care visits was relatively flat over this period, consistent with relative stability in the (per capita) number of primary-care physicians, whereas per capita office-based visits to specialists increased by more than 50%.

Thus a twofold story emerges from Table 2. First, physicians spend less time 'making rounds' and performing procedures in an inpatient setting than they did in the past, decreasing the dependence of physicians on hospitals as the setting for delivering services. Second, the decrease in inpatient care has coincided with a dramatic increase in ambulatory care delivered by medical specialists.

Tables 3 and 4 document how physician employment and practice arrangements have evolved over time. The results in Table 3 report statistics derived from a series of physician surveys conducted by the American Medical Association (AMA) – the Periodic Survey of Physicians (1975), the Socioeconomic Monitoring Study (1983–99), and the Patient Care Physician Survey (2001) – through 2001. More recent trends are documented in Table 4 drawing on data collected by the Center for Studying Health Systems Change (HSC) – four waves of data from the Community Tracking Study (CTS) Physician Survey and the HSC 2008 Health Tracking Physician Survey. Caution is always warranted in evaluating trends across different surveys. A special concern in this case is the sampling methodology of the CTS Physician Survey, which focused on 60 communities in the US. All statistics are weighted to be nationally representative, but trends in these communities may have differed from those in other areas, which could confound some of the cross-study patterns we observe. Nonetheless, these results allow us to draw a number of conclusions.

Physician self-employment has declined, but remains the norm: for much of US history the prototypical physician was self-employed and in solo practice. Even as recently as 1983, more than 40% of physicians fit this model but its numbers have

Table 3 Trends in employment and group size, to 2001

	Year							
	1975[a]	*1983*[a]	*1988*[b]	*1991*[b]	*1994*[b]	*1997*[b]	*1999*[c]	*2001*[c]
Panel A: Employment								
Self-employed		75.8	72.1		57.7		62.0	65.5
Solo practice		40.5	38.5		29.3		26.4	24.4
Employee		24.2	27.9		42.3		38.0	34.5
Group practice			7.8		14.6		8.6	8.4
Institutional			20.1		27.7		29.4	26.1
Hospital			4.1		6.7		7.7	7.7
Medical school			5.6		8.2		7.7	7.4
HMO			2.3		4.1		2.6	1.8
State/local govt			4.7		3.2		3.2	2.4
other			3.3		5.5		8.2	6.8
Panel B: Group Size								
Solo practice	54.2	48.9						
2 Physicians	14.1	12.5						
3–7 Physicians	21.3	24.3						
8–25 Physicians	6.0	8.8						
25+ Physicians	4.5	5.3						
(1988–2001 Categories)								
Solo practice			49.3	45.5	42.9	39.0	37.7	33.2
2–4 Physicians			27.2	29.0	28.2	25.8	25.5	26.4
5–9 Physicians			11.6	13.2	14.7	16.6	15.5	16.3
10–49 Physicians			8.1	8.8	10.7	13.8	15.2	17.0
50+ Physicians			3.8	3.5	3.5	4.8	6.2	7.2

[a](1975) AMA Periodic Survey of Physicians and (1983) AMA Socioeconomic Monitoring System. Employment statistics as reported in Kletke, P. R., Emmons, D. W. and Gillis, K. D. (1996). Current trends in physician practice arrangments. *Journal of the American Medical Association* **276**(7), 555–560. Practice size statistics as reported in Ohsfeldt, R. L. (1983). Changing medical practice arrangements. *Socioeconomic Monitoring Report 2*. Chicago: American Medical Association.

[b]AMA Socioeconomic Monitoring System. Employment statistics as reported in Kletke, P. R., Emmons, D. W. and Gillis, K. D. (1996). Current trends in physician practice arrangments. *Journal of the American Medical Association* **276**(7), 555–560. Practice size statistics as reported in Kletke, P. R. (1998). *Trends in physician practice arrangements. Socioeconomic characteristics of medical practices 1997–98*. Chicago: American Medical Association.

[c]AMA Socioeconomic Monitoring System, as reported by Kane, C. K. (2004a). The practice arrangements of patient care physicians, 1999 (revised). *Physician Marketplace Report*. Chicago: American Medical Association.

[d]AMA Patient Care Physician Survey, as reported by Kane, C. K. (2004b). The practice arrangements of patient care physicians, 2001. *Physician Marketplace Report*. Chicago: American Medical Association.

Note: Employment statistics refer to nonfederal postGME patient care physicians. Self-employed physicians are defined as those with full or part ownership in their main practice. Institutional employee category 'other' includes physicians practicing in community health centers, freestanding clinics, and independent contractors in other institutional settings. Group size statistics additional restricted to physicians in solo or group practice.

Table 4 Recent trends in employment, practice setting, and group size, 1996–2008

	Year				
	1996–97[a]	*1998–99*[a]	*2000–01*[a]	*2004–05*[a]	*2008*[b]
Panel A: Employment					
Self-employed	61.6	56.7	55.9	54.4	56.3
Solo/2-physician practice	37.4	33.6	31.2	28.1	28.5
Employee	38.4	43.3	44.1	45.6	43.7
Panel B: Practice Setting					
Solo/group practice	68.9	64.7	65.4	64.1	72
Hospital	10.7	12.6	12	12	13.1
Medical School	7.3	7.7	8.4	9.3	7.3
HMO	5	4.6	3.8	4.5	3.5
Other	8.3	10.5	10.4	10.1	4.1
Panel C: Group Size					
Solo/2-Physician Practices	59.1	57.8	53.8	50.7	44.4
3 to 5 Physician Practices	17.7	14.8	17.9	15.3	20.1
6 to 50 Physician Practices	19.0	21.9	24.2	27.5	26.9
>50 Physician Practices	4.2	5.4	4.1	6.6	8.5

[a]Community Tracking Study Physician Survey. See Liebhaber, A. and Grossman, J. M. (2007). Physicians moving to mid-sized, single-specialty practices. *Tracking Report*. Washington, DC: Center for Studying Health System Change.

[b]HSC 2008 Health Tracking Physician Survey. See Boukus, E. R., Cassil, A. and O'Malley, A. S. (2009). *A snapshot of US physicians: Key findings from the 2008 health tracking survey*. Data Bulletin. Washington, DC: Center For Studying Health System Change.

Note: Statistics calculated to be nationally representative of all nonfederal physicians who spend at least 20 h a week in direct patient care. Self-employed physicians are defined as those with full or part ownership in their main practice. Practice setting category 'Other' includes physicians practicing in community health centers, freestanding clinics and other settings, as well as independent contractors. Group size statistics restricted to physicians in solo or group practice.

been declining. By 2001, fewer than 25% of physicians were self-employed in solo practices. Overall rates of physician self-employment largely track the decline in solo practitioners, falling from 76% in 1983, to approximately 64% in 2000 (in AMA data), to less than 56% in 2008 (HSC data). In the CTS data, self-employment rates were somewhat lower than in AMA data, but a similar modest decline in self-employment rates is observed.

Institutional employment increased, but then appears to have stabilized (probably): the AMA data indicate that that the share of physicians working as employees of institutions (e.g., hospitals, medical schools, HMOs, etc.) grew from 20% to 28% from 1988 to 1994 – during the ascent of managed care – and then was fairly stable to 2001. The HSC data similarly find that approximately 28% of physicians were employees of institutions in 2008, suggesting little change over the last decade. The CTS data tell a somewhat conflicting story over the decade from 1996 to 2005, with rates of institutional employment seeming to increase from 31% to 36%. Given the nature of the CTS sampling methodology, one is inclined to believe this trend may not be representative of the nation as a whole, and that national rates of institutional employment have probably stabilized at a level less than 30%. It is likely, however, that local healthcare markets are quite heterogeneous in this regard.

Practice groups have gotten larger, but small practices remain the norm: health-system analysts have been predicting the demise of solo and small group practices for decades. Based on AMA data, it appears that a majority of noninstitutional physicians were in solo practice in 1975, declining to a third by 2001. Meanwhile, the share of employment in physician groups with more than 10 physicians grew from 12% to 19% between 1988 and 2001. The CTS data indicate that the share of noninstitutional physicians in solo or two-physician practices declined from 59% in 1996–97 to 51% in 2004–05, whereas the share in groups with six or more physicians increased from 23% to 34%. The historic record, then, is one of decreasing shares of physicians in the very smallest practices, and increasing shares in larger groups. Despite this, however, small groups remain a common feature of the physician labor market. In 2008, 65% of noninstitutional physicians were in practices with five or fewer physicians, accounting for 46% of the physician workforce overall.

Practice size and the prevalence of physician institutional employment (in hospitals and staff-model HMOs) provide an incomplete picture of the extent that individual physicians coordinate activities with one another and with other providers in the healthcare system. Over the last few decades, physician groups have increasingly joined a variety of cross-group and group-hospital organizations intended to facilitate the collective goals of their participants.

Independent practice associations (IPAs) have emerged to provide solo and small group practices many of the benefits associated with larger group practice – economies of scale in insurance contract negotiations, contract oversight, and other administrative functions – while allowing participating physicians greater autonomy over their individual practices. According to the Managed Care Information Center, there are currently approximately 500 IPAs with approximately 264 000 participating physicians, which equates to approximately 55% of the active physicians in group practice.

A variety of organizations have emerged linking physicians and hospitals, with the goal of integrating service delivery and financing. The most common of these, physician-hospital

organizations (PHOs), represent joint ventures between hospitals and private physicians to negotiate and manage insurance contracts. PHOs also frequently operate clinics, employ physicians and staff, and acquire medical practices. Some have established their own insurance products. At the apex of managed care in the mid-1990s, nearly a third of hospitals had one or more PHOs. The fraction has subsequently declined, falling to 13.4% by 2008.

A third type of organization, known as management service organizations (MSOs), are organizations owned by physician groups, by physician-hospital ventures, or by investors in conjunction with physicians. MSOs generally exist to provide practice management and administrative support, thus relieving practices of nonmedical business functions. In some cases, MSOs acquire the facilities and equipment of their client physicians which they then lease back to the physicians.

The widespread existence of such organizations can give the impression that physicians, even those in small practice, are often tightly aligned with one another and frequently integrated with hospitals in their geographic area. However, closer inspection suggests a more nuanced story. The rise of cross-group and group-hospital organizations was largely due to the market pressures imposed by managed care organizations (MCOs). Consolidating business functions allowed groups to achieve economies of scale in nonmedical activities and, more importantly, increased the clout of providers in contract negotiations. These organizations, however, were largely unsuccessful at integrating clinical activities across participating providers. As managed care backed away from capitated contracts, the impetus to integrate clinical activities largely faded.

Agency and Pay for Performance

Physicians know more than patients or insurers about the set of effective treatment options available for an individual with a specific condition. A patient's own physician knows better than other physicians the specifics of the patient's clinical and personal situation. Specialists, with their advanced training and narrow focus on a small set of clinical issues, have a better understanding of treatment options in their specialty than the primary-care providers who refer patients to them. Each of these informational asymmetries creates an agency problem, i.e., a situation in which well-informed physicians recommend or implement courses of action that benefit the well-informed physician at the expense of patients, payers, and other less well-informed parties.

One way to resolve information asymmetries is to design incentive contracts that motivate the best informed agent to 'do the right thing.' Economics offers a well-developed theory for tackling these incentive design problems, the PA model. The canonical PA model considers how a principal might structure incentives to elicit optimal behavior from better informed agents whose interests do not completely conform to the interests of the principal. The principal conditions pay on observed outcomes and the actions taken by the agent reflect the influence of these incentives. The starting point for the vast literature on PA models is a remarkable and quite general result: it is possible to implement a reward structure that can elicit efficient behavior on the part of the agent even when agents are entirely self-interested and even when performance measures are noisy and imperfect.

On the basis of this fundamental result, one might expect that the economist's prescription for efficient healthcare delivery would be widespread use of pay for performance contracts. Unfortunately economic theory does not support such a simple policy prescription. The qualifications on the fundamental PA result are almost as far reaching as the result itself.

The first qualification has to do with the basic statistical properties of performance measures. Clinical outcomes are often influenced by some unknown combination of good actions (taken by the healthcare provider and/or the patient) and good luck. Because of the great diversity of possible medical conditions a patient can manifest and the limited number of patients in a physician's panel of patients, it is not at all clear that an individual physician's practice offers enough observations to reliably distinguish good luck from good medical practice.

Noise in performance measures becomes even more important when agents are risk averse. To see this, consider a setting in which physicians operate under a contract that rewards them for coming in under a threshold level of costs for their entire panel of patients. If costs vary substantially in a year due to actions outside of a physician's control, then large payouts for coming in under target add a considerable component of randomness to a physician's compensation. For risk-averse physicians, the increase in the variability of payments that result from incentive pay imposes a real cost. To make matters worse for incentive design, income variability rises with the intensity of the incentive. As a practical matter, insurers or HMOs that ignore the cost of this increased risk when implementing pay for performance will find they will need to pay more to attract physicians to their networks.

The problems posed by the low statistical power of clinical performance measures and the risk aversion of physicians can be mitigated by pooling information across many physicians. From an economic perspective, pooling or averaging performance measures is problematic because it makes agency problems worse. The larger the physician panels across which outcomes are measured, the less will be the effect of an individual physician's actions on group outcomes. The phenomenon of group incentives weakening as the size of the group increases is well understood in the economics literature where it is often referred to as the 'free-riding' problem.

The relevance of free-riding problems is evident when one looks at the pay practices of physicians groups. Physicians who work in group practices often share revenues among themselves. Revenue sharing has the appeal of allowing physicians to buffer variations in income (doctor A's extra income in a good year will help offset doctor B's poor income in a down year), but this insurance comes at the cost of weakening incentives. If revenues are equally shared, a doctor in a three-person group keeps one-third of each dollar he or she earns. Incentives weaken as the size of the group grows: a doctor in a five-person group keeps one-fifth of the marginal dollar in revenues they earn and so on.

If it is costly or difficult to use incentives to resolve agency problems with meaningful pay for performance systems, organizations might find it profitable to reduce the need for incentives by seeking out physicians who have their principal's

interests at heart. Consider a hypothetical physician who is committed to providing patients with the level of care the patient would choose for themselves if they knew as much as the physician knew and if the marginal cost of care were zero. Finding physicians with such 'altruistic' preferences may not be as hard as it seems. Many aspects of medical education can be understood as efforts to inculcate this attitude into young physicians. It might appear that having physicians with such pro-social, intrinsic motives eliminates the most important agency problem facing doctors: the problem of ensuring that physicians act in the interests of their insured patients. But physicians face more than one agency problem and it is unlikely that the sort of intrinsic motives just described would resolve the agency problem for the insurers and MCOs that have the responsibility of paying for care. Indeed, if insurance contracts were such that the patients themselves had to pay the direct cost of their care, they also might prefer a physician whose internal values moved them to balance the marginal benefit of care against its marginal cost.

A deeper, but more speculative, limitation on the use of intrinsic motives to resolve agency problems is that these preferences may not coexist easily with the use of material incentives. A growing body of theoretical and experimental evidence in economics and psychology points toward the provocative possibility that powerful extrinsic rewards can actually weaken the efficacy of such pro-social motives as altruism, reciprocity, intrinsic motivation, and a desire to uphold ethical norms. If true, this suggests that organizations that rely on a mix of material incentives and intrinsic motivators might be less efficient than organizations that rely solely on one or the other strategy to resolve agency issues.

To sum up, the economics literature suggests that incentives matter, but that high-powered pay for performance schemes may be too blunt a tool for handling the many agency problems that raise the cost and reduce the quality of healthcare. This depressing conclusion is offset by some more recent results in the organizational economics literature suggesting that it may be possible to resolve agency issues with very low-powered incentives by employing physicians in integrated healthcare delivery organizations. It is to that issue that the authors now turn.

Professional Autonomy versus Integration

Hospitals and physicians together deliver the bulk of medical services in the US, yet they are strangely divided from each other. Within hospitals, physician decisions are central to resource allocation and care processes, yet most physicians are quite independent of hospital management, working (as most still do) in small single-specialty groups that they own. Some physicians have 'privileges' at more than one hospital and many more split their time and attention between hospital inpatient care and their office-based practices. It is hard to think of another industry – outside of movie making and construction – that relies so heavily on independent contractors as key decision makers. In virtually every industry that, like hospitals, relies on the mass-production of goods or services, key decision makers are either employees of the enterprise or, much more rarely, its owners. Does this difference matter for the efficiency of the healthcare system? Organizational economics suggests that it might.

For hospitals, the great advantage of having physicians as employees rather than independent contractors is that the employment relationship offers the possibility of resolving agency issues without the distortions created by high-powered incentives. This feature of employment relationships has been most clearly analyzed in the context of multitask models, where agents have more critical tasks to perform than can be included in performance measures. High-powered incentives will, in this context, cause the agent to deliver too much of the metered tasks and not enough of the unmetered tasks. Employment relationships offer a straight-forward fix for these multitask problems because employers have the ability to tell employees the tasks included in their job. By restricting the range of tasks the employee can work on while at work, the employer reduces the opportunity cost of doing the tasks the employer favors. As a result, a small amount of incentive can have a large effect on performance with lower levels of distortion. Put slightly differently, employment relationships differ from market-based relationships in that firms can exert a high degree of influence on employee actions using very little pay for performance. The strength of these low-powered incentives is increased when combined with other features of well-run organizations: the careful selection of new employees combined with their subsequent socialization into the goals and procedures of the enterprise. The effectiveness of the combination of appropriate job design, careful selection, socialization, and low-powered incentives is captured by the term of art used in the management literature, 'high performance human resource systems.'

To see the power of weak incentives in the context of employment relationships imagine that a hospital wishes to improve the way in which surgical tools are sterilized and delivered to operating rooms – a surprisingly complicated process that involves surgeons, operating room nurses, hospital managers, and sterilization technicians. Suppose further that operational efficiency can be improved by reducing the number and variety of surgical tools available to surgeons but that negotiating this change involves meetings and consultations with surgeons. If surgeons are independent contractors paid per operation, any additional meeting takes time away from the next operation. Attending such a meeting then, is a very expensive task for the surgeon and the surgeon requires equally large benefits in order to be induced to participate. This incentive problem is made worse by the fact that the benefits to the independent surgeon of reducing the number of surgical tools in circulation are clearly less than the benefits accruing to the hospital as a whole, especially if the surgeon divides his operating time across a number of hospitals.

Contrast these incentives with those of a surgeon who is employed by a hospital and is paid on salary. In this setting, attending meetings and participating in improving the sterilization process is not nearly so costly to the surgeon because the opportunity cost of his time is relatively low. These incentives to participate are further strengthened by the relationships the physician builds with coworkers and also by the extent of tacit, firm-specific knowledge, acquired over the course of the employment relationship.

Low-powered incentives of the sort discussed can complement higher-powered incentives and may help explain the anomalous findings of the Physician Group Demonstration project, an experiment in pay for performance involving large provider groups. Allowing these groups to keep 80% of their savings (after the first 2% of savings) elicited only small and uneven cost reductions. Very little is known about why some physician groups succeeded and others failed to achieve savings. Free riding can, as seen, undermine the incentive effects of conventional pay for performance – but the low-power organizational incentives discussed in this section can easily 'scale up' for large organizations. It is possible that the variation observed in the demonstration project may be the result of unobserved variations in low-powered incentives that can augment under powered explicit pay for performance incentives.

Given their considerable advantages, why hospitals employing physicians and forming large, integrated care delivery systems are not seen? In most economic settings, the efficiency advantages of integrated systems should enable them to generate the resources to attract large numbers of physicians and members. What prevents this from happening? Surprisingly little attention has been devoted to this important issue. The studies that do address it tend to focus on three potential explanations: the nature of economic competition in healthcare, strategic complementarities between payment systems and healthcare delivery, and the sociology of the medical profession. Each of these in turn are considered.

Regarding economic competition, if payers are unable to measure and reward high value-added producers, then it may be that the enhanced efficiency of integrated systems will not translate into a sustainable competitive advantage. Medicare, the biggest single buyer of healthcare services, does not evaluate the benefits associated with new medical technologies when setting prices, and it is forbidden from using cost-effectiveness analysis and from selectively contracting with more efficient physician groups. Medicare regulatory boards charged with evaluating new technology are concerned primarily with whether new drugs or procedures offer positive benefits. Private insurance coverage is heavily influenced by Medicare coverage. In addition, private payers typically use Medicare prices as a reference point in bargaining, and contracts based on value creation are scarce. Indeed it may be that some employers who purchase insurance for employees are not interested in or capable of evaluating the quality of care their employees receive. The key challenge for the 'failure of competition' explanation for the absence of integrated systems is explaining why competition in healthcare is different than in other sectors where markets do appear able to assess and reward efficient organizational designs.

In contrast, the 'strategic complementarities' explanation for the scarcity of integrated health delivery organizations refers to a generic set of explanations for the failure of advanced production methods to defuse rapidly across industries. Indeed much of this work was originally inspired by the difficulties American manufacturers had in imitating and adopting more efficient 'lean' manufacturing techniques that originated in Japan and the difficulties firms had in realizing productivity gains from the revolution in information. Suppose that managers have identified two complementary innovations, A and B. Each innovation on its own produces a small benefit, but introducing A and B simultaneously yields a big improvement in productive efficiency. For concreteness suppose that innovation A involves redesigning job responsibilities in ways that tap the tacit information and problem solving abilities of front line employees to solve customer problems and that innovation B involves hiring more educated workers. Implementing either of these changes is not easy or inexpensive and so it is reasonable to expect firms to experiment with one or the other innovation rather than implementing them both simultaneously. Thus, a firm might try action A and be disappointed in the result and therefore not follow-up with step B, hiring a more highly educated work force. Similarly, firms might start with B, but not see much productivity gain because they did not implement step A and redesign job responsibilities in ways that allow the more educated workers to use their superior problem solving and communication skills. It is only when firms reorganize and reskill the workforce that the powerful complementarities between the two innovations are realized. Put differently, incremental experimentation might not reveal the full productivity benefits of complementary innovations and so the true value of innovations might not be discovered by managers.

If complementarities can impede innovation within one organization, it becomes even harder when the complementary innovations span multiple organizations, i.e., when innovations are what game theorists call strategic complements. According to this argument the full efficiency gains of integrated care delivery can only be realized under bundled prospective payment systems. But in communities with highly fragmented care delivery, it is hard to find providers who can carry the risks entailed by such payments. As a result, payers do not innovate away from the status quo fee-for-service payment system and there is little competitive advantage for providers to move out of their currently fragmented delivery organizations. One of the interesting implications of the 'strategic complementarities' explanation is that it offers a natural role for public policy. Specifically, the big public payers (Medicare and Medicaid) can force the issue by announcing that they will be moving toward a bundled prospective payment system that will benefit large integrated organizations. This is the intellectual basis for the ACO initiative discussed below.

The third explanation for the relative scarcity of integrated care delivery organizations concerns social norms. The simplest version of the social norms explanation is this: physicians value professional autonomy and do not want to be employed by anyone else. There is considerable historical evidence that physicians as a learned profession did and do value their autonomy. Unfortunately, this fact alone is not likely to support a satisfactory explanation for fragmented delivery systems. If fragmentation between physicians and hospitals was simply the result of a preference for 'being your own boss,' then one should observe that physicians working for integrated systems enjoy a significant wage premium to compensate them for the disutility of their status as employees. One is not aware of any study that documents such a pay differential.

More sophisticated models of social norms, however, offer a more promising line of investigation. In models with a more

sociological flavor, agents compare their actions with a prescribed set of behaviors or with the actions of others in their reference group.

In a conventional microeconomic analysis, if a physician decides to work as an employee at a hospital, only the hospital and the physician are involved in the transaction. All else equal, if the hospital offers a pay differential that exceeds the value the physicians personally place on autonomy, the physicians will choose to abandon the autonomy of their independent practice and go to work as an employee of the hospital. Things work quite differently, however, when norms enter the picture. Norm violating transactions necessarily precipitate actions or changed perceptions (and loss of reputation) by third-party physicians who are not party to the transaction. The involvement of these third parties allows professional norms to persist even when the gains to individuals from violating norms are large relative to their preference for the norm. The involvement of third parties also suggests that stubbornly persistent norms may be greatly weakened by shocks that change the actions or perceptions of many physicians at once.

Just such a change is currently taking place in the medical profession. For most of the twentieth century, professional norms in medicine, law, and other learned professions were shaped by a labor force composed almost entirely of men, and most of these men had stay-at-home wives. In the 1970s, however, women began entering professions in large numbers and today they account for a significant proportion of the labor force in both medicine and law. For our purposes, the significance of this demographic transition is that these new entrants are likely to be influenced by a different set of norms than the male incumbents. Specifically, these women are often married to male professionals who work long hours and they are for this reason quite likely to have to balance norms of medical practice against family responsibilities. To the extent that employment in a hospital or other large integrated delivery organization enables physicians to have shorter and more predictable hours than working as an entrepreneur in a small practice, women might be drawn to these positions and this may have the effect of undermining the norm of professional autonomy that has played such an important historical role in the US healthcare delivery system.

Norms-based models of professions shift attention from a narrow focus on individual incentives to a broader view that also includes incentives for action governing the entire profession. From this perspective it is interesting to observe that in the early-twentieth century the AMA successfully lobbied for the introduction of 'corporate practice of medicine' laws that made it illegal for physicians to be employed by other organizations, especially hospitals. The legal impact of these laws has diminished overtime (as witnessed by the rise of professional hospitalists, an issue taken up below), but their influence is still observed in some states.

If physician professional norms are important for understanding the failure of physician-hospital integration in the US, then they might also be important for understanding other market outcomes in healthcare. Consider, for example, that MCOs compete for patients (who are the paying customers) and for physicians to participate in their network of providers. Patients cannot directly perceive quality and use as a proxy the number of physicians included in the managed care network. The MCOs compete for physicians by offering a combination of salary and cost-containment incentives. MCOs that write contracts with strict cost-control incentives have lower costs and lower premiums but they also have a harder time recruiting physicians to their network of providers than other MCOs do. In equilibrium, MCOs will segment the market. Some will operate with stringent cost-control incentives, small physician panels, and low premiums. MCOs in this part of the market profit by attracting cost-conscious customers. Other MCOs will have more lax cost-control incentives, bigger provider panels, higher premiums, and they profit by attracting customers who put a greater emphasis on provider choice than they do on the cost of insurance. This product differentiation creates disparities in treatment because physicians will use more resources treating policyholders in the high-cost MCOs than the low-cost MCOs. Given the many agency problems in this setting, an increase in competition is likely to cause a decline in medical costs – what some have referred to as a 'race to the bottom.'

Suppose now, that a physician norm against treatment disparities is introduced, perhaps because physicians do not like to deliver care that uses fewer resources than that delivered by other physicians in the market. This norm makes it more difficult for low cost plans to attract physicians and so they must pay them more (while also reducing cost-containment incentives). With low-cost plans behaving more like high-cost plans, there is less product differentiation and no race to the bottom. Indeed heightened competition reduces product differentiation and increases the overall level of resource utilization in the market. In this way, norms of professional practice can help explain why the managed care revolution of the 1990s failed to deliver on its promise to control the rise of medical costs. The model also can account for the absence of the widely predicted 'race to the bottom' in the managed care market of the 1990s.

Coordination, Specialization, and Innovation

The economic ideas discussed so far have been primarily concerned with the problem of motivating physicians. A second less well-developed economics literature focuses on problems of coordinating care among physicians who must specialize in specific aspects of care because no single individual can master all medical knowledge.

In his famous dictum that specialization is limited by the extent of the market, Adam Smith neatly summarized the role that markets play in coordinating the activity of highly efficient, specialized producers. More recent work has augmented Smith's analysis by considering the amount of specialization that will emerge in different economic settings.

Specialization increases the productive efficiency of a team performing complementary tasks. As specialization increases, however, so does the size of the team as well as the costs of coordinating activities among the increasingly specialized producers. These coordination costs are determined by available technologies, especially communication and transportation technologies, but they can also be influenced by agency problems. Increases in the stock of knowledge

increase the payoff to team members of investing in more specialized knowledge. Heightened specialization, it turns out, also increases the payoff to generating new knowledge. Applied to medicine, this suggests a positive feedback in which dramatic increases in medical knowledge coincide with dramatic increases in the number of narrow medical subspecialties.

The trade-off between coordination costs and specialization can be used to analyze the growth of hospitalists. Hospitalists are a new medical subspecialty whose purpose is to care for patients when they are hospitalized and then return them to the care of their primary physicians after discharge from the hospital. Primary-care physicians have superior information about their patient's specific situation and handing off inpatient care to hospitalists creates the risk that key information will not be communicated. For this reason, the rise of the hospitalist specialty creates coordination costs that were not present under the traditional US model in which primary-care doctors supervised their patient's care in both ambulatory and inpatient settings. Improvements in communication technologies have the effect of reducing coordination costs and thus increasing the demand for hospitalists, but this is not the whole story.

Coordination costs are also determined by the switching costs of moving from ambulatory to inpatient settings. It is costly for physicians to switch from office-based care to visiting their hospitalized patients, and some of these costs are fixed (think of the time and effort costs of leaving the office and traveling to the hospital to see patients). In the presence of these fixed switching costs, anything that reduces the number of patients a physician has in the hospital will reduce a primary-care physician's willingness to supervise their patient's inpatient care. For this reason, reductions in hospital length of stay, increases in the use of outpatient procedures in doctors' offices or even a reduction in physician work hours and patient load can have the effect of increasing demand for hospitalists.

The efficiency gains from specialization are not the only gains from the use of hospitalists. Hospitalists are often employed by hospitals but they may also work as contractors employed by outside firms or physician groups. Whatever their formal status, hospitals are likely to have more influence over hospitalist activities than they do over independent, primary-care physicians. Hospitals will therefore find their hospitalists relatively easy to engage in process improvement initiatives. By the same token, however, hospitals might use this heightened influence to encourage their hospitalists to shift costs onto other parts of the healthcare system. Recent findings about the effect of treatment by hospitalists on Medicare patients give some cause for concern in this regard.

The theories of specialization discussed so far are appealing, but they do not consider the referral patterns observed in medicine. In medicine, primary-care providers are generalists trained to recognize and treat common and less difficult conditions. When less common or more difficult patients arrive, the primary-care physicians refer them to specialists who have the extra training and experience required to handle these cases. It follows from this that a fall in the time and effort cost of communication with specialists increases the number of conditions that primary-care physicians will refer to specialists whereas a fall in the costs of learning about rare conditions (e.g., via internet search) broadens the number of cases the primary-care providers will handle themselves.

The process of referral from generalist to specialist creates an agency problem. Consider, for example, a patient who approaches her primary-care doctor for treatment for a rash. The primary-care physician can either refer the patient to a dermatologist or treat the condition themselves and generate extra revenues. If the dermatologists' in-depth knowledge leads to superior and cost-effective treatment, the referral is efficient. Efficient referrals may not occur, however, if the primary-care physician loses too much revenue by referring the patient. Although there may be little concern that an internist will fail to refer a breast cancer patient to a breast surgeon and oncologist, there are a very large number of conditions that fall into a gray area where the skills and knowledge of the generalist and specialist overlap.

In medicine where generalists make the decision to refer to more highly trained specialists, professional partnerships may have a distinct advantage. This is because the revenue sharing agreements in these partnerships allow the referring primary-care doctor to earn some money from the fees the specialist generates. This suggests that to best realize the advantages of efficient referrals, multispecialist groups ought to be composed of physicians working in areas where agency issues are likely to arise. Thus there might be good incentive reasons to include internists and dermatologists in the same group, but not cardiac surgeons.

As already observed, innovation in healthcare has resulted in a division of labor in which specialists with advanced training focus on the most difficult and advanced sort of medical practice. It is also possible, however, that innovations in treating the most common and routine sorts of care might also be very important.

Consider that healthcare delivery in the US must be concerned with treating two very different kinds of medical issues. One the one hand, there are the difficult, hard to assess cases that require sophisticated pattern recognition and nonroutine decision making by the physician (think, here, of the many conditions featured on the TV show 'House' whose etiology or treatment protocol is murky). On the other hand, there are the familiar cases whose treatment can be handled by clear, evidence-based protocols. In the typical physician practice, the responsibility for both of these cases falls to the physician. This division of labor makes some sense as individual patients can unexpectedly acquire one or the other type of condition, and their primary-care physician is in an excellent position to coordinate care across both these types of issues. But this approach to coordinating care also increases costs and dampens important innovation. Care for the protocol-based conditions, if broken out of the physician's practice, will be less expensive because the caregiver is not an expensive or highly trained generalist. In addition, organizations that specialize in protocol-based care for common issues can use the techniques of modern management to implement continuous improvement processes that drive down costs and improve effectiveness. The job of implementing these techniques will be made simpler by the fact that physicians will not play a central role in these organizations. More provocatively there is evidence from other industries suggesting that innovations originating

in the low-cost, low-prestige parts of an industry often end up transforming the production processes required for high-end goods and services as well. If this pattern holds true for medicine, improvements in the delivery of care through 'mini-clinics' and other limited care delivery operations may end up increasing the rate of innovation in the entire industry.

Prospects for Accountable Care Organizations

ACOs are an organizational innovation created as part of the Medicare Shared Savings Program of the Patient Protection and Affordable Care Act that was signed into law by President Obama in 2010. Although ACOs are only a small part of a huge piece of legislation, they have attracted a great deal of attention from policy-makers, physicians, and managers.

ACOs are a network of hospitals and providers that contract with the Center for Medicare and Medicaid Services (CMS) to provide care to a large bloc of Medicare patients (5000 or more). The contracts, which last for 3 years, create a single risk-bearing entity with incentives to control costs. ACOs that come in under their specified cost benchmarks earn a fraction of the savings. To receive these payments the ACO must also meet stringent standards on 65 quality indicators that reflect patient and caregiver experience, care coordination, patient safety, preventative health, and health of at-risk frail and elderly populations.

It is interesting to consider the ACO experiment from the perspective of organizational economics. For the statistical reasons we discussed in the Section on principal-agent models ACOs must enroll large numbers of Medicare patients in order to generate reliable measures of savings. But, as emphasized, implementing pay for performance in large groups creates free-riding problems that can dramatically weaken incentives. Put differently, if the ACO is comprised of independent contractor physicians connected only by a common hospital and a common incentive plan, they are unlikely to achieve the desired changes in provider behaviors. Selection, socialization, training, and careful job design are what gives a large organization the ability to influence the behavior of physicians in large groups. If these elements are missing, it is hard to see ACOs having much effect on the way healthcare is delivered.

To achieve savings, the ACO has to manage the capabilities of hospitals and the primary-care physicians who make up of the ACO. The most straightforward way to manage these very different capabilities would be for hospitals to simply employ physicians, but as discussed there are historical, legal, strategic, and sociological obstacles to achieving this goal. Simply purchasing physician practices, as many hospitals, and PHOs did during the 1990s, will not do the trick, but it may not be necessary for ACOs to employ all their primary-care physicians. Some organizations appear to be able to incorporate a significant number of nonemployed physicians into ACO-like arrangements and this offers some hope for expanding the range of hospital–physician coordination. A critical element in these organizations is to build legitimacy among independent physicians by making them part of the governance of the organization.

Incorporating specialists into the ACO will be challenging because specialists are not required to limit themselves to a single ACO. The economic model of referrals suggests that ACOs can reduce referrals by introducing training and computer-assisted decision support that make it easier for generalists to substitute their own decisions for those of the specialists. It may, for example, be better to train primary-care physicians to treat rashes and acne rather than sending every case of rash or acne to a dermatologist. However, the vast explosion in medical knowledge implies that there are limits to the substitution of generalist for specialist care. In this case, it may be that efficiently managing referrals to specialists will entail bringing some specialists into the ACO. Keeping these specialists fully occupied will also exert upward pressure on the optimal scale of ACOs.

Given their size, it is likely that free-riding issues will cause ACOs to operate with under-powered incentives, i.e., with incentives that are too weak, by themselves, to elicit meaningful changes in behavior. From this perspective it is helpful to think of the ACO's incentive problem as analogous to the provision of effort when effort is a public good. The experimental literature on public goods provision suggests that the effects of incentives on public good provision depend critically on the 'meaning' agents give to the incentive. Well-designed incentives should communicate that they are intended to achieve a socially beneficial outcome rather than threatening individual autonomy or sense of justice. Extending this logic to the case of intrinsically motivated physicians; managing ACOs likely involves paying careful attention to assigning meaning to the payments, but it is unclear if this meaning is more easily constructed within conventional employment relationships or within hybrid organizations in which doctors participate under looser arrangements. Given the medical profession's long history of battling to preserve its status as an autonomous and learned profession, low-powered incentives in ACOs built on a hybrid organizational form might be workable. However, conventional organizations may have greater opportunities to train, screen, and socialize for physicians who might respond well to low-powered incentives.

To the extent that successful ACO's have organizational capabilities that rely on training, screening, socialization, and constructing the 'meaning' of incentives they likely also involve relational contracts. Relational contracts are based on informal trusting arrangements whose credibility is enforced by the continuing value of the relationship between parties. The great advantage of relational contracts for ACOs is that they can complement more formal relationships such as those involved in pay for performance. Incentives that would be under powered in the sense of a principal–agent model may be quite a bit more effective if performance this period determined the continuation of a valuable ongoing relationship. Relational contracts can also be used to reduce some of the distortions of high-powered formal incentives.

Taken together, our analysis suggests that as a policy intervention, ACOs are likely to have the biggest effect where care is already integrated. Advocates of ACOs know this and see ACOs as emerging from five different practice arrangements. In order of ease of implementation these are: integrated delivery systems that combine insurance, hospitals, and physicians; multispecialty group practices; PHOs; IPAs, and virtual physician organizations.

Conclusions

This article applies the conceptual tool-kit of organizational economics to the economics of physician practices. Our discussion has focused on three broad themes from organizational economics: PA problems (both conventionally economic and behavioral); inefficiencies in the market for organizational form (resulting from social norms and various market failures); and the trade-off between the productivity gains from specialization and the coordination costs specialization entails.

These themes have been applied to important features of physician practices. Much of the attention has focused on understanding the stubborn persistence of fragmented care delivery via small, physician-owned practices, but other important issues have been considered as well. These include: the mixed record of pay for performance – especially in large healthcare organizations; the difficulties of achieving efficient levels of referrals between generalized and specialized providers; and the emergence of a fast-growing new medical specialty, hospitalists, as a result of changes in the tradeoffs between specialization and coordination costs. The final section brings all the themes together in an assessment of the prospects for ACOs, an important public policy initiative in the US aimed at reforming both incentive systems and the organizational forms within which care is provided.

In each of the applications it was found that the ideas of organizational economics yielded genuine and sometimes unexpected insights. This gives one some confidence that the idiosyncratic features of physician practices do not invalidate insights gleaned from the study of other, more standard, economic entities. In the long-struggle to improve healthcare efficiency, organizational economics will likely help providers, managers, and policy makers better understand how best to coordinate and motivate the physicians who guide patient care.

This article is a shortened and abridged version of a longer essay 'Organizational Economics and Physician Practices' James B. Rebitzer and Mark Votruba, NBER Working Paper 17 535 (October 2011, updated February 2013). Please consult that essay at the National Bureau of Economics website (www.nber.org) for a full list of relevant citations as well as a more extensive discussion of the literature.

Acknowledgment

In writing this article, we benefited from suggestions provided by Lawton R. Burns, Alan B. Cohen, Keith Ericson, Meredith B. Rosenthal, Mark Rukavina and Victor Fuchs. We are responsible for any errors or omissions. This research was funded in part by a gift from Microsoft Corp.

See also: Demand for Insurance That Nudges Demand. Managed Care. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Physician Labor Supply

Further Reading

American Medical Association (2013). *Physician characteristics and the distribution in the United States*. Washington, DC: American Medical Association Press.

Boukus, E. R., Cassil, A. and O'Malley, A. S. (2009). *A snapshot of US physicians: Key findings from the 2008 health tracking survey*. Data Bulletin. Washington, DC: Center For Studying Health System Change.

Kane, C. K. (2004a). The practice arrangements of patient care physicians, 1999 (revised). *Physician Marketplace Report*. Chicago: American Medical Association.

Kane, C. K. (2004b). The practice arrangements of patient care physicians, 2001. *Physician Marketplace Report*. Chicago: American Medical Association.

Kletke, P. R. (1998). *Trends in physician practice arrangements. Socioeconomic characteristics of medical practices 1997–98*. Chicago: American Medical Association.

Kletke, P. R., Emmons, D. W. and Gillis, K. D. (1996). Current trends in physician practice arrangements. *Journal of the American Medical Association* **276**(7), 555–560.

Liebhaber, A. and Grossman, J. M. (2007). Physicians moving to mid-sized, single-specialty practices. *Tracking Report*. Washington, DC: Center for Studying Health System Change.

National Center for Health Statistics (2011). *Health United States 2010: With special feature on dealth and dying*. Washington, DC: Centers for Disease Control and Prevention.

Ohsfeldt, R. L. (1983). Changing medical practice arrangements. *Socioeconomic Monitoring Report*. Chicago: American Medical Association.

Staiger, D. O., Aurebach, D. I. and Buerhaus, P. I. (2010). Trends in the work hours of physicians in the United States. *Journal of the American Medical Association* **303**(8), 747–753.

Panel Data and Difference-in-Differences Estimation

BH Baltagi, Syracuse University, Syracuse, NY, USA

Introduction

Panel data refer to data sets consisting of multiple observations on each sampling unit. This could be generated by pooling time series observations across a variety of cross-sectional units, including countries, hospitals, firms, or randomly sampled individuals, like nurses, doctors, and patients. This encompasses longitudinal data analysis in which the primary focus is on individual histories. Two well-known examples of the US panel data are the Panel Study of Income Dynamics, and the National Longitudinal Surveys of Labor Market Experience. European panels include the German Socioeconomic Panel, the British Household Panel Survey (BHPS), and the European Community Household Panel (ECHP). Panel data methods in health economics have been used to estimate the labor supply of physicians and nurses; study the relationship between health and wages and health and economic growth; examine the productivity and cost efficiency of hospitals; and estimate the effect of pollutants on mortality. They have also been used to study the relationship between obesity and fast food prices; determine whether beer taxes will reduce motor vehicle fatality rates; and whether cigarette taxes will reduce teenage smoking, to mention a few applications. For example, Askildsen *et al.*, 2003 estimate nurse's labor supply for Norway. The panel data used include detailed information on 19 638 nurses observed over the period 1993–98. The policy question tackled is whether increasing wages would entice nurses to supply more hours of work. Contoyannis and Rice, 2001 estimate the impact of health on wage rates using the first six waves of the BHPS. Abrevaya (2006) utilizes the federal Natality Data Sets (released by the National Center for Health Statistics) from 1990 to 1998, to estimate the causal effect of smoking on birth outcomes. Identification of the smoking effect is achieved in this panel from women who change their smoking behavior from one pregnancy to another. Abrevaya constructs a matched panel data set that identifies mothers with multiple births. With the most stringent matched criterion, this data set contains 296 218 birth observations with 141 929 distinct mothers. Baltagi and Geishecker (2006) estimate a rational addiction model for alcohol consumption in Russia. Their panel data set includes eight rounds of the Russian Longitudinal Monitoring Survey spanning the period 1994–2003. These are four examples of micropanel data applications in health economics and as clear from these data sets, they follow a large number of individuals over a short period of time.

In contrast, examples of macropanels in health economics include Ruhm (1996) who uses panel data of 48 states (excluding Alaska, Hawaii, and the District of Columbia) over the period 1982–88 to study the impact of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates. Greene (2010) who uses the World Health Organization's panel data set to distinguish between cross-country heterogeneity and inefficiency in health-care delivery. This panel follows 191 countries over the period 1993–97. Becker, Grossman, and Murphy (1994), who estimate a rational addiction model for cigarette consumption across 50 states (and the District of Columbia) over the period 1955–85. Baltagi and Moscone (2010) who use a panel of 20 Organization for Economic Co-operation and Development countries observed over the period 1971–2004 to estimate the long-run economic relationship between health-care expenditure and income. Macropanels follow aggregates like countries, states, or regions and usually involve a longer period of time than micropanels. The asymptotics for micropanels has to be for large N, as T is fixed and usually small, whereas the asymptotics for macropanels can be for large N and T. Also, with a longer time series for macropanels one has to deal with issues of nonstationarity in the time series, like unit roots, structural breaks, and cointegration (*see* Chapter 12 of Baltagi, 2008). Additionally, with macropanels, one has to deal with cross-country dependence. This is usually not an issue in micropanels where the households are randomly sampled and hence not likely correlated.

Some of the benefits of using panel data include a much larger data set. This means that there will be more variability and less collinearity among the variables than is typical of cross-sectional or time series data. With additional, more informative data, one can get more reliable estimates and test more sophisticated behavioral models with less restrictive assumptions. Another advantage of panel data is their ability to control for individual heterogeneity. Not controlling for these unobserved individual-specific effects leads to bias in the resulting estimates. For example, consider the Abrevaya (2006) application, where one is estimating the causal effect of smoking on birth weight. One would expect that mothers who smoke during pregnancy are more likely to adopt other unhealthy behavior such as drinking, poor nutritional intake, etc. These variables are unobserved and hence omitted from the regression. If these omitted variables are positively correlated with the mother's decision to smoke, then ordinary least squares (OLS) will result in an overestimation of the effect of smoking on birth weight. Similarly, in the Contoyannis and Rice (2001) study, where one is estimating the effect of health status on earnings, one would expect the health status of the individual to be correlated with unobservable attributes of that individual, which, in turn, affect productivity and wages. If this correlation is positive, one would expect an overestimation of the effect of health status on wages. Cross-sectional studies attempt to control for this unobserved ability by collecting hard-to-get data on twins. However, using individual panel data, one can, for example, difference the data over time and wipe out the unobserved individual invariant ability.

Another advantage of panels over cross-sectional data is that individuals 'anchor' their scale at different levels, rendering interpersonal comparisons of responses meaningless. When you ask people about their health status on a scale of

1–10, Sam's 5 may be different from Monica's 5, but in a cross-sectional regression you assume they are the same. Panel data help if the metric used by individuals is time-invariant. Fixed effects (FE) makes inference based on intra- rather than interpersonal comparisons of satisfaction. This avoids not only the potential bias caused by anchoring but also bias caused by other unobserved individual-specific factors.

Limitations of panel data sets include problems in the design, data collection, and data management of panel surveys. These include the problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period, and time-in-sample bias. Another limitation of panel data sets is the distortion due to measurement errors. Measurement errors may arise because of faulty response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, misrecording of responses, and interviewer effects. Although these problems can occur in cross-sectional studies, they are aggravated in panel data studies. Panel data sets may also exhibit bias due to sample selection problems. For the initial wave of the panel, respondents may refuse to participate or the interviewer may not find anybody at home. This may cause some bias in the inference drawn from this sample. Although this nonresponse can also occur in cross-sectional data sets, it is more serious with panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is high.

The Model

Most panel data applications use a simple regression with error component disturbances:

$$y_{it} = \alpha + X'_{it}\beta + \mu_i + v_{it} \quad i = 1,\ldots,N; \quad t = 1,\ldots,T \qquad [1]$$

with i denoting individuals, hospitals, countries, etc. and t denoting time. The i subscript, therefore, denotes the cross-sectional dimension, whereas t denotes the time series dimension. The panel data are balanced in that none of the observations are missing whether randomly or nonrandomly due to attrition or sample selection. α is a scalar, β is $K \times 1$, and X_{it} is the it-th observation on K explanatory variables. μ_i denotes the unobservable individual-specific effect and v_{it} denotes the remainder disturbances, which are assumed to be independent and identically distributed IID$\left(0,\sigma_v^2\right)$. For example, in the Contoyannis and Rice (2001) study of the impact of health on wage rates using the first six waves of the BHPS, y_{it} is log of average hourly wage, whereas X_{it} contains a set of variables like age, age$_2$, experience, experience$_2$, union membership, marital status, number of children, race, education, occupation, region indicator, etc. The variable of interest is a self-assessed health variable, which is obtained from the response to the following question: "Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?" Contoyannis and Rice constructed three dummy variables: (sahex = 1, if an individual has excellent health), (sahgd = 1, if an individual has good health), and (sahfp = 1, if an individual has fair health or worse). They also included a General Health Questionnaire: Likert Scale score which was originally developed as a screening instrument for psychiatric illness but is often used as an indicator of subjective well-being. Contoyannis and Rice constructed a composite measure derived from the results of this questionnaire which is increasing in ill health (hlghq1).

Fixed Effects

Note that μ_i is time invariant and it accounts for any individual-specific effect that is not included in the regression. If the $\mu_i^{'}$s are assumed as fixed parameters to be estimated and the X'_{it}s are assumed independent of the v_{it} for all i and t, the FE model is obtained. Estimation in this case amounts to including $(N-1)$ individual dummies to estimate these individual invariant effects. This leads to an enormous loss in degrees of freedom and attenuates the problem of multicollinearity among the regressors. Furthermore, this may not be computationally feasible for large micropanels. By the Frisch–Waugh–Lovell Theorem (Baltagi, 2008) one can get this FE estimator by running least squares of $\tilde{y}_{it} = y_{it} - \bar{y}_{i.}$ on the $\tilde{X}_{it}$'s similarly defined, where the dot indicates summation over that index and the bar denotes averaging. This transformation eliminates the μ'_is and is known as the within transformation and the corresponding estimator of β is called the within estimator or the FE estimator. Note that the FE estimator cannot estimate the effect of any time-invariant variable, such as race or education. These variables are wiped out by the within transformation. This is a major disadvantage if the effect of these variables on earnings is of interest. Note that, if T is fixed and $N \to \infty$ as typical in short labor panels, then only the FE estimator of β is consistent; the FE estimators of the individual effects $(\alpha + \mu_i)$ are not consistent because the number of these parameters increases as N increases. This is known as the incidental parameter problem. Note that when the true model is FE, OLS suffers from omission variables bias and inference using OLS is misleading. For the sample of 859 males in the Contoyannis and Rice (2001) study, the OLS estimate for excellent health is 0.065 and significant, whereas the OLS estimate for good health is 0.019 and insignificant (both are contrasted against a baseline of fair, poor, and very poor health). The FE estimates are 0.013 for excellent health and 0.010 for good health, and both are insignificant. The OLS estimate for the General Health Questionnaire: Likert Scale score (hlghq1) is -0.002 and insignificant, whereas the FE estimate is -0.003 and significant. More dramatically, for the Ruhm (1996) study, OLS gets a positive (0.012) and significant effect of real beer taxes on motor vehicle fatality rates, whereas FE obtains a negative (-0.324) and significant effect of real beer taxes on motor vehicle fatality rates.

Janke *et al.* (2009) examine the relationship between population mortality and common sources of airborne pollution in England. The data covers 312 local authorities over the period 1998–2005. They find that higher levels of $PM10$

(particulate matter less than 10 mm in diameter) and ozone (O_3) have a positive and significant effect on mortality rates. The OLS estimate for ($PM10/10$), controlling for three other measures of pollutants (carbon monoxide, nitrogen dioxide, and O_3), smoking rate, employment rate, etc. is 2.33, whereas that for FE is 2.74. The OLS estimate for ($O_3/10$) in the same regression is -0.55, whereas that for FE is 0.80. Only the FE estimates for these pollutants are significant at the 5% level.

One could test the joint significance of the individual effects, i.e., $H_0; \mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0$, by performing an F-test. This is a simple Chow test with the restricted residual sums of squares being that of OLS on the pooled model and the unrestricted residual sums of squares (URSS) being that which includes the $(N-1)$ individual dummies. By the Frisch–Waugh–Lovell theorem (Baltagi, 2008), URSS can be obtained from the within regression residual sum of squares. In this case

$$F_0 = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT-N-K)} \underset{\sim}{H_0} F_{N-1,N(T-1)-K} \qquad [2]$$

For the Contoyannis and Rice (2001) application, This F-statistic is 12.50 and is distributed under the null hypothesis as $F(858, 3406)$. This is significant and rejects H_0. One can infer that the OLS estimates are biased and inconsistent and yield misleading inference.

Difference-in-Differences

Note that the FE transformation ($\tilde{y}_{it} = y_{it} - \bar{y}_i$) is not the only transformation that will wipe out the individual effects. In fact, FD will also do the trick ($\Delta y_{it} = y_{it} - y_{i,t-1}$). This is a crucial tool used in the difference-in-differences (DID) estimator. Before the approval of any drug, it is necessary to assign patients randomly to receive the drug or a placebo and the drug is approved or disapproved depending on the difference in the health outcome between these two groups. In this case, the FDA is concerned with the drug's safety and its effectiveness. However, one runs into problems in setting this experiment. How can one hold other factors constant? Even twins which have been used in economic studies are not identical and may have different life experiences. With panel data, observations on the same subjects before and after a health policy change allow us to estimate the effectiveness of this policy on the treated and control groups without the contamination of individual effects. In simple regression form, assuming the assignment to the control and treatment groups is random, one regresses the change in the health outcome before and after the health policy is enacted on a dummy variable which takes the value 1 if the individual is in the affected (treatment) group and 0 if the individual is in the unaffected (control) group. This regression computes the average change in the health outcome for the treatment group before and after the policy change and subtracts that from the average change in the health outcome for the control group. One can include additional regressors which measure the individual characteristics before the policy change. Examples are gender, race, education, and age of the individual. This is known as the DID estimator in econometrics. Alternatively, one can regress the health outcome y on d_g d_t and their interaction $d_t \times d_g$. d_g is a dummy variable that takes the value 1 if the subject is in the treatment group, and 0 otherwise; d_t is a dummy variable which takes the value 1 for the posttreatment period, and 0 otherwise. In this case, $d_t \times d_g$ takes the value 1 only for observations in the treatment group and in the posttreatment period. The OLS estimate of the coefficient of $d_t \times d_g$ yields the DID estimator. Another advantage of running this regression is that one can robustify the standard errors with standard software.

In economics, one cannot conduct medical experiments. Card (1990) used a natural experiment to see whether immigration reduces wages. Taking advantage of the 'Mariel boatlift' where a large number of Cuban immigrants entered Miami, Card (1990) compared the change in wages of low-skilled workers in Miami with the change in wages of similar workers in other comparable US cities over the same period. Card concluded that the influx of Cuban immigrants had a negligible effect on wages of less-skilled workers. Gruber and Poterba (1994) use the DID estimator to show that a change in the tax law did increase the purchase of health insurance among the self-employed. They compared the fraction of the self-employed who had health insurance before the tax change 1985–86 with the period after the tax change 1988–89. The control group was the fraction of employed (not self-employed) workers with health insurance in those years.

Donald and Lang (2007) warn that the standard asymptotics for the DID estimator cannot be applied when the number of groups is small, as in the case where one compares two states in 2 years or self-employed workers and employees over a small number of years. They reconsider the Gruber and Poterba (1994) paper on health insurance and self-employment and Card's (1990) study of the Mariel boatlift. They show that analyzing the t-statistic, taking into account a possible group error component, dramatically reduces the precision of their results. In fact for Card's (1990) Mariel boatlift study, their findings suggest that the data cannot exclude large effects of the migration on blacks in Miami.

Bertrand *et al.* (2004) argued that several DID studies in economics rely on a long time series. They warn that in this case, serial correlation will understate the standard error of the estimated treatment effects, leading to overestimation of t-statistics and significance levels. They show that the block bootstrap (taking into account the autocorrelation of the data) works well when the number of states is large enough. Readers are advised to refer to Hansen (2007) for inference in panel models with serial correlation and FE and to Stock and Watson (2008) for a heteroskedasticity-robust variance matrix estimator for the FE estimator. Hausman and Kuersteiner (2008) warn that both the DID and the FE estimators are not efficient if the stochastic disturbances are serially correlated. The optimal estimator in this case is generalized least squares (GLS), but this is rarely used in applications of DID studies. Hausman and Kuersteiner (2008) use higher order Edgeworth expansion to construct a size-corrected t-statistic (based on feasible GLS) for the significance of treatment variables in DID regressions. They find that size-corrected t-statistic based on feasible GLS yields accurate size and is significantly more powerful than robust OLS when serial correlation in the level data is high.

Conley and Taber (2011) consider the case where there are only a small number N_1 of treatment groups, say states, that

change a law or policy within a fixed time span T. Let N_0 denote the number of control groups (states) that do not change their policy. Conley and Taber argue that the standard large-sample approximations used for inference can be misleading especially in the case of non-Gaussian or serially correlated errors. They suggest an alternative approach to inference under the assumption that N_1 is finite, using asymptotic approximations that let N_0 grow large, with T fixed. Point estimators of the treatment effect parameter(s) are not consistent as N_1 and T are fixed. However, they use information from the N_0 control groups to consistently estimate the distribution of these point estimators up to the true values of the parameter.

DID estimation has its benefits and limitations. It is simple to compute and it controls for heterogeneity of the individuals or the groups considered before and after the policy change. However, it does not account for the possible endogeneity of the interventions themselves (Besley and Case, 2000). Abadie (2005) discusses how well the comparison groups used in nonexperimental studies approximate appropriate control groups. Athey and Imbens (2006) critique the linearity assumptions used in DID estimation and provide a general changes-in-changes (CIC) estimator that does not require such assumptions.

The DID estimator requires that, in the absence of the treatment, the average outcomes for the treated and control groups would have followed parallel paths over time. This assumption may be too restrictive. Abadie (2005) considers the case in which differences in observed characteristics create nonparallel outcome dynamics between treated and controls. He proposes a family of semiparametric DID estimators which can be used to estimate the average effect of the treatment for the treated. Abadie *et al.* (2010) advocate the use of data-driven procedures to construct suitable comparison groups. Data-driven procedures reduce discretion in the choice of the comparison control units, forcing researchers to demonstrate the affinities between the affected and unaffected units using observed quantifiable characteristics. The idea behind the synthetic control approach is that a combination of units often provides a better comparison for the unit exposed to the intervention than any single unit alone. They apply the synthetic control method to study the effects of California's Proposition 99, a large-scale tobacco control program implemented in California in 1988. They demonstrate that following the passage of Proposition 99, tobacco consumption fell markedly in California relative to a comparable synthetic control region. They estimated that by the year 2000, annual per capita cigarette sales in California were approximately 26 packs lower than what they would have been in the absence of Proposition 99.

Athey and Imbens (2006) generalize the DID methodology to what they call the CIC methodology. Their approach allow the effects of both time and the treatment to differ systematically across individuals, as when new medical technology differentially benefits sicker patients. They propose an estimator for the entire counterfactual distribution of effects of the treatment on the treatment group as well as the distribution of effects of the treatment on the control group, where the two distributions may differ from each other in arbitrary ways. They provide conditions under which the proposed model is identified nonparametrically and extend the model to allow for discrete outcomes. They also provide extensions to settings with multiple groups and multiple time periods. They revisit the Meyer *et al.* (1995) study on the effects of disability insurance on injury durations. They show that the CIC approach leads to results that differ from the standard DID results in terms of magnitude and significance. They attribute this to the restrictive assumptions required for the standard DID methods.

Laporte and Windmeijer (2005) show that the FE and FD estimators lead to very different estimates of treatment effects when these are not constant over time, and treatment is a state that only changes occasionally. They suggest allowing for flexible time-varying treatment effects when estimating panel data models with binary indicator variables. They illustrate this by looking at the effect of divorce on mental well-being using the BHPS. They show that divorce has an adverse effect on mental well-being that starts before the actual divorce, peaks in the year of the divorce, and diminishes rapidly thereafter. A model that implies a constant instantaneous effect of divorce leads to very different FD and FE estimates, whereas a model that allows for flexibility in these effects lead to similar results. In general, the FE estimator is more efficient than the FD estimator when the remainder disturbance $v_{it} \sim \text{IID}(0,\sigma_v^2)$. The FD estimator is more efficient than the FE estimator when the remainder disturbance v_{it} is a random walk (Wooldridge, 2002). These estimators are affected differently by measurement error and by nonstationarity (Baltagi, 2008).

Certainly, this analysis can be refined to account for perhaps better control and treatment groups. If a policy is enacted by state s to reduce teenage smoking or motor vehicle fatality due to alcohol consumption or healthcare service for the elderly, then, for the two periods case, d_t takes the value 1 for the postpolicy period, and 0 otherwise; d_s takes the value 1 if the state has implemented this policy, and 0 otherwise; and d_g takes the value 1 for the treatment group affected by this policy like the elderly, and 0 otherwise. In this case, one regresses health-care outcome on d_t, d_s, d_g, $d_t \times d_g, d_t \times d_s, d_s \times d_g$ and $d_t \times d_s \times d_g$. The OLS estimate of the coefficient of $d_t \times d_s \times d_g$ yields the difference-in-difference-in-differences estimator of this policy. This estimator computes the average change in the health outcome for the elderly in the treatment state before and after the policy is implemented, and then subtracts from that the average change in the health outcome for the elderly in the control state, as well as the average change in the health outcome for the nonelderly in the treatment state.

Carpenter (2004) studied the effect of zero-tolerance (ZT) driving laws on alcohol-related behaviors of 18–20-year olds, controlling for macroeconomic conditions, other alcohol policies, state FE, survey year and month effects, and linear state-specific time trends. ZT Laws make it illegal for drivers under age of 21 years to have measurable amounts of alcohol in their blood, resulting in immediate license suspension and fines. Carpenter uses the Behavioral Risk Factor Surveillance System, which includes information on alcohol consumption and drunk driving behavior for young adults over the age of 18 years for the years 1984–2001. He estimates the effects of ZT Laws using the DID approach. The control group is

composed of individuals aged 22–24 years who are otherwise similar to treated individuals (18–20-year olds) but who should have been unaffected by the ZT policies. Let d_{ZT} be a dummy variable that takes the value 1 if the state has ZT in that year, and 0 otherwise; and d_g is a dummy variable that takes the value 1 if the subject is in the treatment group, and 0 otherwise. Alcohol consumption is regressed on d_{ZT}, d_{1820}, $d_{ZT} \times d_{1820}$, and other control variables mentioned above. The OLS estimate of the coefficient of $d_{ZT} \times d_{1820}$ yields the DID estimator of the ZT laws. Carpenter's results indicate that the laws reduced heavy episodic drinking (five or more drinks at one sitting) among underage males by 13%. For a recent review of DID health economics applications as well as a summary table of these applications, *see* Jones (2012).

Random Effects

There are too many parameters in the FE model and the loss of degrees of freedom can be avoided if μ_i can be assumed random. In this case, $\mu_i \sim \text{IID}(0,\sigma_\mu^2), v_{it} \sim \text{IID}(0,\sigma_v^2)$ and the μ_i is independent of the v_{it}. In addition, the X_{it} is independent of the μ_i and v_{it}, for all i and t. This random-effects (RE) model can be estimated by GLS, which can be obtained using a least squares regression of $y_{it}^* = y_{it} - \theta \bar{y}_{i.}$ on X_{it}^* similarly defined. $\theta = 1 - (\sigma_v/\sigma_1)$ where $\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2$. The best quadratic unbiased estimators of the variance components depend on the true disturbances, and these are minimum variance unbiased under normality of the disturbances. One can obtain feasible estimates of these variance components by replacing the true disturbances by OLS or FE residuals (*see* Chapter 2 of Baltagi (2008) for details).

Under the assumption of normality of the disturbances, Breusch and Pagan (1980) derived a Lagrange multiplier (LM) test to test H_0; $\sigma_\mu^2 = 0$. The resulting LM statistic requires only OLS residuals and is easy to compute. Under H_0, this LM statistic is asymptotically distributed as a χ_1^2 (*see* Chapter 4 of Baltagi (2008) for details.) For the Contoyannis and Rice (2001) application, this LM statistic is 3355.26 and is significant. This means that heterogeneity across individuals is significant and ignoring it as OLS does will lead to misleading inference. The RE estimates are 0.028 for excellent health, 0.013 for good health, and -0.002 for the General Health Questionnaire: Likert Scale score (hlghq1), with only the good health estimate being statistically insignificant.

Hausman Test

A specification test based on the difference between the FE and RE estimators is known as the Hausman test. The null hypothesis is that the individual effects are not correlated with the X'_{it}s. The basic idea behind this test is that the FE estimator $\tilde{\beta}_{FE}$ is consistent, whether or not the effects are correlated with the X'_{it}s. This is true because the within transformation $\tilde{y}_{it}$ wipes out the μ_i's from the model. However, if the null hypothesis is true, the FE estimator is not efficient under the RE specification because it relies only on the within variation in the data. However, the RE estimator $\hat{\beta}_{RE}$ is efficient under the null hypothesis but is biased and inconsistent when the effects are correlated with the X'_{it}s. The difference between these estimators $\hat{q} = \hat{\beta}_{FE} - \hat{\beta}_{RE}$ tends to zero in probability limits under the null hypothesis and is nonzero under the alternative. The variance of this difference is equal to the difference in variances, $\text{var}(\hat{q}) = \text{var}(\tilde{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})$, because $\text{cov}(\hat{q}, \hat{\beta}_{RE}) = 0$ under the null hypothesis. Hausman's test statistic is based on $m = \hat{q}'[\text{var}(\hat{q})]^{-1}\hat{q}$ and is asymptotically distributed as χ_K^2 under the null hypothesis. For the Contoyannis and Rice (2001) application, Hausman's test statistic is 322.39 and is distributed as χ_{29}^2. But the $\text{var}(\hat{q})$ is not positive definite. Using an alternative computation of this Hausman (1978) test based on an artificial regression, the null hypothesis is rejected and one can infer that the RE estimator is inconsistent and should not be used for inference.

Powell (2009) uses four waves of the 1997 National Longitudinal Survey of Youth and external data to examine the relationship between adolescent body mass index (BMI), fast food prices, and fast food restaurant availability. The OLS estimate of the fast food price elasticity of BMI is -0.095, whereas the RE estimate is -0.084. The latter is closer to the FE estimate of -0.078, but the RE estimator is rejected by the Hausman test. The number of fast food restaurants per capita was not found to be significant.

Hausman and Taylor Estimator

The RE model is rejected because it assumes no correlation between the explanatory variables and the individual effects. The FE estimator, however, assumes that all the explanatory variables are correlated with the individual effects. Instead of this 'all or nothing' correlation among the X and the μ_i, Hausman and Taylor (1981) consider a model where some of the explanatory variables are related to the μ_i. In particular, they consider the following model:

$$y_{it} = X'_{it}\beta + Z'_i\gamma + \mu_i + v_{it} \qquad [3]$$

where the Z_i is cross-sectional time-invariant variable. Hausman and Taylor (1981), hereafter HT, split X and Z into two sets of variables: $X = [X_1; X_2]$ and $Z = [Z_1; Z_2]$ where X_1 is $n \times k_1$, X_2 is $n \times k_2$, Z_1 is $n \times g_1$, Z_2 is $n \times g_2$, and $n = NT$. X_1 and Z_1 are assumed exogenous in that they are not correlated with μ_i and v_{it}, whereas X_2 and Z_2 are endogenous because they are correlated with μ_i, but not with v_{it}. The Within transformation sweeps the μ_i and removes the bias, but in the process it would also sweep the Z'_is and hence the Within estimator will not give an estimate of γ. To get around that, HT suggest obtaining the FE residuals and averaging them over time:

$$\hat{d}_i = \bar{y}_i - \bar{X}'_i\tilde{\beta}_{FE} \qquad [4]$$

Then, one can run 2SLS of $\hat{d}_i$ on Z_i with the set of instruments $A = [X_1, Z_1]$ to get a consistent estimate of γ which is called $\hat{\gamma}_{2SLS}$. For this to be feasible, the order condition for identification has to hold ($k_1 \geq g_2$). This means that there has to be as many time-varying (X_1) exogenous variables as there are time-invariant endogenous variables (Z_2). The intuition here is that every X_{it} can be written as the sum of $\tilde{X}_{it} = (X_{it} - \bar{X}_i)$ and $\bar{X}_i$. It is the latter term that contains μ_i as it is swept away from the former. If X_2 is correlated with μ_i, it must be in $\bar{X}_2$, which makes $\tilde{X}_2$ the ideal instrument. HT use X_1 twice because it is exogenous, once as $\tilde{X}_1$ and another time as $\bar{X}_1$. Z_1

is exogenous and Z_2 can be instrumented by the additional instruments gained from X_1. With consistent estimates of the disturbances obtained from $\tilde{\beta}_{FE}$ and $\hat{\gamma}_{2SLS}$, one can obtain consistent estimates of the variance components and hence θ. This, in turn, allows us to compute $y_{it}^* = y_{it} - \theta\bar{y}_i$ and X_{it}^* and $Z^* = (1-\theta)Z$. HT suggest an efficient estimator that can be obtained by running 2SLS of y_{it}^* on X_{it}^* and Z^* using $A_{HT} = [\tilde{X}, \bar{X}_1, Z_1]$ as instruments.

1. If $k_1 < g_2$, then the equation is underidentified. In this case, $\hat{\beta}_{HT} = \tilde{\beta}_{FE}$ and γ cannot be estimated.
2. If $k_1 = g_2$, then the equation is just-identified. In this case, $\hat{\beta}_{HT} = \tilde{\beta}_{FE}$ and $\hat{\gamma}_{HT} = \hat{\gamma}_{2SLS}$.
3. If $k_1 > g_2$, then the equation is over-identified and the HT estimator is more efficient than the FE estimator.

A test for over-identification is obtained by computing

$$\hat{m}_2 = \hat{q}_2'[\text{var}(\tilde{\beta}_{FE}) - \text{var}(\hat{\beta}_{HT})]^{-}\hat{q}_2 \qquad [5]$$

with $\hat{q}_2 = \tilde{\beta}_{FE} - \hat{\beta}_{HT}$ and $\hat{\sigma}_v^2 \hat{m} \overset{H_0}{\rightarrow} \chi^2_{k_1 - g_2}$.

Contoyannis and Rice (2001) applied the HT estimator, choosing race to be exogenous (the only time-invariant Z_1) and education to be endogenous (the only time-invariant Z_2). They also chose the health variables that are time varying to be endogenous (sahex, sahgd, and hlghq1) as well as (prof, manag, skllnm, and skllm). The HT estimates are 0.013 for excellent health, 0.010 for good health, and -0.003 for the General Health Questionnaire: Likert Scale score (hlghq1), with only the latter estimate being statistically significant.

Dynamic Panel Data Models

Many economic relationships are dynamic in nature and one of the advantages of panel data is that they allow the researcher to better understand the dynamics of adjustment. For example, a key feature of the rational addiction theory studied by Becker, Grossman, and Murphy (1994) is that consumption of cigarettes is addictive and will depend on future as well as past consumption. Consumers are rational if they are forward-looking in the sense that they anticipate the expected future consequences of their current actions. They recognize the addictive nature of their choices but they may elect to make them because the gains from the activity exceed the costs through future addiction. The more they smoke the higher is the current utility derived. However, the individual recognizes that he or she is building up a stock of this addictive good that is harmful. The individual rationally trades off these factors to determine the appropriate level of smoking. Finding future consumption statistically significant is a rejection of the myopic model of consumption behavior. In the latter model of addictive behavior, only past consumption stimulates current consumption, because individuals ignore the future in making their consumption decisions.

More formally, dynamic relationships are characterized by the presence of a lagged dependent variable among the regressors, i.e.,

$$y_{it} = \delta y_{i,t-1} + x_{it}'\beta + \mu_i + v_{it} \quad i = 1,\ldots,N \quad t = 1,\ldots,T \qquad [6]$$

where δ is a scalar, x'_{it} is $1 \times K$ and β is $K \times 1$, where $\mu_i \sim \text{IID}(0,\sigma_\mu^2)$ and $v_{it} \sim \text{IID}(0,\sigma_v^2)$ independent of each other and among themselves. This dynamic panel data regression model is characterized by two sources of persistence over time. Autocorrelation due to the presence of a lagged dependent variable among the regressors and individual effects characterizing the heterogeneity among the individuals. As y_{it} is a function of μ_i, it immediately follows that $y_{i,t-1}$ is also a function of μ_i, Therefore, $y_{i,t-1}$ is correlated with the error term. This renders the OLS estimator biased and inconsistent even if the v_{it} are not serially correlated. For the FE estimator, the Within transformation wipes out the μ_i, but $(y_{i,t-1} - \bar{y}_{i-1})$ where $\bar{y}_{i-1} = \sum_{t=2}^{T} y_{i,t-1}/(T-1)$ will still be correlated with $(v_{it} - \bar{v}_{i.})$ even if the v_{it} are not serially correlated. This is because $y_{i,t-1}$ is correlated with $\bar{v}_{i.}$ by construction. The latter average contains $v_{i,t-1}$ which is obviously correlated with $y_{i,t-1}$. In fact, the Within estimator will be biased of $O(1/T)$ and its consistency will depend on T being large (Nickell, 1981). Therefore, for the typical micropanel where N is large and T is fixed, the Within estimator is biased and inconsistent. It is worth emphasizing that only if $T \rightarrow \infty$ will the Within estimator of δ and β be consistent for the dynamic error component model. For macropanels, some researchers may still favor the Within estimator arguing that its bias may not be large. Judson and Owen (1999) performed some Monte Carlo experiments for $N = 20$ or 100 and $T = 5$, 10, 20, and 30 and found that the bias in the Within estimator can be sizable, even when $T = 30$. This bias increases with δ and decreases with T. But even for $T = 30$, this bias could be as much as 20% of the true value of the coefficient of interest.

Arellano and Bond (1991) suggested FD model to get rid of the μ_i and then using a Generalized Method of Moments (GMM) procedure that utilizes the orthogonality conditions that exist between lagged values of y_{it} and the disturbances v_{it}. It is illustrated with the simple autoregressive model with no regressors. With a three-wave panel, i.e., $T = 3$, the differenced equation becomes:

$$y_{i3} - y_{i2} = \delta(y_{i2} - y_{i1}) + (v_{i3} - v_{i2})$$

In this case, y_{i1} is a valid instrument because it is highly correlated with $(y_{i2} - y_{i1})$ and not correlated with $(v_{i3} - v_{i2})$ as long as the v_{it} are not serially correlated. But note what happens if the fourth wave is obtained:

$$y_{i4} - y_{i3} = \delta(y_{i3} - y_{i2}) + (v_{i4} - v_{i3})$$

In this case, y_{i2} as well as y_{i1} are valid instruments for $(y_{i3} - y_{i2})$ because both y_{i2} and y_{i1} are not correlated with $(v_{i4} - v_{i3})$. One can continue in this fashion, adding an extra valid instrument with each forward period, so that for period T, the set of valid instruments becomes $(y_{i1}, y_{i2}, \ldots, y_{i,T-2})$. The optimal Arellano and Bond (1991) GMM estimator of δ utilizes all these moment conditions weighting them by a sandwich heteroskedasticity auto-correlation estimator of the variance–covariance matrix of the disturbances. Arellano and Bond (1991) propose testing for serial correlation for the disturbances of the first-differenced equation. This test is important because the consistency of the GMM estimator relies on the assumption of no serial correlation in the v'_{it}s. Additionally, Arellano and Bond (1991) suggest a Sargan test for over-identifying. One has to reject the existence of serial correlation in the v'_{it}s and not reject the over-identifying

restrictions. Failing these diagnostics renders this procedure inconsistent.

Using Monte Carlo experiments, Bowsher (2002) finds that the use of too many moment conditions causes the Sargan test for overidentifying restrictions to be undersized and have extremely low power. The Sargan test never rejects when T is too large for a given N. Zero rejection rates under the null and alternative were observed for the following (N,T) pairs (125,16), (85,13), and (40,10). This is attributed to poor estimates of the weighting matrix in GMM. Using Monte Carlo experiments, Ziliak (1997) found that there was a bias/efficiency trade-off for the Arellano and Bond (1991) GMM estimator as the number of moment conditions increase and that one is better off with suboptimal instruments. Ziliak attributes the bias in GMM to the correlation between the sample moments used in estimation and the estimated weight matrix.

Blundell and Bond (1998) attributed the bias and the poor precision of the first difference GMM estimator to the problem of weak instruments. They show that an additional mild stationarity restriction on the initial conditions process allows the use of a system GMM estimator which captures additional nonlinear moment conditions that are ignored by the Arellano and Bond (1991) estimator. These additional nonlinear moment conditions are described in Ahn and Schmidt (1995) and can be linearized by adding a set of equations in levels on top of the set of equations in first differences of Arellano and Bond, hence a system of equations (*see* Baltagi, 2008, Chapter 8, for details). In this case, one uses lagged differences of y_{it} as instruments for equations in levels, in addition to lagged levels of y_{it} as instruments for equations in first differences. The system GMM estimator is shown to have dramatic efficiency gains over the basic first-difference Arellano and Bond GMM estimator as $\delta \to 1$, i.e., as the process tends to unit root and nonstationarity.

Baltagi *et al.* (2000) estimate a dynamic demand model for cigarettes based on panel data from 46 American states over the period 1963–92. The estimated equation is:

$$lnC_{it} = \alpha + \beta_1 lnC_{i,t-1} + \beta_2 lnP_{i,t} + \beta_3 lnY_{it} + \beta_4 lnPn_{it} + \mu_i + \lambda_t + \nu_{it} \quad [7]$$

where the subscript i denotes the i-th state ($i=1, \ldots, 46$), and the subscript t denotes the t-th year ($t = 1,\ldots,30$). C_{it} is real per capita sales of cigarettes by persons of smoking age (14 years and older). This is measured in packs of cigarettes per head. P_{it} is the average retail price of a pack of cigarettes measured in real terms. Y_{it} is real per capita disposable income. Pn_{it} denotes the minimum real price of cigarettes in any neighboring state. This last variable is a proxy for the casual smuggling effect across state borders. μ_i denotes the state-specific effects, and λ_t denotes the year-specific effects. OLS, which ignores the state and time effects, yields a low short-run price elasticity of -0.09. However, the coefficient of lagged consumption is 0.97 which implies a high long-run price elasticity of -2.98. The FE estimator with both state and time effects yields a higher short-run price elasticity of -0.30, but a lower long-run price elasticity of -1.79. Both state and time dummies were jointly significant with an observed F-statistic of 7.39 and a p-value of .0001. This is a dynamic equation and the OLS and FE estimators do not take into account the endogeneity of the lagged dependent variable. The Arellano and Bond (1991) GMM estimator yields a lagged consumption coefficient estimate of 0.70 and an own price elasticity of -0.40, both highly significant (Baltagi, 2008). The two-step Sargan test for over-identification does not reject the null, but this could be due to the bad power of this test for $N=46$ and $T=28$. The test for first-order serial correlation rejects the null of no first-order serial correlation, but it does not reject the null that there is no second-order serial correlation. This is what one expects in a first-differenced equation with the original untransformed disturbances assumed to be not serially correlated. Blundell and Bond (1998) system GMM estimator yields a lagged consumption coefficient estimate of 0.70 and an own price elasticity of -0.42, both highly significant, but with higher standard errors than the corresponding Arellano and Bond estimators. Sargan's test for over-identification does not reject the null, and the tests for first- and second-order serial correlation yield the expected diagnostics for system GMM.

Scott and Coote (2010) applied the dynamic panel data system GMM estimator to estimate the effect of regional primary-care organizations on primary-care performance. They utilize a panel of 119 Divisions of General Practice in Australia observed quarterly over the period 2000–05. Using four different measures of primary-care performance, a high level of persistence was found. The results show that Divisions were more likely to influence general practice infrastructure than clinical performance in diabetes, asthma, and cervical screening. Other applications of dynamic panel data GMM estimation methods include Baltagi and Griffin (2001) to a rational addiction model of cigarettes and Suhrcke and Urban (2010) to the impact of cardiovascular disease morality on economic growth.

Ng *et al.* (2012) study the relative importance of diet, physical activity, and health behavior of smoking and drinking on weight for a set of Chinese males, using panel data from the China Health and Nutrition Survey. The authors use a dynamic panel system GMM approach that explicitly includes time and spatially varying community-level urban city and price measures as instruments, to obtain estimates for the effects of diet, physical activity, drinking, and smoking on weight. Results show that approximately 5.4% of weight gain is due to declines in physical activity and 2.8–3.1% is due to dietary changes over time.

Limited Dependent Variable Panel Data Models

In some health studies, the dependent variable is binary. For example, individual i may be in good health at time t, i.e., $y_{it}=1$ with probability p_{it}, or in bad health, $y_{it}=0$ with probability $1-p_{it}$. Good health occurs when a latent unobserved index of health y^*_{it} is positive

$$\begin{aligned} y_{it} &= 1 \quad \text{if } y^*_{it} > 0 \\ &= 0 \quad \text{if } y^*_{it} \le 0 \end{aligned} \quad [8]$$

with $y^*_{it} = x'_{it}\beta + \mu_i + \nu_{it}$. So that

$$\Pr[y_{it}=1] = \Pr[y^*_{it} > 0] = \Pr[\nu_{it} > -x'_{it}\beta - \mu_i] = F(x'_{it}\beta + \mu_i) \quad [9]$$

where the last equality holds as long as the density function describing F is symmetric around zero. This is true for the logistic and normal density functions which are the most used in practice. This is a nonlinear panel data model because F is a cumulative density function, and one cannot get rid of the individual effects as in the linear panel data case with a within transformation. Hsiao, 2003 showed that unlike the linear FE panel data case, where the inconsistency of the μ'_is did not transmit into inconsistency for the β's. For the nonlinear panel data case, the inconsistency of the μ'_is renders the maximum likelihood estimates of the β's inconsistent. The usual solution around this incidental parameters problem is to assume a logistic function and to condition on $\sum_{t=1}^{T} y_{it}$, which is a minimum sufficient statistic for μ_i maximizing the conditional logistic likelihood function

$$L_c = \prod_{i=1}^{N} Pr\left(y_{i1}, \ldots, y_{iT} / \sum_{t=1}^{T} y_{it}\right) \qquad [10]$$

yields the conditional logit estimates for β. By definition of a sufficient statistic, the distribution of the data given this sufficient statistic will not depend on μ_i. In contrast to the FE logit model, the conditional likelihood approach does not yield computational simplifications for the FE probit model. But the probit specification has been popular for the RE model. In this case, $u_{it} = \mu_i + v_{it}$ where $\mu_I \sim \text{IIN}(0,\sigma_\mu^2)$ and $v_{it} \sim \text{IIN}(0,\sigma_v^2)$ independent of each other and the x_{it}. Because $E(u_{it}u_{is}) = \sigma_\mu^2$ for $t \neq s$, the joint likelihood of $(y_{1t}, \ldots, y_{Nt})$ can no longer be written as the product of the marginal likelihoods of the y_{it}. This complicates the derivation of maximum likelihood which will now involve T-dimensional integrals. Fortunately, Butler and Moffitt (1982) showed that for the probit case, the maximum likelihood computations involve only one integral which can be evaluated using the Gaussian–Hermite quadrature procedure. For an early application of the RE probit model, Sickles and Taubman (1986), who estimated a two-equation structural model of the health and retirement decisions of the elderly using five biennial panels of males drawn from the Retirement History Survey. Both the health and retirement variables were limited dependent variables and MLE using the Butler and Moffitt (1982) Gaussian quadrature procedure was implemented. Sickles and Taubman found that retirement decisions were strongly affected by health status and workers not yet eligible for social security were less likely to retire.

Contoyannis *et al.* (2004) utilize seven waves (1991–97) of the BHPS to analyze the dynamics of individual health and to decompose the persistence in health outcomes in the BHPS data into components due to state dependence, serial correlation, and unobserved heterogeneity. The indicator of health is defined by a binary response to the question: "Does your health in any way limit your daily activities compared to most people of your age?" A sample of 6106 individuals resulting in 42 742 panel observations are used to estimate static and dynamic panel probit models by maximum simulated likelihood methods. The dynamic models show strong positive state dependence.

Hernández-Quevedoa *et al.* (2008) use eight waves of the ECHP over the period 1994–2001 to estimate a dynamic nonlinear panel data model of health limitations for individuals within the Member States of the European Union. The RE probit specification conditions on previous health status and parameterizes the unobserved individual effect as a function of initial period observations on time-varying regressors and health (Wooldridge, 2005). Results reveal high state dependence of health limitations, which remains after controlling for measures of socioeconomic status. There is also heterogeneity in the socioeconomic gradient across countries. The importance of regarding health as a dynamic concept has implications for policy development. They imply that medical interventions or health improvement policies that create health gains, will have multiplier effects in the long run.

Further Readings

The panel data econometrics literature has exhibited phenomenal growth and one cannot do justice to the many theoretical contributions to date. Space limitations prevented the inclusion of many worthy topics including attrition, sample selection, semiparametric, nonparametric, and Bayesian methods using panel data. Unbalanced panels, problems associated with heteroskedasticity, serial as well as spatial correlation in panels, measurement error, duration, and quantile panel data models to mention a few. More extensive treatment of these and other topics are given in textbooks on the subject by Baltagi (2008), Wooldridge (2002), and Hsiao (2003). Also see the survey by Imbens and Wooldridge (2009) for an extensive discussion of alternative econometric methods to program evaluation besides the DID method. Also see Angrist and Pischke (2009) for a textbook discussion of DID. For recent applications of panel data methods to health economics, see the special issue of *Empirical Economics* edited by Baltagi *et al.* (2012) and Jones (2012).

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies* **72**, 1–19.

Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* **105**, 493–505.

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* **21**, 489–519.

Ahn, S. C. and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* **68**, 5–27.

Angrist, J. D. and Pischke, J-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton: Princeton University Press.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277–297.

Askildsen, J. E., Baltagi, B. H. and Holmas, T. H. (2003). Wage policy in the health care sector: A panel data analysis of nurses' labour supply. *Health Economics* **12**, 705–719.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74**, 431–497.

Baltagi, B. H. (2008). *The econometrics of panel data.* Chichester: Wiley.

Baltagi, B. H. and Geishecker, I. (2006). Rational alcohol addiction: Evidence from the Russian longitudinal monitoring survey. *Health Economics* **15**, 893–914.

Baltagi, B. H. and Griffin, J. M. (2001). The econometrics of rational addiction: The case of cigarettes. *Journal of Business and Economic Statistics* **19**, 449–454.

Baltagi, B. H., Griffin, J. M. and Xiong, W. (2000). To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand. *Review of Economics and Statistics* **82**, 117–126.
Baltagi, B. H., Jones, A. M., Moscone, F. and Mullahy, J. (2012). Special issue on health econometrics: Editors' introduction. *Empirical Economics* **42**, 365–368.
Baltagi, B. H. and Moscone, F. (2010). Health care expenditure and income in the OECD reconsidered: Evidence from panel data. *Economic Modelling* **27**, 804–811.
Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 249–275.
Besley, T. and Case, A. (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *Economic Journal* **110**, F672–F694.
Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115–143.
Bowsher, C. G. (2002). On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* **77**, 211–220.
Breusch, T. S. and Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* **47**, 239–253.
Butler, J. S. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* **50**, 761–764.
Card, D. (1990). The impact of the Mariel boat lift on the Miami labor market. *Industrial and Labor Relations Review* **43**, 245–253.
Carpenter, C. (2004). How do zero tolerance drunk driving laws work? *Journal of Health Economics* **23**, 61–83.
Conley, T. G. and Taber, C. R. (2011). Inference with "Difference-in-Differences" with a small number of policy changes. *Review of Economics and Statistics* **93**, 113–125.
Contoyannis, P., Jones, A. M. and Rice, N. (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics* **19**, 473–503.
Contoyannis, P. and Rice, N. (2001). The impact of health on wages: Evidence from the British Household Panel Survey. *Empirical Economics* **26**, 599–622.
Donald, S. G. and Lang, K. (2007). Inference with difference in differences and other panel data. *Review of Economics and Statistics* **89**, 221–233.
Greene, W. (2010). Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* **13**, 959–980.
Gruber, J. and Poterba, J. (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Quarterly Journal of Economics* **109**, 701–734.
Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics* **140**, 670–694.
Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251–1271.
Hausman, J. A. and Kuersteiner, G. (2008). Difference in difference meets generalized least squares: Higher order properties of hypotheses tests. *Journal of Econometrics* **144**, 371–391.
Hausman, J. A. and Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica* **49**, 1377–1398.
Hernández-Quevedoa, C., Jones, A. M. and Rice, N. (2008). Persistence in health limitations: A European comparative analysis. *Journal of Health Economics* **27**, 1472–1488.
Hsiao, C. (2003). *Analysis of panel data*. Cambridge: Cambridge University Press.
Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**, 5–86.
Janke, K., Propper, C. and Henderson, J. (2009). Do current levels of air pollution kill? The impact of air pollution on population mortality in England. *Health Economics* **18**, 1031–1055.
Jones, A. M. (2012). Panel data methods and applications to health economics. In Terence, C., Mills and Patterson, Kerry (eds.) *Handbook of econometrics volume II: Applied econometrics*, pp. 557–631. Basingstoke: Palgrave MacMillan.
Judson, R. A. and Owen, A. L. (1999). Estimating dynamic panel data models: A guide for macroeconomists. *Economics Letters* **65**, 9–15.
Laporte, A. and Windmeijer, F. (2005). Estimation of panel data models with binary indicators when treatment effects are not constant over time. *Economics Letters* **88**, 389–396.
Meyer, B., Viscusi, K. and Durbin, D. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review* **85**, 322–340.
Ng, S. W., Norton, E. C., Guilkey, D. K. and Popkin, B. M. (2012). Estimation of a dynamic model of weight. *Empirical Economics* **42**, 413–433.
Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica* **49**, 1417–1426.
Powell, L. M. (2009). Fast food costs and adolescent body mass index: Evidence from panel data. *Journal of Health Economics* **28**, 963–970.
Ruhm, C. J. (1996). Alcohol policies and highway vehicle fatalities. *Journal of Health Economics* **15**, 435–454.
Sickles, R. C. and Taubman, P. (1986). A multivariate error components analysis of the health and retirement study of the elderly. *Econometrica* **54**, 1339–1356.
Stock, J. H. and Watson, M. W. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* **76**, 155–174.
Suhrcke, M. and Urban, D. (2010). Are cardiovascular diseases bad for economic growth? *Health Economics* **19**, 1478–1496.
Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*, XXIII, pp. 752. Cambridge, Mass: MIT Press.
Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**, 39–54.
Ziliak, J. P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics* **15**, 419–431.

Patents and Other Incentives for Pharmaceutical Innovation

PV Grootendorst and A Edwards, University of Toronto, Toronto, ON, Canada
A Hollis, University of Calgary, Calgary, AB, Canada

Introduction

Innovation – the discovery of ways to get more value from limited resources – is critically important for society's health and material standard of living. Despite the salience of innovation, there is no consensus on how much investment in innovation there should be, or how to ensure that the optimal investments are made, or how much the public sector should pay. These issues are particularly contentious in the pharmaceutical industry, given the high cost of drug development and the extraordinary importance of pharmaceuticals for human health.

Pharmaceutical research and development (R&D) takes place within both not-for-profit and for-profit organizations. The not-for-profit sector – comprising both academic and public sector labs – tends to focus on basic research, whereas for-profit drug companies exploit this basic research, as well as their own research, to develop, test, and market new drugs. Public support of basic research is usually by way of grants, administered by charitable foundations and public agencies, such as the Wellcome Trust in the UK and the National Institutes of Health in the US. Government support for the subsequent steps in the drug development process is composed of three core approaches: (1) tax subsidies for clinical trials and other R&D costs; (2) patents and other forms of intellectual property (IP) protection that provides the innovator with an exclusivity period in sales of newly developed products; and (3) prescription drug subsidies (often extended to seniors and those with high drug costs relative to income). These policies increase the profits to firms that undertake R&D, either by decreasing R&D costs (in the case of tax subsidies) or increasing revenues earned on the sale of new drugs (prescription drug subsidies, IP protection). Policies that decrease firms' costs of drug development are commonly referred to as 'push' incentives, whereas policies that increase the revenues accruing to firms that manage to bring new drugs to market are termed 'pull' incentives.

During the past decade or so, there has been considerable interest by academics, activists, politicians, and others in the manner in which governments support pharmaceutical R&D undertaken by for-profit firms. Of particular concern is the declining productivity of the pharmaceutical R&D sector. Recent academic studies place the capitalized cost of drug development to be between US$1.5 and US$1.8 billion. Other estimates are even higher. Different types of push and pull incentives, it is argued, would yield more therapeutically novel drugs per dollar of public support. But although there is agreement that reform is needed, there is less agreement over what should be done to improve matters.

This article reviews the contours of the academic debate, focusing on the advantages and deficiencies of existing forms of government support for drug R&D, and the features of the alternative arrangements that their proponents suggest will improve in the current system.

Intellectual Property: Advantages and Disadvantages

Much of the analysis of the defects of current arrangements focus on the system of IP protection afforded to new drugs. This is not surprising, considering that IP privileges constitute one of the most expensive of all the forms of public support extended to the pharmaceutical sector.

There are two distinct kinds of IP currently available for pharmaceuticals: patents and data exclusivity. A patent provides its owner with a 20-year period of exclusive use of the invention disclosed in the patent. Typically, drugs will be protected by a number of patents, each of which may have a number of claims disclosing distinct inventions. Because the innovation process is long, some of the relevant patents are filed many years before the drug comes to market. The result is that the average period of exclusivity owing to patents is roughly 10 years.

Data exclusivity is another important IP tool. The patent protects the invention, and can therefore protect the drug if it embodies the invention and there is no way of circumventing the patent. Data exclusivity protects the data produced by the innovator for the purpose of obtaining regulatory approval. Regulators such as the US Food and Drug Administration (FDA) require extensive testing of drugs, and running the clinical trials to obtain data that will demonstrate safety and effectiveness is extremely costly – typically approximately half of the total cost of drug development. During the term of data exclusivity, which begins after the drug receives regulatory approval and varies by country from five to ten years, no generic drug is approvable based on a reference to the clinical trial data of the innovator's drug.

IP allows the innovator to earn more sales revenue than would be possible without exclusivity. The variable costs of production and distribution of drugs are typically only a small fraction of the brand drug price, ranging approximately 5–25%. After covering other costs, such as marketing and administration, the remainder (often called the mark-up or the margin) can be paid as dividends or reinvested into R&D. IP thus creates very powerful incentives to develop therapeutically novel drugs (among other innovations). The chief appeal of this system of market exclusivity is that it automatically generates a relationship between the reward to the innovator (the mark-up times, the volume of sales) and the value of the innovation to society, as Adam Smith himself noted. More therapeutically valuable drugs are expected to earn greater sales revenues. IP, however, also has some drawbacks, five of which are focused in the following sections.

Drug R&D Costs

The first drawback is that the IP system can increase the cost of drug R&D for two reasons. Basic research conducted in academic or public sector labs will occasionally identify cellular

 doi:10.1016/B978-0-12-375678-7.01219-0

proteins, known as 'targets,' implicated in disease pathways. Multiple companies will then attempt to develop drugs that act on these targets. But drug development is very difficult because human biology is very complicated. As a result, many drugs simply do not work as expected and many of those that successfully disrupt the disease pathway either show no therapeutic benefit or do so only at doses that are toxic. It is also difficult to deliver an agent to its target in sufficient concentration and for an appropriate duration to achieve the desired therapeutic effect.

Often the viability (or otherwise) of a drug candidate or a disease mechanism becomes apparent only after much time and money have been spent. Ideally, drug companies would share this information. Sharing would reduce duplication of R&D costs, would eliminate unnecessary experimentation on human subjects, and would advance the understanding of human pathophysiology and pharmacology. But commercial drug R&D has historically been conducted in a culture of secrecy. Some of this secrecy, no doubt, is due to normal competitive behavior, present to varying degrees in all technology-intensive industries. But pharmaceutical companies are hesitant to share information, at least before patent filing, owing to the risk that this information may be used by a competitor that is developing patents in the same area. Such a competitor may preemptively patent a class of molecules with therapeutic promise, or worse, attempt to patent the target or pathway itself.

Of course, patents, once filed, publicly disclose the claimed inventions, so they do help to disseminate knowledge. Also the 'first to file' rule for patents gives incentives to patent, and therefore disclose, early. And companies routinely report scientific and clinical progress at academic and medical meetings. But the information that is disclosed in patents or at meetings is typically incomplete, and certainly does not prevent multiple drug companies from pursuing leads that are known by other competitors to be dead ends.

IP can hinder drug development in another way. Many research inputs, such as disease-linked human genes and techniques to manipulate deoxyribonuclic acid and proteins, are patented. Innovating firms must therefore conduct R&D cognizant of the landscape of existing patents. One way is to conduct R&D in ways that do not infringe on existing patents. But if it is not possible (or prohibitively costly) to use a circuitous technique, the firm must anticipate the threat of legal action by patent holders. One way to deal with such threats is to pay licensing fees, assuming that the entrant can strike a mutually beneficial deal with possibly numerous patent holders. Another approach is to wait until relevant patents have expired. The potential entrant might also mount a legal challenge to the validity of patents perceived as being weak. Yet another tactic is to amass a portfolio of patents so that the firm can credibly threaten to counter-sue for infringement of some of its own patents. Each of these strategies can increase the costs of research substantially.

'Follow-on' Drugs

A successful new 'first-in-class' drug will often face competition from a series of 'me-too' or 'follow-on' drugs that are therapeutically similar to the first-in-class drug. Often, follow-on drugs are simply the natural outcome of simultaneous research programs into the same therapeutic target. In other cases, they are the result of an intentionally imitative research program. The angiotensin converting enzyme (ACE) inhibitors, a class of drugs routinely used to manage high blood pressure, is illustrative of this. The first ACE inhibitor, captopril, was introduced in the US in 1981. Since then, over 10 ACE inhibitors have been launched, the last one in the mid-1990s. It appears that some of these later arriving ACE inhibitors were launched to capitalize on the commercial success (and clinically proven mechanism of action) of the earlier ACE inhibitors.

The proliferation of follow-on drugs is the subject of some debate. Proponents note that some follow-on drugs are therapeutically superior to the pioneer. Indeed, this appears to be the case for the ACE inhibitors. Moreover, if patients respond idiosyncratically to any one in a group of similar drugs, it is very useful to have alternatives. But the threat of imitative drug development also reduces the incentive for firms to develop first-in-class drugs. The reason is that follow-on drugs decrease the expected sales revenues and increase the costs of the pioneer. Lichtenberg and Philipson show that competition from follow-on drugs decreases the pioneer drug's revenues more than the competition from generics after patent expiry, in large part because the competition from follow-on drugs occurs early in the life of the pioneer drug. Costs increase because the pioneer firm typically spends on marketing and promotion to defend market share from capture by competitive products. Follow-on drugs may therefore dull the incentive to develop first-in-class drugs. As evidence of this, critics point to the protein kinases; these cellular proteins represent the most common targets for drug discovery. However, although there are 518 protein kinases in the human genome, more than half the current drug discovery programs focus on the handful of kinases for which there is already an existing drug.

A related criticism concerns the large outlays on the marketing and promotion of branded pharmaceutical drugs, more generally. Estimates of promotional expenditures are somewhat uncertain, but data compiled by Gagnon and Lexchin suggest that in 2004 the US pharmaceutical industry spent at least as much on promotion as it did on R&D. (Gagnon and Lexchin's estimate includes two particularly notable components of promotional cost: They include samples, representing 27.7% of total promotional dollar value; and estimated 'unmonitored' expenditures, representing 25%. The samples are valued at the retail prices, which are probably on average at least 10 times the cost of manufacture; and the estimate of unmonitored expenditure is highly uncertain. Removing these components reduces promotional expenditures to an amount closer to the amount spent on R&D.) Economic theory predicts that firms will continue to spend on promotion as long as the last dollar spent results in a compensatory increase in unit sales and gross profits. Thus IP, to the extent that it increases the margins earned on unit sales, encourages promotion.

No doubt some, perhaps the majority, of this promotion is socially valuable, alerting consumers about the availability of effective therapies for an untreated health condition, or providing prescribers with information on the properties of new

drugs. However, critics charge that a significant portion is more persuasive than informative, and in some cases is misleading. This kind of advertising is socially wasteful, as it represents a zero-sum competition between firms for market share; in the worst cases it might lead to worse health outcomes if clinicians are persuaded to prescribe drugs that are unnecessary or inferior to other therapies.

Drug Pricing

In most markets, consumer willingness and ability to pay is an important constraint on the price charged. Pharmaceutical markets are different. Most consumers in developed countries have insurance that covers some or all of the cost of prescribed drugs. If consumers do not pay for their drugs, what constrains prices? The answer, of course, is that insurers set limits on what they will pay. Indeed, many drug plans wield substantial bargaining power on account of their large size. France, UK, and Australia, for instance, operate national drug plans that account for the majority of drug sales. Federal states without national drug plans typically have large-scale plans operated by regional governments. These plans increasingly exploit their negotiating power to extract price concessions from drug companies that wish to have their products listed on the drug plan formulary. These price concessions directly reduce the margins that are ostensibly there to recoup R&D costs. Negotiation costs and costs of applying for formulary listing can also be substantial; Cohen reports that drug manufacturers often need to contract with hundreds of different drug plans in the US.

The price concessions are sometimes directly negotiated with insurers. For instance, the public drug plan operating in the province of Ontario, Canada, extracts confidential discounts off list prices. Other insurers set a maximum price that they are willing to pay, not for tablets or pills, but for expected units of health generated by the use of a new drug. These health units are usually denominated in 'quality-adjusted life-years' (or 'QALYs') – which measure both survival and quality of life gains. Typically, insurers' willingness to pay for a QALY are well below consumers' expressed willingness to pay. The UK National Institute for Health and Clinical Excellence (NICE), for instance, uses a threshold of £20 000 to £30 000 per QALY. Consensus estimates of consumers' valuation of a life year in normal health in the US are closer to US$100 000 (~£65 000).

The UK's NICE is perhaps the most well known national health technology assessment body, but it is not the only such initiative. Other countries that have created such agencies include Australia (Pharmaceutical Benefits Advisory Committee), Canada (Common Drug Review), Scotland (Scottish Medicines Consortium), Sweden (TLV), and Germany (IQWiG). Although the US has no national-level body, Cohen notes that many insurers informally consider cost effectiveness when making formulary decisions.

One reason that public insurers are willing to pay less than consumers is that public health-care budgets are constrained. The relevant consideration for NICE is its opportunity cost: if the threshold payment per QALY is too high, it will displace other, more cost-effective nonpharmaceutical interventions. A relatively low willingness to pay for drugs is thus appropriate if the public health authority wishes to maximize health given a constrained budget, although this raises questions about the adequacy of the health budget.

The use of QALY assessments reflects a growing tendency on the part of insurers to assess value for money when considering whether, and under what conditions, they will reimburse a drug. The insurers that used to cover almost all new drugs that received regulatory approval are now becoming much more selective consumers. An innovator, of course, can elect to forgo formulary listing in a given insurance plan; beneficiaries always have the option of paying cash for drugs that are not insured. But if the insurer has a large market share (as is typical for most public drug plans), then exclusion from the formulary will markedly reduce sales. There is a double effect from exclusion: nonformulary drugs tend not to be used, not only because of the cost to consumers, but also because physicians are not accustomed to prescribing them.

In summary, IP affords innovators some market power and this market power is complemented by widespread insurance coverage, which renders consumer demand less price sensitive. This has naturally resulted in high prices. However, more and more insurers are exercising countervailing market power and, by so doing, are reducing innovators' margins and hence the incentive to conduct R&D. Many payers now require innovators to demonstrate that their new drugs provide sufficient value for money as a condition for reimbursement. This has reduced margins both directly (when low willingness to pay thresholds are applied) and indirectly (by requiring firms to incur the time and expense of conducting economic appraisal studies).

Drug Access

Scholars distinguish between two separate drug access issues. The first is that IP may result in less drug use, relative to a world in which the drug is available for sale at a competitive (generic) price. More precisely, some consumers, who can be labeled as 'price sensitive,' are unable or unwilling to pay the brand price but may be willing and able to pay the marginal production cost, which would be the generic price in a well-functioning generic market. These sales are valued at more than their resource cost, so society would gain if the drug company lowered its price for the price-sensitive consumers. But to make these sales, the firm may need to reduce its price for everyone and, by so doing, may lose more revenues on its 'price-insensitive' customers – those who are willing to pay the brand price – than it earns on its price-sensitive customers. It would be profitable to sell at a lower price just to its price-sensitive consumers if it could prevent resale of the product to price-insensitive customers. But it is costly to prevent resale; indeed this appears to be the reason that drug companies were reluctant to sell acquired immunodeficiency syndrome (AIDS) drugs at discounted prices in low-income countries. There is also pressure from higher-income countries to get the same discounted prices offered in low-income countries. Recently some companies have used 'tiered pricing' in which the price in the lowest-income countries is essentially just the manufacturing cost. However, even then, as observed by Flynn,

Hollis, and Palmedo, within many developing countries the most profitable price may be the one that targets chiefly price-insensitive, high-income consumers.

Although access problems are most acute in developing countries, they also affect insured residents of developed countries. With prices of over US$50 000 for many new biologic therapies, insurers (and especially government drug plans) are not listing some products in their formularies.

The second access issue is that IP directs R&D into therapeutic areas where expected revenues exceed anticipated drug development costs. Thus diseases with limited market demand have traditionally received little attention from drug companies. This includes diseases affecting large numbers of individuals in poor countries (such as drug resistant TB, malaria, and other tropical diseases). This inequity in the distribution of R&D effort is morally problematic. It is true that this problem is not caused by IP. Diseases of poverty will tend to be neglected regardless of whether or not society extends IP protections to commercial drug developers. For such diseases to be the focus of R&D effort necessarily requires that affluent people underwrite R&D costs. However, if the decision has been made for the affluent to subsidize the costs of developing drugs with limited market demand, then it does not follow that IP is the best way to incentivize such R&D.

'Profit Raiding'

The final drawback of IP is that some portion of the innovator's potential sale revenues will simply be lost. The reason is that the high profit margins provided by market exclusivity attract 'raiders' who attempt to appropriate these margins. The potential profits from IP protection therefore decline, both by the profits actually appropriated by raiders and by the resources expended by the innovator to fend off raiders. Hence the threat posed by raiders dulls the financial incentive to innovate in the first place.

Counterfeiters, the clearest example of a profit raider, are attracted to patented drugs owing to their high margins and low transport costs. Historically, drug companies ignored the problem given that most contraband was sold in low-income countries, where potential profits were low. This has changed. Advances in counterfeit technology, the entry of organized crime syndicates into the counterfeit industry, and the introduction of patent protection (and hence higher drug prices) in several emerging markets following the 1994 ratification of the Trade-Related Aspects of Intellectual Property Rights (TRIPs) agreement, resulted in large increases in counterfeit sales. This counterfeit is increasingly difficult to distinguish from the genuine product and is infiltrating developed country markets. Drug companies have responded by changing the design of their pills, tablets, and packaging to make imitation more costly; they have also invested in radio-frequency identification and other technologies to secure their distribution channels from infiltration. Despite these efforts, losses from counterfeiting have been estimated to be as high as US$45 billion (US) annually. Lybecker reports that counterfeiting remains a pervasive problem "impacting nations of every size and income level and drugs of every description."

Drug resellers are another type of profit raider. International differences in price regulation regimes and national income, as well as exchange rate fluctuations and other factors, result in differences in the maximum price that a multinational drug company can charge in different markets. These variations present drug companies with a dilemma. On the one hand, profit maximization requires that they charge as much as each market will bear, so that less affluent countries will pay less toward the cost of R&D than richer countries. But drug resellers are quick to exploit price differences. Moreover, price regulators in Canada and elsewhere mandate that they pay no more than what is paid in a set of comparator countries. So listing at a low price in one country might cannibalize profits elsewhere. Faced with this tradeoff, a drug company might sacrifice profits in a country with limited willingness to pay (by delaying listing or listing at a higher than optimal price) to preserve more substantial profits in a country with greater willingness to pay. Nevertheless, it is difficult to eliminate all arbitrage opportunities. For instance, Bart presents estimates of the value of the drugs resold in the EU as being in the order of EUR€ 5–6 billion in 2006.

Generic competition can also be a form of raiding, to the extent that it undermines the legitimate and expected exclusivity period. Generic firms have an obvious financial incentive to enter large markets as soon as possible and will mount legal challenges to patents perceived as being weak. Brand firms have responded by increasing the number of patents listed on each drug. Indeed, Frank reports that branded drug firms in the US now carry an average of 10 patents for each drug – as compared with an average of 2 a decade earlier. Two studies have examined the impact of these developments on exclusivity periods. Grabowski and Kyle examine drugs in the US for which there was generic entry in the period 1995–2005, and find that there was an increase in generic challenges in both small and large 'blockbuster' pharmaceutical markets, which reduced market exclusivity periods. Hemphilll and Sampat, examining data for 2001–10, confirm increased numbers of challenges by generics, especially those that occur within the first five years of a drug being on the market. In contrast, however, they find that exclusivity periods have not changed significantly. Their explanation for this is that generic firms are chiefly challenging low-quality patents, so that generic challenges are simply 'maintaining' the traditional patent life by preventing low-quality patents from extending exclusivity.

In addition to strategic patenting, brand drug companies have used two other strategies to mitigate profit loss from generic competition. First, a brand firm may launch a generic version of its branded drug product – a so-called 'authorized generic' – to compete with independent generics for price-sensitive consumers. Second, if generic entry is likely, there may be an arrangement between the firms to delay generic entry. Such arrangements can be profitable because the brand firm, should it retain market exclusivity, can typically earn a higher margin on the generic firm's unit sales than the generic could itself earn. So the brand firm can pay the generic firm the margins that it would have made and still have money left over. These arrangements have attracted considerable attention under antitrust laws; naked pay-for-delay settlements are now prohibited in the US. More recently the US Federal Trade

Commission has challenged settlements in which the compensation to the generic for delayed entry is a promise of less aggressive competition.

Finally, noting that the market entry of follow-on drugs can be considered as a form of profit raiding, although unlike the case of counterfeiters, follow-on drugs provide benefits to consumers, in the form of expanded treatment choices, some of which are therapeutically superior to the pioneer drug.

Alternative Push and Pull Schemes

Analysts have proposed a variety of different push and pull schemes that are claimed to yield more therapeutically valuable drugs per dollar of public support. These proposals are outlined as follows.

Alternative Push Programs

Push programs aim to reduce the cost of conducting R&D to drug companies. Governments are already heavily involved in basic medical research, and this research substantially reduces the cost of bringing new drugs to market. However, there has been much interest recently in push programs that extend beyond basic clinical research, and into development problems that have traditionally been the domain of pharmaceutical companies. Of these, two proposals have received the most attention: (1) public subsidies for translational research and (2) public subsidy of Phase III clinical trials.

Public subsidies for translational research

A key determinant of the cost of bringing new drugs to market is the rate of failure of drugs in late stage clinical trials (where drugs are tested on large number of subjects with the disease). Drug candidates that are abandoned at this stage can be enormously costly. A high profile example was the failure of Pfizer's drug torcetrapib in Phase III trials in 2006; the development costs on this ultimately unsuccessful product totaled US$1 billion.

The failure of drug candidates is due to an incomplete understanding of human pathophysiology and pharmacology. Two aspects of the reward systems that drive commercial drug discovery inhibit learning. First, as was noted earlier, drug companies keep the results of their drug development programs secret, at least before patents are filed; this secrecy is in part due to the IP system. Second, due to the high risk of failure in pursuing risky hypotheses, companies tend to focus only on those targets that are well-studied in academic labs. But academia focuses on only a small fraction of potential targets. There are approximately 3000 targets in the human genome that are potentially susceptible to a drug. Of these, by 2006 only a few hundred targets had been shown to be therapeutically useful and modifiable by metabolically accessible, nontoxic drugs. Drug companies are understandably reluctant to invest significantly in their own target validation programs given the high risk of failure and concerns that competitors will use this information to develop competing drugs.

Rai *et al.* and Edwards *et al.* concur that this work requires the expertise and resources of both the academic and industrial pharmaceutical sectors, and therein lies the problem. Academic researchers collectively may have greater insight of the disease relevance of targets than do individual drug companies, given that the public sector collectively spends far more than industry does on understanding basic disease mechanisms. Drug companies have the expertise and resources needed to carry out the systematic steps of drug discovery. They also hold three other key inputs for target validation: (1) proprietary collections of small molecules that are needed to assess the functional attributes of proteins, (2) the expertise of medicinal chemists needed to produce new ones, and (3) the expertise and resources to develop and test biologics. To date, most collaborations between industry and academia are conducted within closed IP frameworks, in part because drug companies are reluctant to share their knowledge and resources widely, lest they benefit competitors.

To accelerate the process of target validation, Rai *et al.* propose that a trusted agency provide a sort of matchmaking service between academics and drug companies. The agency would assess targets discovered by academics using the small molecule libraries owned by drug companies and notify both parties if a target was hypothesized to have therapeutic potential. If both parties wanted to deal, the agency would help broker an IP agreement. Edwards *et al.* suggest that target validation, which only occurs after extensive clinical trials, is best conducted as part of an open-access, not-for-profit collaboration between the academic and commercial sectors. Placing research findings in the public domain in real time and unencumbered by IP restrictions would disseminate findings rapidly and widely, avoid duplication of effort, and conserve the considerable time and energy that is required to allocate IP rights over basic scientific discoveries. The open-access model would also prevent the exposure of research subjects to experimental drugs that are often ineffective, and known to be so by at least a few commercial players and regulators.

But why would academics and industry collaborate? Edwards and colleagues argue that drug companies that contribute their equipment, molecular libraries, and the expertise of their scientists would gain more from the collaboration (i.e., access to novel drug targets and influence over research directions) than they would lose from sharing their resources with potential competitors. To prevent free-riding, they propose that membership should be restricted to organizations that make a meaningful, and agreed-upon, contribution. For academics to commit fully, the collaboration must offer an attractive opportunity to conduct intellectually satisfying research and to receive peer recognition.

There is some evidence that such collaborations between academic and industrial scientists can be successful. The Structural Genomics Consortium (SGC), founded in 2003, accounts for approximately 15% of all the human protein structures in the public databases and is using this information to collaborate with industry medicinal chemists to generate open-access chemical inhibitors of new drug targets. The SGC is funded by the Canadian and Ontario governments, the Wellcome Trust, and eight drug companies, and all research output is released without restriction under an

open-access policy. Edwards and his colleagues are initiating a parallel effort in which a consortium of academic institutions and major drug companies collaborates openly to test the disease relevance of novel drug targets in humans.

Public subsidies for clinical research

Several commentators, including Lewis, Reichman and So, Baker, Boldrin and Levine, and Jayadev and Stiglitz, have advocated the public funding of Phase III clinical trials. Public funding of clinical trials would relieve drug companies of the single largest cost of drug R&D. At the same time, public spending on clinical trials would be relatively modest, for two reasons. First, governments already subsidize clinical trials through the use of tax subsidies. Second, governments likely face a cost of capital less than the 11% cost faced by the pharmaceutical industry. Because clinical trials must be conducted before marketing approval, development costs are very sensitive to the cost of capital. In addition to being relatively economical, publicly funded safety and efficacy trials can produce information that is more credible and clinically useful than industry-funded trials, which are naturally designed to maximize the expected profits rather than the expected clinical benefits.

The public agency responsible for the trials would presumably need a way of deciding which drug candidates are eligible for public funding. One concern is that the agency's choice of drugs whose trial costs are eligible for public subsidy may be subject to undue political interference. Moreover, the agency may not be well informed of the most promising drug candidates. These issues could be dealt with to some extent by providing public subsidies only to those drugs that cleared the clinical trials.

Product-development partnerships

As noted earlier, there are very weak incentives for commercial drug companies to invest in the development of drugs targeting malaria, tuberculosis, and other diseases prevalent in developing countries. Governments and private foundations, however, have made substantial investments in research that addresses the gap. But pharmaceutical companies themselves have the crucial advantages of substantial expertise, technological capabilities, and large libraries of potentially interesting compounds. So it was natural for 'product-development partnerships' (PDPs) to arise between governments or foundations and for-profit pharmaceutical companies. The difficulty was that for-profit companies wanted to be able to earn a return on their investment, whereas governments and foundations were especially interested in ensuring that the prices of products were low enough to enable widespread use in poor countries.

A key innovation in developing these PDPs was identifying how pharmaceutical companies could be rewarded for their participation, in a way that also allowed the 'public' partner to achieve its goals. Modern PDPs have typically satisfied the divergent goals of their partners by splitting the market for the product into a commercial one, left to the industry partner, and a humanitarian one, in which some arrangement was made to achieve wider access, usually through at-cost pricing or through licensing to competitive producers.

Thus, during the early 2000s, several important PDPs have been established, and some have been successful in delivering products to market. These PDPs have become a central component of the war on neglected diseases. The most substantial PDPs, according to funding received, are the International AIDS Vaccine Initiative and the Medicines for Malaria Venture (MMV), which are both funded at close to US$100 m annually. MMV has a large portfolio of products at different stages of development. Unlike a traditional drug company, which has a portfolio of drugs in different therapeutic areas, MMV is focused only on malaria. This approach would create undesirable risk for a for-profit company; but it is efficient for MMV because its advisory committee has the opportunity to compare many different prospective products and to choose a portfolio optimized to achieving success in addressing medical needs. Most other PDPs focus on one or two therapeutic areas, a feature that distinguishes them from private companies and offers a strategic advantage for development work.

Alternative Pull Programs

The push programs reviewed above aim to reduce the cost of conducting R&D to drug companies. Pull programs, conversely, provide rewards for the end products of the R&D process – new drugs. The alternative pull programs differ from the IP model in one key respect: instead of granting the innovator the right to exclude others from selling the drug, innovators are rewarded with payments that are proportional to the drug's value. This means that under the alternative models, competitive entry occurs sooner, so that sales volumes are greater and prices lower. With prices closer to competitive levels, drug access would be improved and resale, counterfeiting, and other forms of profit competition would be rendered less lucrative.

These proposals can be categorized along several dimensions.

Measurement of drug value

Under the current IP system, drugs that the market deems to be more valuable earn greater profits. But the market's measure of value – willingness to pay – is a noisy measure of a drug's value. In most markets, consumers assess whether a good or service is worth the price; consumer willingness to pay in such markets is a reasonable estimate of social value. Pharmaceutical markets are extraordinary because the consumer neither chooses the medicine (the physician does) nor pays for it (the insurer does). Market demand for drugs thus reflects physicians' prescribing decisions, insurers' coverage decisions (and related cost containment policies), and consumer willingness to pay for insurance and amounts not covered by insurers. Although physicians act as expert agents on behalf of patients, many physicians are doubly protected from pricing concerns, because they do not make even copayments.

Different schemes use different measures of drug value. Sanders' Medical Innovation Prize Fund Act would grant a public agency the power to decide on reward amounts and identify priority disease areas. The agency would be bound only by various guidelines, such as the guideline that more effective drugs should earn larger rewards. DiMasi and

Grabowski express concern that under a discretionary system, 'political rent seeking' and lobbying may distort research directions. Moreover, they suggest that for-profit drug developers are best able to identify and pursue the scientific opportunities that will lead to socially valuable products.

Other proposals would rely on forecasts of the profitability of new drugs. Kremer proposed that a public agency assess drug value by auctioning off IP rights to a new drug. In most cases – say, nine of ten auctions – the winning bid would be used only to set the reward payment to the innovator and the patents would then be placed in the public domain. In a randomly selected tenth auction, the winning bidder would receive the IP rights at the bid price. A defect of this scheme is that 10% of all new drugs remain under patent. However, this is necessary if auction participants are to take the bidding seriously.

David Levine has proposed a mechanism that could be useful in cases where a public sector agency or perhaps an open source drug discovery consortium has identified a compound with some therapeutic promise, but whose properties have not been subject to any clinical testing. Levine proposes that drug companies bid for the rights to these compounds. Bids consist of royalty rates that would accrue to the winner from all firms selling the drug, should the drug clear all the clinical trials and gain regulatory approval. The lowest bidder earns royalty income but is responsible for covering trial costs.

Levine's proposal is akin to the compulsory licensing schemes that have been used in Canada and elsewhere, but with one major difference. Historically, regulators set compulsory license royalty rates at some arbitrary amount, whereas Levine would let firms bid on the royalty rate. A firm's bid would depend on its ability to operate clinical trials and its expectations are: the likelihood that the drug will clear regulatory hurdles, the therapeutic value of the drug (vis-à-vis current therapies), the anticipated market size, and the number of competing firms.

How would drug companies fare under these proposed schemes relative to the existing IP system? In theory, Kremer's proposal would give innovators the present discounted value of their anticipated monopoly profits, so over the long term, firms would fare just as well as in the existing system, if expectations are unbiased. Indeed, because rewards are less variable under the Kremer approach, it might even be preferred to the IP system. Should Levine's proposal be adopted, competition among firms would decrease the rewards to the firms' opportunity cost – the most that they could earn in some other venture. Kremer therefore rewards the market value of the innovation, which itself is determined by the willingness to pay on the part of drug plans and consumers whereas Levine covers firms' cost of innovation.

The proposed Health Impact Fund (HIF) offers a mechanism that rewards firms in a way related to both value and cost. A drug enrolled in the HIF would be sold at cost but would earn payments proportional to its measured impact on population health in each of the 10 years following market launch. The proposal also allows for supplementary rewards for 5 years should its sponsor receive approval to use the drug for new indications. Each annual payment represents a share of a reward fund; the reward fund share for an enrolled drug in a given year is equal to the drug's share of the global health produced by all participating drugs in that year. Health impacts would be measured using many of the same health technology assessment procedures as currently used by drug plans when deciding whether or not to reimburse a new drug. For example, if all participating drugs were estimated to have produced 20 million QALYs in a given year, and if an enrolled drug had produced 2 million of these QALYs, then it would receive 10% of the fund. Contributions to the HIF reward fund by donor countries would be proportional to the donor's Gross National Income.

Participation in the HIF by drug companies would be voluntary; a drug developer could elect to exercise its IP privilege or relinquish high prices in exchange for the reward payments. By making the scheme optional, developers could earn at least as much as they would under the existing system. At the same time, because firms would compete over a fixed pool of rewards, the expected reward must be equal to the cost of development for the firm with a marginal project. Thus, this system makes rewards depend explicitly on the marginal cost of innovation.

Because the IP system, as well as the systems proposed by Kremer and Levine, are market-driven, firms have little incentive to conduct R&D into important diseases afflicting chiefly the poor. The HIF, in contrast, could be used to reward the development of drugs with large health impacts, even if the beneficiaries are themselves not funding the reward payments. The HIF could similarly incentivize the development of new uses of older drugs for which there would otherwise be no significant reward.

The HIF is but one approach that could be used to fund the development of drugs that are intended for use in low-income regions. The Advance Market Commitment is another. Several governments and the Bill & Melinda Gates Foundation have funded a US$1.3 billion program to subsidize the provision of the pneumoccal vaccine in the poorest countries. The subsidy is at a fixed rate per vaccine delivered, and firms were intended to compete for a share of the US$1.3 billion by accelerating the development of vaccines that would treat the most common strains of pneumonia in developing countries and by rapidly scaling up production.

A criticism of the HIF is that it requires a relatively complex centralized system of health impact assessment. However, many drug plans presently require forecasts of the health impacts of new drugs being considered for formulary inclusion, so the criticism is somewhat misplaced. It is true that the HIF requires measurements of actual – rather than anticipated – heath impacts, and also requires a standardized measure of health impact. These measures would likely vary somewhat by country, depending on the institutional features of health care, as well as health risks specific to each setting. So health impact assessment is front and center of the HIF approach whereas under existing assessment procedures, drug plans appear to be satisfied with a lower standard of evidence (i.e., anticipated, not actual health impacts), especially if the price charged is attractive. A pilot of the HIF could involve a performance-based reward applied in a single country to one or more drugs, to test the ability to measure impact in a credible way, and also to see whether firms would respond to the incentives inherent in the system.

It is also useful at this point to note some drawbacks of the auction mechanisms contemplated by Kremer and Levine. The auction format is widely used to elicit private information.

But Grinols and Henderson question the utility of the auction mechanism for reward determination given the substantial uncertainty over the profitability of a new drug. They argue that it is difficult to forecast profits owing to the introduction of competing drugs, and changes in disease prevalence and severity. Auctions are also subject to gaming by bidders, so they need to be carefully designed.

Kremer's proposal has one additional defect. The innovating firm receives a lump sum payment before the drug is actually sold and receives nothing thereafter. Hence there is little incentive for the firm to promote its drug (which is often an important component of achieving widespread sales) or to investigate new uses for the drug. The HIF, and the approach advocated by Levine, however, reward innovators in proportion to their market sales and therefore retain incentives for promotion and on-going product development.

Financing reward payments

The mechanisms proposed by Kremer, Sanders, and Hollis and Pogge would finance rewards using public funds, whereas Levine's scheme is self-financing.

Public funding has both pros and cons. On the one hand, because the technology embodied in a new drug is a classic public good, prices should ideally be close to competitive levels. With publicly financed rewards, drug prices would be lower than in royalty-based schemes, assuming that the funding scheme set limits on pricing, for example, by requiring generic licensing. Public funding could also simultaneously address equity goals through the distribution of the financial burden of R&D across taxpayers. In a privately financed scheme, this burden is distributed among drug users. The difference in drug prices would likely not be large for high volume drugs, but they could be for low volume drugs (including drugs used to treat rare disorders).

Public funding also has drawbacks. Public funding of investment into innovation requires taxation, which distorts labor-leisure choices and consumption decisions. This drawback should not be overstated, however, most prescription drug spending in developed countries is already publicly funded, either directly (through a public drug plan) or indirectly (through tax subsidies for private drug coverage). Under a rewards scheme, drug prices would drop markedly – likely by a larger percentage than the percentage increase in unit volume. Publicly funded drug spending should therefore decrease, and the savings could be directed toward the reward fund.

Both the existing IP system and the proposed publicly financed rewards system rely on contributions by different jurisdictions to finance drug R&D. One important advantage of a publicly financed reward system is that each jurisdiction's contribution to the rewards fund would be transparent, making it easier to ensure that financial commitments are being honored. International contributions to R&D in the existing patents system, by contrast, are opaque. The reason is that whereas the TRIPs agreement requires uniformity of patent length and nondiscrimination, it fails to prevent countries from negotiating aggressively on the prices of new drugs, or reducing the period of market exclusivity by delaying formulary approval. Ideally, countries would contribute toward innovation in proportion to their ability to pay. Such an allocation of contribution is not only ethically attractive, but it is also likely to be roughly efficient in a Ramsey sense.

A dedicated, publicly financed international reward fund has an additional advantage over the IP system. National governments are responsible for setting IP policy, but do not bear the full burden of higher drug prices; these costs are often borne by regional government drug plans or private plans. These plans do not receive much political kudos for supporting innovation, and, indeed, are rewarded by plan sponsors for reducing prices. Plan sponsors presumably care about innovation but if there is any impact of their price control on drug innovation, this is likely small and indirect and occurs only after a considerable lag. As a result these plans tend to focus myopically on cost control. A national government, conversely, could benefit politically from financing a drug innovation fund: it would create lower drug prices, no doubt popular with constituents, and it would support drug innovation in a very direct, visible way. In addition, it is the national government that has relationships with other national governments, and these relationships can be used to help deter free-riding.

Publicly financed reward schemes, then, are in some ways more transparent than existing arrangements. As a result, countries would be less able to shirk their responsibility to finance drug R&D. But this very transparency could make it difficult to strike a deal in the first place. Indeed, some commentators suggest that it would be very difficult for national governments to agree on a division of R&D costs and a means of enforcing the agreement. Moreover, these commentators suggest that it would be difficult to devise a sharing rule that is responsive to changes in countries' willingness and ability to pay.

Nevertheless, the calls for a global R&D treaty continue. The World Health Organization's 'Consultative Expert Working Group' on R&D Financing has expressed preliminary support for a recommendation that countries begin negotiations toward such a treaty, but because the possible components are disparate and vague (and indeed the proposal includes no specific items that should be included in the treaty) it is difficult to know what is being proposed.

Although prospects for an R&D treaty are unclear, other international agreements appear to be feasible. For example, Lawrence Gostin and colleagues have proposed a research program for a 'global health governance framework' which would set out minimal responsibilities for countries in terms of meeting health requirements of their citizens, as well as international obligations to help build the capacity of low- and middle-income states.

In summary, the alternative pull programs described here reward new drugs in differing ways. Sanders would vest a public agency in the US with the power to decide on rewards based on domestic sales only. Kremer would use anticipated market demand and Levine would use anticipated market demand and the risk adjusted cost of running the clinical trials required by the regulatory authorities. The HIF would set rewards in proportion to the health gains generated by the use of the new drug globally. The HIF proposal is unique in that it is intended to supplement, rather than replace the IP system. Levine's proposal is unique in that it is self-financing; any firm could sell a newly approved drug, but each seller would be required to pay a royalty to the innovator on each unit sold. This royalty rate essentially reflects the anticipated risk adjusted

cost of generating the evidence needed to gain regulatory approval. The remaining approaches require a dedicated, publicly financed reward fund to remunerate innovators. The funds required vary between proposals. Because the HIF is a supplement to the IP system, the financing required would be less than in the proposals advanced by Sanders and Kremer. A wholesale replacement of the IP system would undoubtedly require contributions by different countries. This requires that an international agreement be struck, an outcome that some analysts view as politically intractable.

Discussion

The proposed mechanisms described here hold the promise to enhance the effectiveness of public support for the drug discovery enterprise. Implementation of these initiatives, however, requires that a variety of issues be addressed.

These include the following: Which reforms are feasible? Of critical importance, revenue streams must be predictable if firms are to commit funds to R&D projects. Should the IP system be replaced with a rewards scheme, then there needs to be an agreement that binds governments to commit resources. An agreement that reallocates the hundreds of billions of dollars spent annually on pharmaceutical R&D at this time is likely not politically feasible. However, more modest reforms might be possible. In particular, push-type programs that seek to reduce the private cost of drug discovery and commercialization, and pull-type programs that supplement, rather than replace, the IP system would be less disruptive to the status quo, would have more predictable consequences on the firm's profitability, and would involve smaller public sector financial commitments. At the same time, push-type policies, should they be successful, would reduce private drug development costs and hence the reliance on IP to recover development costs. This may, in turn, facilitate further reforms to the IP system.

The public–private consortiums engaged in translational research are particularly promising. However, some critical problems need to be considered. How can the consortium satisfy the many competing interests? In particular, how should the consortium decide on which therapeutic areas to investigate? Should this be decided by a vote or by consensus? Should there be sanctions applied to members that are found to not cooperate fully? For example, it may become apparent that a drug company withheld from the consortium research output that would have been useful. Should this result in expulsion from the consortium? Which drug companies would conduct clinical trials on drug candidates that emerge from a consortium? What restrictions will there be on the pricing of such drugs, given the contributions by academics and public funders? More generally, how should a consortium interact with the existing IP regime? Resolution of these issues would appear to be vital to the success of a consortium.

Another promising avenue is public funding of Phase III trials. If there were public funds, however, there would likely be no shortage of drug candidates seeking funding. Should public funding be linked to the ultimate success of the trial, or simply to the promise demonstrated before the trial? How should conflicting priorities among different disease advocacy groups and among different jurisdictions be resolved?

Finally, more work is needed to operationalize the HIF's health impact measurement technology. Because the HIF relies on assessment of health impact, it is important to know how such an assessment would be performed and how firms would respond to being paid based on impact. A pilot trial could be done for a single drug in a country or region. The HIF also requires further analysis of antitrust issues and evaluation of its likely effectiveness.

See also: Patents and Regulatory Exclusivity in the USA. Pharmaceutical Parallel Trade: Legal, Policy, and Economic Issues. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Value of Drugs in Practice

Further Reading

Boldrin, M. and Levine, D. K. (2008). *Against intellectual monopoly*. 1st ed. Cambridge: Cambridge University Press.

DiMasi, J. A. and Grabowski, H. G. (2007). Should the patent system for new medicines be abolished? *Clinical Pharmacology & Therapeutics* **82**, 488–490.

Edwards, A. (2008). Open-source science to enable drug discovery. *Drug Discovery Today* **13**, 731–733.

Edwards, A. M., Bountra, C., Kerr, D. J. and Wilson, T. M. (2009). Open access chemical and clinical probes to support drug discovery. *Nature Chemical Biology* **5**, 436–440.

Gostin, L. O., Heywood, M., Ooms, G., et al. (2010). National and global responsibilities for health. *Bulletin of the World Health Organization* **88**, 719–719A.

Grootendorst, P., Hollis, A., Levine, D. K., Pogge, T. and Edwards, A. M. (2011). New approaches to rewarding pharmaceutical innovation. *Canadian Medical Association Journal* **183**, 681–685.

Kremer, M. (1998). Patent buyouts: A mechanism for encouraging innovation. *The Quarterly Journal of Economics* **113**, 1137–1167.

Lewis, T., Reichman, J. and So, A. (2007). The case for public funding and public oversight of clinical trials. *Economists' Voice* **4**, 1–4.

Love, J. and Hubbard, T. (2007). The big idea: Prizes to stimulate R&D for new medicines. *Chicago-Kent Law Review* **82**, 1519–1554.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203–214.

Rai, A. K., Reichman, J. H., Uhlir, P. F. and Crossman, C. (2008). Pathways across the valley of death: Novel intellectual property strategies for accelerated drug discovery. *Yale Journal of Health Policy, Law & Ethics* **8**, 1–36.

Stargardt, T. and Vandoros, S. (2013). Reimbursement and price regulation in Europe. In Culyer, A. (ed.) *Encyclopedia of health economics*, 1st ed., Ch. 12. New York: Elsevier.

Towse, A. (2013). Measuring value: Implementation. In Culyer, A. (ed.) *Encyclopedia of health economics*, 1st ed., Ch. 12.5. New York: Elsevier.

Relevant Websites

http://www.policynetwork.net/uploaded/pdf/Global_Medical_Research_web.pdf
A Global Medical Research and Development Treaty.

http://ssrn.com/abstract=1830404
Evergreening, Patent Challenges, and Effective Market Life in Pharmaceuticals.

http://www.keionline.org/misc-docs/SandersRxPrizeFundBill19Oct2007.pdf
Medical Innovation Prize Fund Act of 2007.

http://www.who.int/pmnch/topics/economics/20100505_medicinesaccessible/en/index.html
The Health Impact Fund.

Patents and Regulatory Exclusivity in the USA

RS Eisenberg, University of Michigan Law School, Ann Arbor, MI, USA
JR Thomas, Georgetown University Law Center, Washington, DC, USA

Introduction

Research and development (R&D) for pharmaceuticals, medical devices, and other healthcare technologies require not only the creation of new products, but also the production of information about their effects through preclinical research and clinical trials. Firms might underinvest in this costly, risky, and time-consuming R&D in the absence of subsidies or rewards. But intellectual property rights are a two-edged sword in the struggle to improve health. They both motivate the efforts that result in healthcare innovation, and limit access to the fruits of those labors. This contribution examines and critiques the patent and regulatory regimes that offer innovators in healthcare technologies temporary shelter from competition. The primary emphasis is on the US law pertaining to pharmaceuticals, with some attention to medical devices and diagnostics as well as to laws in other jurisdictions.

Pharmaceutical Patents

In comparison to firms in other industries, drug companies rely heavily upon the patent system to capture rewards from R&D. Most new drugs are covered by a series of patents with staggered terms pertaining to their active ingredients, particular pharmaceutical formulations, manufacturing techniques, methods of medical treatment, and related products and processes. Critics decry this patent layering strategy as 'evergreening,' whereas brand-name firms explain that they often improve their initial marketplace offerings through patentable innovation.

Patent Acquisition

To be protectable under US law, an invention must consist of patentable subject matter and be new, useful, and nonobvious. 35 US Code §§ 101, 102, 103, 112. The inventor must file a patent application that provides a written description of the invention and that enables a person skilled in the field to make and use it. 35 US Code § 112. Failure to satisfy these requirements may lead to rejection of a patent application or to invalidation of an improperly issued patent.

Patent law encourages innovative firms to file patent applications promptly. Patentability depends on how an invention compares to the 'prior art,' including publications, patents, and public knowledge or use. Patentability is denied if the prior art discloses an invention, either explicitly or inherently, or makes it obvious. Because the prior art is constantly expanding, broad patent claims in follow-on patents are often invalid.

For example, in Schering versus Geneva, 339 F.3d 1373 (Fed. Cir. 2003), the patent at issue claimed descarboethoxyloratadine (DCL), a metabolite of loratadine. The prior art included an expired patent on loratadine. Schering sought to enforce the DCL patent against sellers of generic loratadine by arguing that patients who ingested loratadine would necessarily produce DCL in their guts, thereby infringing the DCL patent. The court held the DCL patent invalid, reasoning that if, as both parties agreed, administering loratadine to patients inherently causes the production of DCL, DCL became prior art when the earlier loratadine patent disclosed administering loratadine to patients. A narrower claim to the metabolite in isolated form might have survived this challenge, but such a claim would not preclude competitors from selling loratadine.

Patents claiming modest changes that are sufficient to avoid a novelty challenge have often been held invalid for obviousness, including claims to new dosages, new formulations, combination products that package the drug with another familiar ingredient in a single capsule, and single enantiomers or diastereomers isolated from a racemate or other mixture of stereoisomers. To avoid invalidation for obviousness, the patent holder must show either that a person working in the field would not have found the modification obvious or that the modified version has surprising properties not present in the prior art. This has proven to be a significant obstacle for many follow-on patents.

Despite the advantages of early filing, the time lag between the discovery of a new molecule and the development of information necessary to use that molecule as a drug may make early filing difficult. First, without some research into the properties of a molecule, it may be difficult to satisfy the utility requirement. The patent statute limits protection to inventions that are 'useful,' a standard which the courts have explained falls far short of US Food and Drug Administration (FDA) approval standards of safety and efficacy for use in humans. As a general matter, promising *in vitro* test results have been held to support patentable utility. In re Brana, 51 F.3d 1560 (Fed. Cir. 1995). The Supreme Court has held, however, that it is not enough to show that as of the filing date the invention is merely the subject of research. Brenner versus Manson, 383 US 519, 536 (1966). In some cases courts have required clinical testing to support controversial claims of therapeutic utility.

Second, if a new drug is structurally similar to a prior art molecule, further research may be necessary to satisfy the nonobviousness requirement. The US Patent and Trademark Office (PTO) deems a new but structurally similar variation of a known molecule to be *prima facie* obvious. The patent applicant may rebut *prima facie* obviousness by showing that the new molecule has surprising properties (or particularly advantageous properties) not present in the prior art molecule. In re Dillon, 919 F.2d 688 (Fed. Cir. 1990). To make this showing, it may be necessary to perform tests on both the new molecule and the prior art molecule. If the obviousness challenge does not arise until after the product has been thoroughly tested, the patent holder may use studies

performed after the filing date to show surprising or advantageous properties. Knoll Pharmaceutical versus Teva Pharmaceuticals USA, 367 F.3d 1381 (Fed. Cir. 2004). But if the PTO rejects the application at the outset for prima facie obviousness, the need to develop rebuttal evidence may cause delays. Meanwhile, firms might hesitate to invest in costly testing while patentability remains in doubt.

Recent judicial developments in the US have called into question whether patentable subject matter includes certain biotechnology products obtained from nature (such as DNA molecules) and certain medical diagnostic methods that involve correlating observed biomarkers with a patient's condition, prognosis, or treatment. The US Supreme Court held in Association for Molecular Pathology v. Myriad Genetics, 133 S. Ct. 2107 (2013) that isolated DNA molecules that are otherwise identical to naturally occurring DNA are not patent-eligible, but that complementary DNA (cDNA) molecules that are not naturally occurring but are created artificially are patent-eligible. Previously, the US Supreme Court had held in Mayo Collaborative Services versus Prometheus Laboratories, Inc., 132. Ct. 1289 (2012), that a process patent claim directed to a method of determining whether a given dosage level for a drug was too high or low by comparing observed levels of a drug metabolite with reference values set forth in the claim was an invalid attempt to patent a natural law. The implications of these cases for other biopharmaceutical and diagnostic patents are not yet clear.

Patent Term

As required by the World Trade Organization (WTO) Agreement on Trade Related Aspects of Intellectual Property Rights (hereinafter TRIPS Agreement), patents on most inventions expire 20 years after their application filing dates. 35 US Code § 154. Pharmaceutical patents may be entitled to term extensions, however, under the Hatch–Waxman Act. In particular, drug patents may be extended for up to 5 years to compensate for some of the time lost during clinical trials and regulatory review. 35 US Code §§ 155, 156. The remaining patent life after extension may not exceed 14 years beyond the date of FDA approval. The period of extension includes one-half of the time spent in clinical trials and all of the time between submission and approval of the new drug application (NDA). Both periods are reduced by any time attributable to an applicant's lack of diligence. Only the first approval of a new active ingredient qualifies for a patent term extension, and only one patent may be extended per new active ingredient. The patent to be extended must be in force on the date of approval and must cover either the product, a method of using the product, or a method of manufacturing the product. These provisions present a firm with a strategic dilemma: Should it extend an early-filed patent that expires sooner but is more likely to survive a validity challenge, or a follow-on patent that potentially confers more years of exclusivity, but is narrower and more vulnerable to patent-defeating prior art?

Patent enforcement

The Hatch–Waxman Act fundamentally altered drug patent enforcement in the US. Before Hatch–Waxman, generic versions of previously approved drugs faced two major entry barriers: First, the FDA approval process, which generally required the same showing of safety and efficacy for a generic product as for a pioneering drug; and second, patents, which private owners had the burden of enforcing through infringement actions in the courts. The FDA entry barrier was usually sufficient to defer generic entry long after relevant patents had expired, because the costs of clinical trials were prohibitive for generic products that would be sold at competitive prices. The Hatch–Waxman Act lowered the regulatory entry barrier considerably by allowing approval of a generic product that is 'bioequivalent' to a previously approved product under an Abbreviated New Drug Application (ANDA), without requiring duplication of safety and efficacy trials. An ANDA does not require full reports of clinical trials to show safety and efficacy, so long as the conditions of use, active ingredients, route of administration, and strength are the same as a previously approved 'listed product,' the ANDA product is 'bioequivalent' to the listed product, and the labeling of the two products is the same. 21 US Code § 355(j)(2).

At the same time, the Hatch–Waxman Act fortified the patent entry barrier by creating a system within FDA for tracking drug patents and by deferring the approval of ANDAs during the patent term. The Hatch–Waxman Act set up a complex process to divert disputed patent issues to the courts and to motivate potential competitors to challenge the validity of patents. The FDCA requires that an NDA disclose any patent that claims the drug or a method of using the drug 'with respect to which a claim of patent infringement could reasonably be asserted if a person not licensed by the owner engaged in the manufacture use, or sale of the drug.' 21 US Code § 355(b)(1). Upon approval of the NDA, the FDA publishes this information, updated to include later patents, in a publication (available on the FDA website) called 'the Orange Book.' The statute requires that an ANDA include one of four prescribed 'certifications' with respect to each patent in the Orange Book for the previously approved 'listed drug,' 21 US Code § 355(j)(2)(A)(viii): a 'paragraph I certification' indicating that no patent information has been filed; a 'paragraph II certification' indicating that the patent has expired; a 'paragraph III certification' indicating the date on which the patent will expire; or a 'paragraph IV certification' indicating 'that such patent is invalid or will not be infringed by the manufacture, use, or sale of the new drug for which the application is submitted.' 21 US Code §§ 355(j)(2)(A)(7)(vii)(1)-(4). If no relevant patents remain in force, the ANDA may be approved without further delay, assuming it is otherwise approvable. If a patent is still in force, the ANDA may be approved upon its expiration date. The consequences of a paragraph IV certification are complex. An applicant making a paragraph IV certification must give notice within 20 days to each owner of the patent and to the holder of the approved NDA including 'a detailed statement of the factual and legal basis of the opinion of the applicant that the patent is invalid or will not be infringed.' 21 US Code § 355(j)(5)(B)(iii). The ANDA may then be approved immediately, unless a patent infringement action is brought within 45 days. The filing of a patent infringement action triggers a 30-month stay of approval of the ANDA, which may be adjusted by the court. The

statute gives the first firm to file an ANDA with a paragraph IV certification for a listed drug a 180-day market exclusivity period before the FDA will approve another ANDA for the same product. These provisions have been litigated extensively as firms have explored their strategic implications.

The Hatch–Waxman Act in effect shifts some of the burden of patent enforcement from patent owners to FDA by directing FDA to refrain from approving ANDAs during the patent term. Unless there is a paragraph IV certification, FDA will use its regulatory gatekeeper role to exclude competitors until the patents listed in the Orange Book expire, without the need for infringement litigation. Patent holders need not monitor competitors to detect infringement; the burden is on firms seeking to enter the market to address infringement of listed patents when they file ANDAs. Even when an ANDA filer challenges the patent in a paragraph IV certification, a patent holder who brings an infringement action gets an automatic 30-month stay of approval of the ANDA, without having to meet judicial standards for a preliminary injunction. These features of the statute fortify the exclusionary effect of patents beyond ordinary judicial remedies for patent infringement through the use of heightened regulatory entry barriers during the patent term.

A distinct advantage for patent holders of the FDA-administered remedies is that, unlike a court, FDA makes no effort to determine whether patents listed in the Orange Book are valid and infringed, or even whether they claim the listed drug or its use. FDA relies on the NDA sponsor to identify which patents are appropriate for listing in the Orange Book, and if anyone disputes the accuracy of the patent information or the propriety of listing a particular patent, FDA relies upon the sponsor to decide whether to change the listing or to leave it as is. FDA does not itself consider the merits of paragraph IV certifications, but simply defers approval of ANDAs for 30 months pending further instructions from the courts. All listed patents get the same administrative treatment regardless of their validity or scope.

Typically the patent holder has much more at stake in ANDA litigation than the generic challenger. If the patent holder prevails, the court will likely direct the FDA to defer approval of the ANDA until the end of the patent term. Assuming no other ANDA filer successfully challenges the patent, the patent holder could remain the sole source of the drug for the remaining patent life, an outcome that could be worth billions of dollars in the case of a blockbuster product. If the generic challenger prevails, its ANDA will be approved, and if it was the first ANDA with a paragraph IV certification for that product, it may be the only ANDA-approved generic on the market for 180 days. Because the first generic competitor in the market for a drug is typically able to charge higher prices and capture a larger market share until a second generic competitor enters the market, this period of generic exclusivity has significant value. The value of generic exclusivity is diminished if multiple ANDA filers on the same date share the exclusivity, or if the NDA holder decides to launch its own competing 'authorized generic' during the generic exclusivity period. Teva Pharmaceuticals versus Crawford, 410 F.3d 51, 54 (D.C. Cir. 2005). A recent US Federal Trade Commission (FTC) report on authorized generics reports that, on average, expenditures at wholesale prices of a generic during the 180-day exclusivity period equal 61% of expenditures on the brand name product during a comparable period prior to generic entry (Federal Trade Commission, 2009). Once the generic exclusivity period expires and more generic competitors enter, price competition is likely to reduce profits considerably. Even with generic exclusivity, the profits that a generic challenger hopes to gain in the patent challenge are thus likely to be a fraction of the profits the patent holder hopes to preserve.

Given uncertainty as to the outcome and litigation costs, this gap between the value and risks to the patent holder and to the generic challenger may tempt the parties to try to reach a settlement. Some settlement agreements have provoked antitrust scrutiny by providing for 'reverse payments' from the patent holder to the generic challenger in exchange for agreement by the challenger to defer market entry (Federal Trade Commission, 2010). Before 2003 statutory amendments, these 'pay-for-delay' agreements could potentially preclude all generic entry so long as any patents remained in the Orange Book for the listed drug. This is because under previous statutory provisions, FDA could not approve subsequent ANDAs with paragraph IV certifications until 180 days after either (1) the first commercial marketing of the product by the first ANDA filer or (2) a court decision holding the challenged patents invalid or not infringed. Settlements could prevent either of these triggers from occurring, thus postponing indefinitely the time when competing ANDAs could be approved. The revised statute addresses this problem with several different patches. Specifically, the revised statute allows multiple ANDA filers on the same date to share the 180-day exclusivity, 21 US Code § 505(J)(5)(B)(iv)(I), (II); redefines the trigger that begins the 180-day period to be the first commercial marketing by any of the first ANDA filers, including an 'authorized generic,' 21 US Code § 505(j)(5)(B)(iv)(I); and provides for forfeiture of the 180-day exclusivity period if an applicant fails to market the drug within specified periods, withdraws the application, amends the certification, fails to obtain tentative approval for the ANDA, enters into an agreement that is adjudicated to be in violation of the antitrust laws, or if the relevant patents expire. See 21 US Code §505(j)(5)(D).

The FTC, along with a number of antitrust scholars, sees most settlements that involve reverse payments to defer generic entry as agreements in restraint of trade in violation of the antitrust laws (Federal Trade Commission, 2010; Hovenkamp *et al.*, 2003). Although owners of valid patents are entitled to exclude competitors from the market until the end of the patent term, antitrust authorities may suspect that settlements with reverse payments signal weak patent claims. But in the ANDA litigation context, the inference of a weak patent is not as compelling as might first appear. Even a victorious patent holder could not recover damages from a defendant that is not yet selling a product, and even an optimistic patent holder would presumably give up some portion of expected profits in settlement to reduce the risk of a litigation loss that could bring patent-protected profits to an abrupt halt. One might therefore expect settlement of even strong patent infringement claims to involve reverse payments. The US Supreme Court held in Federal Trade Commission versus Actavis, 133 S. Ct. 2223 (2013) that reverse payments might sometimes be

justified to settle a patent infringement action, but that a large, unexplained reverse payment might indicate that the patent infringement suit is weak, and that courts should apply "rule of reason" analysis to determine whether such settlement agreements violate the antitrust laws.

In 2003, Congress implemented reforms suggested by the FTC to minimize anticompetitive abuses, including revisions to the 180-day exclusivity period, and ensured continuing antitrust oversight by requiring that ANDA litigation settlement agreements be filed with the FTC and the Justice Department within 10 days of execution. Medicare Prescription Drug, Improvement and Modernization Act of 2003, Title XI, § 1112 *et seq*. The scrutiny of antitrust authorities presumably reduces the expected value of settlements and prolongs litigation.

Current law reflects repeated compromises to rebalance the interests of innovators and generic competitors as Congress and FDA seek to block abuses, with each legislative patch inevitably introducing a new set of unintended consequences. This complex regime, which provokes costly litigation with unpredictable results, is clearly far from optimal.

Regulatory Exclusivity for Drugs

Congress has repeatedly acted outside the patent system to shelter pharmaceutical innovators from competition by controlling the timing of regulatory entry barriers. These nonpatent exclusivity provisions promote certain forms of R&D, such as the development of orphan drugs or new chemical entities or the testing of approved drugs for new uses or for use in children. The terms of protection vary. Sometimes regulatory exclusivity runs concurrently with patent protection and sometimes it extends beyond the patent term. Even when shorter in duration than patents, most forms of regulatory exclusivity are better synchronized with the timeline of drug development than patents at the front end, so that the entire period of exclusivity starts to run only once the product is launched rather than ticking away before product launch. When patent validity and infringement are contested and uncertain, regulatory exclusivity can provide a minimum period of exclusivity that is less vulnerable to challenge than patent protection. Moreover, legislators enjoy greater latitude in designing regulatory exclusivity provisions to meet the needs of the pharmaceutical marketplace, while the patent system applies essentially the same rules to all fields of technology.

Orphan Drugs: Market Exclusivity

The Orphan Drug Act of 1983 grants 7 years of market exclusivity for products to treat rare diseases and conditions affecting fewer than 200 000 patients in the US. Available for both drugs and biologics, Orphan Drug exclusivity does not merely defer the use of an abbreviated approval pathway (ANDA). It entirely prohibits approval of another application 'for such drug for such disease or condition' for 7 years after the initial product approval, even if the later applicant conducts its own clinical trials. It does not, however, preclude approval of either (1) another drug for the same disease or condition or (2) the same drug for another disease or condition. Genentech, Inc. versus Bowen, 676 F. Supp. 301 (D.D.C. 1987); Sigma-Tau Pharms. versus Schwetz, 288 F.3d 141 (4th Cir. 2002). FDA interprets the statutory language 'such drug' narrowly to permit approval during the 7-year term of a 'clinically superior' product that uses the 'same active moiety.' 21 C.F. R. § 316.3(b)(13)(i), (ii).

Market exclusivity under the Orphan Drug Act is somewhat like a patent on a particular use of a drug, enforced by FDA, with the drug narrowly defined to exclude 'clinically superior' formulations. Although Congress might have thought it was facilitating only products with markets too small to be lucrative, many products qualifying for orphan drug exclusivity for one indication have had large and profitable markets, usually as a result of nonorphan, larger indications. The Orphan Drug Act can also provide a nonpatent source of exclusive rights for new uses of old drugs that were taken off the market and for which early patents have expired, such as thalidomide. But because it does not preclude approval of other applications to sell the same product for other uses, Orphan Drug exclusivity is of little value for products that competitors are free to sell for other indications.

Hatch–Waxman: Data Exclusivity

In the 1984 Hatch–Waxman Act, Congress provided 5 years of data exclusivity for the first approval of a new chemical entity (NCE), 21 US Code § 355(j)(5)(F)(ii), and 3 years of data exclusivity for a supplemental NDA approval making changes in a previously approved product that required new clinical trials, 21 US Code § 355(j)(5)(F)(iii). In contrast to the Orphan Drug Act provisions, these Hatch–Waxman Act provisions do not prevent a competitor that is willing to conduct its own clinical trials from obtaining approval of its own NDA. They merely prevent competitors from relying upon the innovator's prior showing of safety and efficacy by using an ANDA. In this sense they look less like patents on products and more like proprietary rights in regulatory data. But if the costs of a full NDA are prohibitive for a product that will be sold at generic prices, the practical effect is to defer generic competition. The term 'data exclusivity' is often used to refer to periods of delay before a follow-on product may reference the originator's data as part of an abbreviated approval application, and the term 'market exclusivity' is used to refer to a more comprehensive exclusivity such as that available under the Orphan Drug Act, but usage is not entirely consistent.

NCE exclusivity

The period of exclusivity for a new chemical entity, which begins with first market approval, often runs concurrently with patent protection, although in some cases it may last longer. Although sometimes referred to as '5-year exclusivity,' the effective period of protection is generally longer than 5 years. During the statutory period a competitor may not even submit an ANDA to FDA; effective exclusivity continues thereafter until FDA approves the ANDA, a process that takes an average of 19.2 months (Food & Drug Administration, 2007). An ANDA with a paragraph IV certification may be submitted as early as 4 years after approval of the NCE, but if the patent

holder responds by bringing a timely infringement action, the 30-month stay is extended to give a total of 7.5 years from initial NDA approval to the time of ANDA approval. 21 US Code § 355(j)(5)(F)(ii). With these adjustments, in practice the period of exclusivity from first approval of the NCE until approval of an ANDA is likely to range from a low of 5.5 years (if an ANDA with a paragraph IV certification is filed after 4 years, no infringement action is filed, and FDA takes 1.5 years to approve the ANDA) to a high of 7.5 years (if an infringement action is filed and the 30-month stay is extended in accordance with the statute), although approval times vary. The court in the infringement action has statutory authority to lengthen or shorten the stay. In the absence of patents, the period of exclusivity is 5 years plus approval time, or approximately 6.5 years. Relatively few NDAs do not involve patents, but some involve invalid patents. NCE exclusivity provides 4 years before such a product might face a patent challenge and 7.5 years before a patent challenger can enter the market under an ANDA.

Supplemental NDA exclusivity

FDA approval of a supplemental NDA is necessary to market a drug for a new indication, or in a different dosage form or formulation, or to sell the drug over-the-counter (OTC) rather than by prescription only. When FDA approves a supplement to a previously approved NDA that required further clinical trials for approval, the applicant is entitled to a 3-year period of exclusivity for the supplemental approval. 21 US Code § 355(j)(5)(F)(iv). The 3-year period begins with approval of the supplemental NDA, making it advantageous to defer filing a supplemental NDA until other forms of exclusivity are about to end in order to prolong the total period of exclusivity. As a product approaches the end of its patent life, a firm might, for example, seek approval to switch from prescription to OTC sales, thereby gaining 3 years before it faces generic competition in the OTC market. FDA may not approve an ANDA for the same change during the exclusivity period, but it may receive and review applications and grant tentative approvals that become effective when the exclusivity expires. The additional years of exclusivity are only available if additional clinical trials were necessary to get the supplement approved.

The exclusivity thereby gained is limited to the terms of the supplemental approval, and will not prevent a competitor from using an ANDA to get approval to sell the product as previously approved. This is a significant limitation on the exclusive rights conferred by a supplemental NDA to market a drug for a new indication. Such exclusivity does not stop a generic competitor from getting approval to sell its product for the original indication. Once the generic version is on the market, physicians may prescribe it off-label for the new indication, and pharmacists may substitute the generic version unless physicians expressly require that the brand-name product be used. Indeed, unless the new indication involves a different formulation of the product, state generic substitution laws may require substitution of the cheaper generic product at the pharmacy. A similar problem limits the value of both Orphan Drug exclusivity and method of use patents for products with multiple therapeutic uses.

A 3-year exclusivity remains commercially valuable in circumstances where generic substitution is less likely, such as a prescription to OTC switch. If the branded product becomes available OTC and the generic product is available only by prescription because FDA may not yet approve it for OTC sales during the period of exclusivity, patients may buy the branded product directly at the pharmacy rather than going to a physician for a prescription, or a physician may advise the patient that the product is available without a prescription. Either way, the generic product is at a competitive disadvantage because it is available by prescription only. Further, insurance would generally not cover the product once it is available OTC. Three years of exclusivity in the OTC market may give the branded product a dominant position with consumers that persists even after generic entry in the OTC market.

In one controversial episode, 3-year exclusivity led to a sharp rise in the price of colchicine, an ancient remedy for gout that had previously been marketed for decades without FDA approval (Kesselheim and Solomon, 2010). In 2006, FDA launched an Unapproved Drugs Initiative to get manufacturers of old drugs that came to market before modern premarket approval requirements to test the drugs for safety and efficacy and to seek formal regulatory approval (Food & Drug Administration, 2011). FDA had previously approved two combination products including colchicine as one of multiple ingredients, but had never approved a single-ingredient colchicine product. Mutual Pharmaceutical submitted NDAs for the use of colchicine to treat familial Mediterranean fever and acute gout flares. FDA approved both applications and awarded 7 years of orphan drug exclusivity for the use of colchicine to treat familial Mediterranean fever and 3-year exclusivity for the treatment of acute gout flares. FDA subsequently announced its intention to take enforcement action against unapproved single-ingredient colchicine products (Food & Drug Administration, 2010). Its new position as sole source of this ancient drug enabled Mutual Pharmaceutical to increase the price of its colchicine product, Colcrys®, from $0.09 to $4.85 per tablet (Kesselheim and Solomon, 2010). Although regulatory exclusivity normally affects only products that have not yet become available at competitive prices, this unusual case provides a stark illustration of the very real costs that exclusivity imposes on patients and payers for products that might otherwise have been supplied more cheaply in competitive markets.

Pediatric Exclusivity: Prolonging Existing Rights

The Food and Drug Administration Modernization Act of 1997 added 6 months of incremental exclusivity for conducting pediatric trials of drugs. This 6-month period of 'pediatric exclusivity' is not contingent upon approval of the drug for use in children and is not limited to such use. The only requirements are that FDA must request the pediatric studies and they must be completed and submitted within the timeframe specified by FDA.

In contrast to the Hatch–Waxman and orphan drug exclusivities, the 6-month pediatric exclusivity does not start running immediately upon FDA approval. It simply adds 6 months to the end dates of any existing forms of exclusivity held by the submitter, whether under a patent, the Orphan

Drug Act, or Hatch–Waxman Act provisions, 21 US Code § 355a, further deferring the time when FDA may approve a competing generic product. Because it does not run concurrently with other forms of exclusivity, there is no advantage to be gained by deferring pediatric trials until other forms of exclusivity are at an end. The result has been a significant increase in available information about the effects of drugs in children, but critics have questioned whether 6 months of exclusivity is an excessive reward for rather modest expenditures on pediatric trials.

Regulatory Exclusivities in Other Countries

Longer periods of regulatory exclusivity are available in the European Union, dating back to a time when some members of the European Union did not allow patents on pharmaceuticals (Junod, 2004). Council Directive 87/21/EEC of 22 December 1986, amending Directive 65/65/EEC on the Approximation of Provisions Laid Down by Law, Regulation, or Administrative Action Relating to Proprietary Medicinal Products. The European regime currently provides 8 years of exclusivity before authorization for a generic may be submitted and 2 further years before it may be approved, and extends each of these dates by an additional year if, during the first 8 years, the holder of the authorization obtains further authorization for one or more new therapeutic indications for the product. By decoupling the duration of exclusivity from the timing of supplemental approval, the European approach encourages the testing and submission for approval of new indications and formulations earlier in the life cycle of a drug, while the US approach makes it advantageous to defer such testing and submission in order to maximize the duration of exclusivity.

The pharmaceutical industry has sought to establish regulatory exclusivity regimes throughout the world in the terms of trade agreements, with mixed success (Reichman, 2009; IFPMA, 2011). The TRIPS Agreement, in lieu of a proposed requirement for a minimum of 5 years of regulatory exclusivity, ambiguously requires WTO members to protect undisclosed data against 'unfair commercial use' and against 'disclosure, except where necessary to protect the public.' TRIPS Agreement, Article 39.3. Subsequent regional and bilateral free trade agreements more clearly specify time periods during which national regulators may not approve generic drugs on the basis of bioequivalence or otherwise rely upon data provided by the originator (Reichman, 2009). Some nations have implemented regulatory exclusivity in their national laws, while in other cases treaty provisions for regulatory exclusivity are self-executing.

Medical Devices

Medical devices, which range from bandages to artificial heart valves, vary significantly in the levels of risk they pose, with products in different risk categories facing different regulatory entry barriers. The Medical Device Amendments Act of 1974 divides devices into three classes. 21 US Code § 360c. Class III devices pose the greatest risks, and generally require premarket approval of a comprehensive NDA-like application before they may be sold. 21 US Code § 360e. Some Class III devices and most of the less risky Class II devices may instead get to market through a less onerous pathway in which they are 'cleared' under a '510(k)' submission, 21 US Code § 360k, based on a showing that the proposed device is 'substantially equivalent' to a legally marketed device, meaning that it is at least as safe and effective as that device. The 510(k) submission may include results of clinical trials. Some Class II devices and most Class I devices are exempt from the 510(k) process.

Although premarket review of new devices is generally more abbreviated than that of new drugs, FDA also regulates devices through postmarket surveillance and controls. 21 US Code § 360l. A pure generic industry does not exist in this sector; rather, some firms focus on 510(k) applications that they anticipate will not require clinical trials because they are similar to existing products. The absence of pure generics may reflect the much lower R&D costs and much shorter product life cycles for medical devices than for drugs. With incrementally improved products coming out every couple of years, by the time an originator product's patents expired and a generic version could legally enter the market, the original product design would be economically obsolete. Medical devices, like drugs, are eligible under the Hatch–Waxman Act for patent term extensions to compensate for marketing approval delays. 35 US Code §156(g). They obtain no regulatory exclusivity, however, perhaps due to differences in the nature of FDA oversight of the drug and device fields and/or to the absence of a market for true generics due to the short product life cycles.

Biosimilars

Regulatory exclusivity is a significant focus in legislation for the regulation of follow-on biological products, reflecting its growing importance to the biopharmaceutical industry. As discussed earlier, Congress provided for periods of regulatory exclusivity for pharmaceuticals ranging from 4 years to 7.5 years in the 1984 Hatch–Waxman Act. In 2010, Congress enacted a new regulatory approval pathway for biological products that are 'biosimilar' to and/or 'interchangeable' with previously licensed biological products, preceded by a 12-year period of regulatory exclusivity for the reference products, a period that may extend beyond the expiration of relevant patents. Some critics have argued that the 12-year period of exclusivity is excessive, while others have argued that it is necessary to compensate for deficiencies in patents and have argued for even longer periods.

Decoupling Regulatory Exclusivity from Patent Protection

The biosimilars legislation departs from the Hatch–Waxman model by decoupling regulatory exclusivity from patent protection. While under the Hatch–Waxman Act the dates when an ANDA may be filed or approved turn in part on the expiration dates of patents in the Orange Book for the listed product and on the status of litigation between the parties over those patents, these dates are fixed for biosimilars: Irrespective

of patents, an application for a biosimilar license may not be filed for 4 years, and its approval may not be made effective for 12 years, after the first licensing of the originator reference product. 42 US Code §§ 262(k)(7)(A), (B). FDA is not charged with maintaining an archive of relevant patents in the Orange Book. FDA receives notice and a copy of the complaint in any patent infringement action and publishes such notice in the Federal Register, but there is no provision for FDA to enter a stay of regulatory approval of the biosimilar license pending resolution of the litigation. Instead, the biosimilar applicant must give notice to the reference product sponsor 180 days before the first commercial marketing of its product, allowing the reference product sponsor to seek relief from the court in an infringement action.

In the place of the Hatch–Waxman Act's patent certification system, the biosimilars legislation creates an extraordinarily elaborate set of provisions for resolving patent disputes. These provisions entail considerable prelitigation activity, including disclosure of the biosimilar application to the reference product sponsor, ad hoc identification of relevant patents by each party, a negotiation process regarding which patents will be litigated, and a simultaneous double-blind exchange of patents designated for litigation. The meaning and implications of these provisions have yet to be tested, and many years will likely pass before relevant parties develop the familiarity firms, attorneys, and jurists now possess with the provisions of the Hatch–Waxman Act.

This approach nonetheless holds some advantages over the Hatch–Waxman regime. It does not put FDA in the role of patent enforcer, but leaves this task to the courts. Innovators thus have no incentive to list dubious patents in the Orange Book to obtain a 30-month stay of regulatory approval from an agency that is unwilling to evaluate patent claims. Remedies for infringement follow adjudication in the courts rather than arising automatically at the outset based upon the patent holder's allegations. The timing and duration of regulatory exclusivity are certain and do not depend on the vicissitudes of either infringement litigation or regulatory lag times (unless regulatory approval takes longer than the 8 years between the date when an application may be filed and the date when its license may become effective).

Supplemental Exclusivity

In contrast to the Hatch–Waxman Act, the biosimilars legislation does not offer additional exclusivity for minor product changes requiring supplemental approval. An important exception is made for 'a modification to the structure of the biological product' that results 'in a change in safety, purity, or potency.' Such a modification gets its own full 12-year period of regulatory exclusivity. The legislation also adds 6 months of 'pediatric exclusivity' both to the 4-year period before an application for a biosimilar license may be filed and to the 12-year period before the license may become effective, under the same conditions applicable to pediatric exclusivity for drugs.

The different approach to supplemental exclusivity in biosimilars legislation may reflect the limited value that biopharmaceutical innovators have found in supplemental exclusivity under the Hatch–Waxman Act. Although that Act nominally awards 3 years of additional exclusivity for supplemental approvals for changes such as new indications, the terms of exclusivity are limited to the terms of the supplemental approval, allowing approval of competing products for the old indications. If supplemental exclusivity does not defer FDA approval of a generic for the older indications, the generic will usually be substituted when the reference product is prescribed for the new indication.

On the other hand, for a structural change to the product that changes its safety, purity, or potency, generic substitution is unlikely and additional exclusivity is likely to be valuable. In these circumstances, the new legislation not only provides exclusivity, but expands it to the full 12-year term given for new products. This approach may both overreward structural changes and undermotivate other changes that require investments in clinical trials.

Limits of abbreviation in approval pathway for biosimilars

The biosimilars legislation does less to lower the regulatory entry barrier to follow-on competition for biologics than the Hatch–Waxman Act did for generic versions of small molecule drugs. Although a follow-on competitor need not make the showing required for an ANDA that the active ingredient is 'the same' as that of the listed drug, it must demonstrate that its product is 'biosimilar to a reference product' based upon data derived from analytical studies, animal studies, and one or more clinical studies 'that are sufficient to demonstrate safety, purity, and potency' in a use for which the reference product is licensed, unless FDA determines that one of these elements is unnecessary. 42 US Code §§ 262(k)(2)(A)(i), (ii).

While a showing of 'biosimilarity' is sufficient to obtain a biologics license, a more extensive showing is necessary to get an agency determination that the follow-on product is 'interchangeable' with the reference product. A determination of interchangeability requires showing that the biosimilar product 'can be expected to produce the same clinical result as the reference product in any given patient,' and for products administered more than once, that 'the risk in terms of safety or diminished efficacy of alternating or switching between use of the biological product and the reference product is not greater than the risk of using the reference product without such alternation or switch.'

Without a determination of interchangeability, additional marketing may be necessary to get physicians and pharmacists to switch their patients from the reference product to the biosimilar product. By contrast, a determination of 'bioequivalence' for a generic drug under an ANDA is enough to permit or even compel pharmacists to substitute the less expensive generic version unless the physician directs otherwise. Since most biologics are infusions that are dispensed by physicians in their offices, rather than by pharmacists, substitution of biosimilars will depend on incentives and decisions of physicians rather than on decisions of pharmacists.

The cost to show biosimilarity and interchangeability to the satisfaction of FDA is unclear at this point, but informed observers expect that the costs will be considerably higher than necessary to show bioequivalence in an ANDA. Moreover, further marketing costs may be necessary to persuade physicians, pharmacists, and payers to switch to these

products. This may in part be an inevitable consequence of differences between drugs and biologics, but it is also now in part a function of deliberate legislative policy in the US.

Exclusivity for Follow-on Products

The first follow-on product to receive a determination of interchangeability is entitled to a period of exclusivity before FDA will make a determination of interchangeability for a competing product. 42 US Code § 262(k)(6). A biosimilarity determination does not trigger exclusivity for the follow-on product, and exclusivity for the first product determined to be interchangeable with a reference product does not preclude licensure of a competing product that is determined to be merely biosimilar to the reference product but not interchangeable with it. Follow-on exclusivity ends at the earlier of 1 year after first commercial marketing, 18 months after a final court decision in a patent infringement action against the applicant or dismissal of such an action, 42 months after approval if the applicant has been sued and the litigation is still ongoing, or 18 months after approval if the applicant has not been sued.

These provisions reflect an effort to address some of the perceived flaws with generic exclusivity in the Hatch–Waxman Act. Rather than encouraging litigation by rewarding the first applicant to challenge a patent, the biosimilars legislation rewards the first applicant to receive a determination of interchangeability. This encourages applicants to make the more difficult showing of interchangeability in addition to biosimilarity, and encourages alacrity in securing a license from FDA rather than promising rights on the basis of application filing dates. Moreover, the proposal limits the duration of exclusivity in ways that may not be avoided by delays in marketing, litigation, or settlement agreements.

Whether the exclusivity for follow-on biologics will prove as valuable as Hatch–Waxman generic exclusivity remains to be seen. Although the periods of exclusivity appear longer under the biologics legislation than the 180 days provided under the Hatch–Waxman Act, other follow-on products may obtain licenses during the exclusivity period based on a showing of biosimilarity to the reference product, potentially eroding sales and profits for the first interchangeable product. As a result, incentives to introduce follow-on biologics may be limited.

Regulatory Exclusivities versus Patents

The growing duration of regulatory exclusivities in the US raises questions about their role relative to that of the patent system within the healthcare industry. Economic incentives for innovation are traditionally the province of the patent system. Failings of the patent system might better be addressed through patent law reform, rather than by creating additional sources of exclusivity outside the patent system. Moreover, in the healthcare context intellectual property barriers to generic competition are in tension with the competing interest in promoting affordable access to medicine. Perhaps regulatory exclusivities should do no more than compensate sponsors for their FDA approval expenses, rather than promoting biopharmaceutical R&D.

Yet reforming the patent system may be more challenging than fortifying regulatory exclusivity. The patent system remains a one-size-fits-all legal regime that applies essentially the same rules to inventions arising in biopharmaceutical research, automotive engineering, information technology, semiconductors, rocket science, and even business methods, although the need for patent protection across these fields differs greatly. Tailoring the patent laws to address the environment for innovation within the pharmaceutical industry might upset the balance of protection and competition in other industries.

Membership in the WTO also limits the ability of member nations to tailor the patent system to specific industries. The WTO TRIPS Agreement requires signatories to provide patent protection 'without discrimination as to the place of invention, the field of technology and whether products are imported or locally produced' TRIPS Agreement Article 27. The brand-name pharmaceutical industry favored the prohibition in the TRIPS agreement against discrimination as to field of technology, because it would require member states to eliminate provisions in national laws that weakened drug patents (such as compulsory licensing provisions). But the treaty language seems to prohibit discrimination in favor of drug patents as well as against them. On the other hand, the TRIPS Agreement places few restrictions on the award of regulatory exclusivities by WTO members.

Enhanced regulatory exclusivity offers other advantages for brand-name drug companies over stronger patent protection. First, patents provide not so much the right to exclude as the right to sue to exclude. Generic firms frequently make successful arguments that the brand-name firm's patents are invalid or not infringed. In contrast, regulatory exclusivity keeps competitors off the market without the need for owners to bring costly and risky infringement actions. This may be particularly advantageous in countries that do not have well-functioning institutions for patent enforcement, but even in the US the costs and risks of infringement litigation are considerable.

Second, apart from its more limited duration, regulatory exclusivity is a better temporal fit with the life cycle of a pharmaceutical product. Regulatory exclusivity periods typically do not begin until a product is on the market, while much or all of a patent term may run before that time. Finally, the scope of regulatory exclusivity generally corresponds better to relevant product markets than do patents. Regulatory exclusivity tracks the terms of FDA product approvals, while patent claims, drafted to distinguish an invention from the prior art, may not correspond as closely to any actual commercial product.

From a political economy perspective, it may be easier for interest groups to influence policy initiatives that focus narrowly on a particular industry (such as modifications to drug regulation) than it is to influence policy initiatives that have a broader impact (such as modifications to patent law), because the broader the implications of the policy, the more likely they are to encounter competition from other interest groups. The pharmaceutical industry has sometimes found itself in opposition to the financial services and information technology

industries, for example, in legislative debates about patent law reform (Kahin, 2007). However, the biopharmaceutical industry was quite successful in shaping the terms of recent biosimilars legislation to secure generous benefits for its members (Greenwood, 2010). Perhaps an industry-specific approach increases the risk that policy makers will be unduly influenced by industry rent-seeking to the detriment of other interests that are less vigorously represented in policy debates.

The proposed Modernizing Our Drug & Diagnostics Evaluation and Regulatory Network Cures Act, or MODDERN Cures Act, H.R. 3497, 112th Cong., 1st Sess., provides an example of a framework of innovation incentives that emphasizes regulatory exclusivity over patents when the sponsor asserts that patent protection would be inadequate to support development of the product. Under the proposed legislation, which has been endorsed by a long list of disease advocacy groups, professional associations, and biopharmaceutical firms, a drug sponsor could submit a request to FDA for 'dormant therapy' designation for a therapy that fulfills an 'unmet medical need' and that has 'prospectively insufficient patent protection.' The request must include a list of patents covering the therapy and a conditional waiver of the right to enforce those patents after the termination of regulatory exclusivity. If the FDA agrees that the indication for which approval is sought addresses an unmet medical need, it will grant the dormant therapy designation and the patent waiver will become effective. The sponsor then obtains 15 years of marketing exclusivity. All of the identified patents are given an extended term of up to 15 years after the product is approved, but pursuant to the patent waiver, the sponsor of the disclaims any patent term after the 15-year exclusivity period. Whether the MODDERN Cures Act will become law remains to be seen, but the proposal itself represents a significant shift in understandings of the roles of patents and regulatory exclusitivies in promoting innovation.

Regulatory exclusivities have other disadvantages. They place public health officials in the uncomfortable position of denying patients access to safe and effective generic substitutes for unpatented medications. They require FDA to devote considerable time and manpower toward drafting regulations, issuing guidance documents, and adjudicating disputes involving multiple regulatory exclusivity regimes. These resources might be more effectively spent in pursuit of the agency's core mission of protecting public health.

Regulatory exclusivities lack important limitations that are built into patent law. The US Supreme Court recently explained that "the results of ordinary innovation are not the subject of exclusive rights under the patent laws. Were it otherwise, patents might stifle rather than promote the progress of useful arts." KSR Int'l. versus Teleflex, 550 US 398, 427 (2007). Yet regulatory exclusivities are available for old products based upon the completion of routine clinical trials that would not qualify for additional patent rights. Perhaps the same considerations that justify withholding patent rights from 'ordinary innovation' in order to promote progress also caution against awarding regulatory exclusivity for 'ordinary innovation' in the context of drug testing. No other industry receives comparable guarantees of an exclusive market to induce them to bring their products to market. Perhaps the extraordinary protections afforded to the pharmaceutical industry are diminishing the motivation of drug-developing firms to become more efficient innovators.

Conclusion

The current overlapping legal protections for exclusivity in the pharmaceutical marketplace reflect a series of political compromises, repeatedly renegotiated to correct for unintended consequences in the last version of the rules. It is not easy to design a set of rules that provides the optimal balance between incentives for innovation and barriers to competition. On one view, patent law reform provides a more appropriate response to concerns over particular innovation environments than the creation of a gallery of additional regulatory exclusivities. The failure to make the patent system responsive to this intensely science-based industry threatens to allow the patent regime that has long served as the engine of innovation to become antiquated. It may also maintain deficiencies in patent doctrine for the great majority of innovative industries that do not benefit from regulatory exclusivities.

Another view is that a simpler and more effective legal regime would rely less upon patent protection and more upon well-designed regulatory exclusivity to support incentives for new drug development. It makes little sense for lawmakers and trade negotiators to extend the Byzantine Hatch–Waxman system into new legal regimes, either by duplicating it for biosimilars or by binding our trading partners to adopt similar systems in their national laws. Recent biosimilars legislation in the US, although controversial in specifics, corrects some problematic structural features of the Hatch–Waxman Act, disentangling regulatory exclusivity from patents and offering greater incentives to develop products that may be inadequately protected by a patent system that is often asynchronous with biopharmaceutical product development.

See also: Biopharmaceutical and Medical Equipment Industries, Economics of. Biosimilars. Patents and Other Incentives for Pharmaceutical Innovation. Research and Development Costs and Productivity in Biopharmaceuticals

References

Federal Trade Commission. (2009). Authorized generics: An interim report. Available at: http://www.ftc.gov/os/2009/06/P062105authorizedgenericsreport.pdf (accessed 27.10.09).

Federal Trade Commission. (2010). Pay-for-delay: How drug company pay-offs cost consumers billions. Washington, DC: FTC. Available at: http://www.ftc.gov/os/2010/01/100112payfordelayrpt.pdf (accessed 02.03.10).

Greenwood, J. (2010). 'State of the industry address from 2010 BIO International Convention.' Available at: http://biotech-now.org/2010/05/10/state-industry-address-from-2010-bio-international-convention (accessed 15.07.10).

Hovenkamp, H., Janis, M. and Lemley, M. A. (2003). Anticompetitive settlement of intellectual property disputes. *Minnesota Law Review* **87**, 1719–1766.

International Federation of Pharmaceutical Manufacturers & Associations (IFPMA) (2011). Data exclusivity: Encouraging development of new medicines. Available at: http://www.ifpma.org/fileadmin/content/Publication/IFPMA_2011_Data_Exclusivity__En_Web.pdf (accessed 11.07.13).

Junod, V. (2004). Drug marketing exclusivity under United States and European Union Law. *Food and Drug Law Journal* **59**, 479–518.
Kahin, B. (2007). Patents and diversity in innovation. *Michigan Telecommunications and Technology Law Review* **13**, 389–399.
Kesselheim, A. S. and Solomon, D. H. (2010). Incentives for drug development – The curious case of colchicine. *New England Journal of Medicine* **362**, 2045–2046.
Reichman, J. H. (2009). Rethinking the role of clinical trial data in international intellectual property law: The case for a public goods approach. *Marquette Intellectual Property Law Review* **13**, 1–68.

Further Reading

Eisenberg, R. S. (2007). The role of the FDA in innovation policy. *Michigan Telecommunications and Technology Law Review* **13**, 345–388.
Engelberg, A. B. (1999). Special patent provisions for pharmaceuticals: Have they outlived their usefulness? A political, legislative and legal history of US law and observations for the future. *IDEA: Journal of Law and Technology* **39**(3), 389–428.
Engelberg, A. B., Kesselheim, A. S. and Avorn, J. (2009). Balancing innovation, access and profits – Market exclusivity for biologics. *New England Journal of Medicine* **361**, 1917–1919.
Food & Drug Administration. (2010). Single-ingredient oral colchicine products; enforcement action dates. *U.S. Federal Register* **75**, 60768–60771.
Food & Drug Administration. (2011). Guidance for FDA staff and industry: Marketed unapproved drugs: compliance policy guide. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070290.pdf (accessed 25.03.12).
Grabowski, H. G. and Kyle, M. (2007). Generic competition and market exclusivity periods in pharmaceuticals. *Managerial and Decision Economics* **28**, 491–502.
Institute of Medicine (IOM) (2011). *Medical devices and the public's health: The FDA 510(k) clearing process at 35 years*. Washington, DC: The National Academies Press.
Institute of Medicine (IOM) (2012). *Safe and effective medicines for children: Pediatric studies conducted under the best pharmaceuticals for children act and the pediatric research equity act*. Washington, DC: The National Academies Press.
Mossinghoff, G. J. (1999). Overview of the Hatch–Waxman Act and its impact on the drug development process. *Food and Drug Law Journal* **54**, 187–194.
Schacht, W. H. and J. R. Thomas. (2009). Follow-on biologics: Intellectual property and innovation issues. *Congressional Research Service*. Washington, DC. Available at: http://www.ipmall.info/hosted_resources/crs/RL33901_090320.pdf (accessed 31.05.09).
Thomas, J. R. (2006). Authorized generic pharmaceuticals: Effects on innovation. *Congressional Research Service*. Washington, DC. Available at: http://www.orangebookblog.com/files/thomas_j.,%20Authorized%20Generics.pdf (accessed 26.10.09).

Pay for Prevention

A Oliver, London School of Economics and Political Science, London, UK

Introduction

The idea of paying people to engage in healthy activities, and to refrain from unhealthy ones, gained some traction in the health policy discourse in several developed and developing countries toward the end of the 2000s. The concept itself is simple and is informed by one of the most basic features of standard economics, the relative price mechanism, that is, if you pay someone to do something, there is an expectation that the person is more likely to do it, and the higher the cash incentive, the greater the effect. Here, payments to people to encourage healthier behaviors are defined as user financial incentives (UFIs), and might be expected to have a positive effect if the utility that an individual gains from the payment outweighs the personal disutility consequent on the behavior change.

UFIs exist when an individual can expect a monetary transfer, which is made conditional on them for acting in a particular way. Mostly, they are positive rather than negative financial incentives, rewards not punishments. Taxes on certain harmful products, such as cigarettes, are, however, a form of negative financial incentive and can be quite effective in changing people's consumption patterns. Moreover, this intervention has gained considerable exposure in recent policy debates internationally, with respect to, for example, the so-called 'fat taxes' (i.e., additional taxes on food with high calorific content) and 'soda taxes.' The imposition of such measures affects all those engaged in the targeted activity though: those who alter their behavior, by the very fact of them altering their behavior, and those who continue as before, when they are now subject to higher prices. It could be argued, therefore, that taxes are heavily paternalistic. UFI are more libertarian by comparison, in that those who continue to smoke, or eat too much, or refrain from exercise, are not directly affected by the payments at all. It is for this reason – even though UFI ought to be seen as a complement to, rather than a substitute for taxes – that the political leanings of those who currently govern much of the world may lead them to take an interest in positive financial incentives.

Aside from the simplicity of the theory that informs UFI, one must ask oneself, do they work? To answer this, one must turn to the evidence.

Evidence

Numerous experiments with UFI have been, and are being, conducted at the local practical policy level in a number of countries, including the UK, with the use of financial and payment in kind (iPods, hotel breaks, helicopter trips) incentives by healthcare purchasers, and the US, in employer-based wellness programs. These local level pilots are, however, rarely well evaluated. Fortunately, there is a reasonable amount of evidence reported in the academic literature on the effectiveness of mostly quite small UFIs to enable us to reach some informed conclusions.

Behavior change can be categorized as either complex or simple. Complex behavior change requires sustained effort over a length of time; simple behavior change requires single actions at a point in time. Some preventive activities – for example, smoking cessation and healthier eating – clearly require a sustained effort, whereas others, such as attending doctors' appointments and participating in vaccination programs, are more likely to be relatively simple. Sustained and 'one shot' behavior change require qualitatively different responses from the individual, and thus it makes sense to dichotomize the evidence according to these categories.

Sustained Behavior Change

The behavior changes discussed here are those associated with smoking cessation and weight loss. There have been systematic reviews in both areas, and the evidence is not auspicious. In relation to smoking, abstinence is measured through use of a biochemical test that records cotinine levels in saliva or urine. In 2008, Cahill and Perera reported a systematic review of studies that have analyzed the effectiveness of UFI for smoking abstinence, wherein only 1 of the 17 studies has demonstrated significantly higher cessation rates for those to whom incentives are offered as compared with those in control groups beyond 6 months from the start of the intervention. Unfortunately, cost information is usually absent from UFI studies and thus even when effectiveness is observed, the discernment of cost effectiveness is difficult.

A 2007 meta-analysis of nine randomized trials on the use of UFI to reduce obesity rates after 12 months following the initiation of the incentive as reported by Paul-Ebhohimhen and Avenell has concluded that an incentive of less than 1.2% of personal disposable income is associated with a zero mean weight change. Financial incentives of at least 1.2% of disposable income were, compared with no incentive, associated with a mean weight loss of 2.4 lb at 12 months and 1.5 lb at 18 months. Thus, the effect by 18 months postinitiation was small and dissipating, and cost-effectiveness information was, again, missing. Indeed, in the domain of weight loss interventions, it is probably difficult to gauge cost effectiveness without knowing the health implications of the loss in weight, which will occur (if at all) many years in the future, and are therefore likely to be confounded by an individual's broader behaviors, environment, and genetic profile. Rewarding people for weight loss could also feasibly incentivize some quite unhealthy behaviors.

One Shot Behavior Change

As noted above, various types of medical adherence require single, or limited, acts. In 1997, Giuffrida and Torgerson reported a systematic review of UFI to motivate medical

adherence (including adherence to a tuberculosis medical regimen, dental care for children, immunization, postpartum appointments, etc.). They identified 11 randomized controlled trials, of which, 10 demonstrated a positive effect. There have been additional studies in the intervening years, many (although by no means all) of which have shown similar promise. For example, a 2003 study by Seal and colleagues reported a randomized controlled trial on a population of hard to reach intravenous drug users in San Francisco. All of the drug users were given the first of three required hepatitis B vaccine doses and then they were divided into two groups, an 'outreach group' and an 'incentive group.' The third vaccine dose was administered 6 months after the first dose, and the outreach group was assigned a weekly contact with an outreach worker; the incentive group, however, received a monthly US$20 monetary incentive if they remained in the vaccine program. It turned out that 69% of those in the incentive group received all three doses of the vaccine as against only 23% in the outreach group.

In another study, this time from 2005, Slater and colleagues administered two types of mail-based interventions to women aged 40–64 years to encourage them to undergo mammography. Both of the interventions offered a free mammogram if the respondent rang a toll-free number, with one of the interventions also offering a small financial incentive if a person actually underwent the mammogram within 1 year. More than four times as many calls were received for the mail-plus-incentive intervention than the mail-only intervention, and the subsequent mammogram rate was significantly higher in the former intervention than the latter intervention, which in itself produced a significantly higher rate than 'do nothing.' As with the interventions to encourage sustained behavior change, however, those that focus on medical adherence are generally silent regarding value for money.

Summary

There is evidence to suggest that UFIs are potentially useful in many areas of medical adherence, but in terms of policy areas that demand more sustained efforts from the targeted groups, the effectiveness of this intervention has been generally poor. In short, for smoking cessation and weight loss, any early success tends to dissipate when the incentives are no longer offered. It may be the case that many studies have been somewhat underpowered, in terms of the size of the incentives offered and the length of time they are offered for. Indeed, a large trial reported in 2009 by Volpp and colleagues at the Center for Health Incentives and Behavioral Economics at the University of Pennsylvania, would seem to suggest that larger incentives may work.

In the trial, 878 people were randomized to either a control group or an incentives group. At baseline, all participants tended to smoke approximately one packet of cigarettes daily. The participants in both groups received information regarding smoking cessation programs, but the incentives group additionally received US$100 for completing a program, a further US$250 for exercising abstinence for 6 months into the trial, and an additional US$400 if they remained abstinent until 6 months thereafter. At 12 months from the initiation of the trial, the cessation rate in the incentive group was significantly higher than that for the controls (14.7% vs. 5%), and although there was some relapse in both groups, this pattern persisted beyond the lifetime of the incentive at 18 months (9.4% vs. 3.6%). Moreover, using incentives in excess of US$100 over a 4-week period, Charness and Gneezy have provided some evidence that gym attendance may be sustained at a significantly higher rate than would otherwise be the case post trial, at least in the relatively short term.

Generally, however, offering larger incentives over longer periods may not be feasible, particularly if financed by the public sector and targeted at broad population levels, given the current global fiscal environment. Therefore, it would make sense to examine whether the payment mechanism in UFI can be redesigned so as to improve its effect.

Strengthening the Payment Mechanism

It may be possible to improve the strength of the UFI payment mechanism by appealing to the findings of behavioral economics. For instance, theoretically, requiring participants to commit their own money (a 'deposit contract'), with the intention of receiving their money back if they achieve the target behavior, might be expected to improve effect. This is because in the behavioral economics literature, it has been observed that losses loom larger than gains; that is, people attach a greater magnitude of disutility to losing a particular good than the utility that they attach to winning the same good. Given this general observation of loss aversion, we might expect the loss associated with giving up money in deposit contracts to make them more effective than the conventional practice of simply giving people money if they meet their target behaviors. Moreover, some have highlighted the behavioral economic observation that people tend to be attracted to large rewards that have a small probability of occurring, and that therefore, instead of offering the target population a small financial incentive for certain, they ought to be offered, if they successfully change their behavior, a lottery that has the same expected value as the certain payment, but entails some probability of winning a relatively large monetary amount. In short, compared with conventional UFI, deposit contracts and lottery payment mechanisms do not necessitate increases in the average payment, but they may trigger cognitive effects that make respondents perceive the incentives to be more substantial.

In a selection of small studies that have not been specifically informed by behavioral economics, the performance of deposit contracts and lottery payment mechanisms in motivating sustained behavior change has been mixed, at best. There has, however, been at least one UFI for weight loss study that was informed directly by behavioral economics, with some interesting, if not spectacular, results. In this 2008 study by Volpp and colleagues, participants were assigned to one of three arms: (1) a weight monitoring program that required a monthly weigh-in; (2) the weight monitoring program plus a deposit contract, where at the beginning of each month participants could deposit between 1 cent and US$3 per day, with the deposited amount matched by the investigators in addition to a fixed payment of US$3 per day, with all refundable if the participant met the targeted weight loss at the end of each

month; and (3) the weight monitoring program plus a lottery incentive, where, if they met their weight loss target, participants played a daily lottery that had an expected value of approximately US$3, with some of the lotteries comprising of a large payoff with small probability but most comprising of a small payoff with larger probability. The trial end point target weight loss for all participants was 16 lb at 16 weeks. At 16 weeks, the mean weight losses in the control group, the deposit incentive group and the lottery incentive group were 3.9, 14, and 13.1 lb, and the percentage of those in each group achieving the 16 lb weight loss target were 10.5%, 47.4%, and 52.6%, respectively: At that point in time, the weight loss was statistically higher in both incentive groups, as compared with the controls. However, at 7 months, although both the incentives groups still on average weighed significantly less than they did at the initiation of the study, the mean weight losses in the control group, the deposit incentive group, and the lottery incentive group were now 4.4, 6.2, and 9.2 lb, respectively, a nonsignificant difference across the three groups. In both incentives arms, therefore, the earlier effectiveness had dissipated considerably.

More experimentation with different UFI payment mechanism designs, informed by the findings of behavioral economics, is warranted. If they can be shown to effect sustained as well as simple behavior change, then they might prove to be a very useful addition to the preventive health policy armory. For this, however, evidence of effect is necessary, but not sufficient. There are many moral and practical objections to the use of UFI that should not be ignored.

Objections

Moral Objections

Some believe that UFIs are unethical. One argument is that trading money for health (or, presumably, health-related behaviors) involves incommensurable values, in the same way that, for example, selling a child for an electoral vote is unacceptable. The commodification of health-related behaviors, according to this view, may lead to their denigration, devaluation, and/or corruption. UFI may denigrate the person's choices by failing to respect sufficiently the decision that the individual has reached, assuming that he or she has taken into account all of the pros and cons of their actions, which might be particularly dangerous in the field of mental health, where the voice of the patient has often traditionally been ignored.

Thus, there are concerns that UFI, by potentially interfering with the rights of self governance, may undermine some conceptions of fairness and justice. Moreover, in the area of personal lifestyle behaviors, such as smoking, diet, and exercise, it is possible that the general public will resent general tax revenues being used to reward people for doing what the majority believe they should be doing anyway. However, the general acceptance of UFI may prove context specific. For example, paying people to take their medications might garner general societal support, especially if it is perceived that patients underestimate the pros and/or overestimate the cons of medication. If patients do not fully appreciate the benefits of medication, and if their carelessness toward medication poses avoidable harm for themselves or for the communities in which they live, then the societal view may be that the use of UFI is justifiable.

If UFIs are targeted at the relatively poor, as they often are, they could generate further ethical problems, because even small monetary rewards may deter poor patients from terminating treatment when they feel that it is causing them harm. Furthermore, the offer of a financial incentive could be judged as a bribe when directed toward those with limited means. Conversely, it may be contended that UFIs, whereby those targeted can refuse to participate for any reason, are a model of respectful and equal exchange, that the notion of coercion is more usually associated with punishment than reward, and that it is hard for some to accept how a transparent financial inducement to take medicine in order to remain well undermines society's notion of fairness and justice.

Unintended Consequences of UFI

Other objections to UFI focus less on whether they are morally wrong, and more on their potential to have undesirable unintended effects. For example, some worry that the introduction of UFI for some aspects of medication adherence will encourage patients to stop taking their treatment so as to receive the payment for taking their treatment again. Similarly, in relation to broader lifestyle behaviors, it is possible that a few will temporarily initiate unhealthy activities, such as smoking, so as to receive payment for quitting.

A number of experts, across a range of disciplines, have warned that external rewards to act in a particular way may in the long run crowd out the intrinsic desire to alter one's behavior. For instance, if rewards are offered to people to quit smoking, those payments could potentially become an expectation, and thus people might be less willing to abstain under their own personal motivation. Thus, those who are targeted for UFI must be made aware that any financial incentive is there to serve as merely temporary support to help them achieve a personal goal. Similarly, perhaps, financial incentives to discourage particular unhealthy actions might crowd in other undesirable activities: for instance, an ex-smoker motivated by money rather than health may be more likely than a health-aware reformed smoker to substitute doughnuts for cigarettes.

A monetary payment, if administered from the doctor to the patient, might also crowd out the traditional trust-based nature of the doctor–patient relationship, possibly damaging the principal–agent interaction in this regard if some patients no longer follow advice in the absence of financial rewards. Moreover, attempts to change people's lifestyle choices are potentially patronizing and condescending to the targeted group; the obese, for instance, may feel unfairly stigmatized beyond the levels to which they have already been stigmatized. There are clearly a host of objections to the use of UFI; whether or not they are insurmountable in all contexts requires a broad public policy debate.

Conclusion

For some aspects of medical adherence, modest financial incentives can have an effect on people's behavior, although

little is yet known regarding the cost effectiveness of these interventions. However, there is currently little evidence of a positive sustained effect on changing lifestyle behaviors when associated with smoking or weight loss. Nonetheless, all hope in this domain should not yet be abandoned; further research that tests the differential impact – and value for money – of different incentive mechanisms that are informed by the findings of behavioral economics is warranted.

There are many ethical and practical objections to the use of UFI. Although the use of UFI is not intended to threaten individual liberty in that participants self-select into incentives programs, which target behavior change that they themselves are meant to desire, there is perhaps a legitimate concern that these interventions violate spheres of privacy and autonomy. For instance, if UFI are used as a public policy tool, it seems not unreasonable to worry that they may patronize people, and might single out a subset of the population (e.g., the obese), stigmatize them further, and treat them as if they lacked full use of reason. Clearly, more public debate is required in order to reach a consensus on the limits of acceptability to the use of UFI, assuming of course that they prove to demonstrate both reasonable effect and value for money in the areas toward which they are targeted.

See also: Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Pricing and User Fees

Further Reading

Bryan, G., Karlan, D. and Nelson, S. (2012). Commitment devices. *Annual Review of Economics* **2**, 671–698.

Burns, T. (2007). Is it acceptable for people to be paid to adhere to medication? Yes. *British Medical Journal* **335**, 232.

Cahill, K. and Perera, R. (2008). Competitions and incentives for smoking cessation. *Cochrane Database of Systematic Reviews*, **2**, 1–31.

Charness, G. and Gneezy, U. (2009). Incentives and exercise. *Econometrica* **77**, 909–931.

Cookson, R. (2008). Should disadvantaged people be paid to take care of their health? Yes. *British Medical Journal* **337**, a589.

Giuffrida, A. and Torgerson, D. J. (1997). Should we pay the patient? Review of financial incentives to enhance patient compliance. *British Medical Journal* **315**, 703–707.

Gneezy, U., Meier, S. and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives* **25**, 1–21.

Jochelson, K. (2007). *Paying the patient: Improving health using financial incentives*. London: The King's Fund.

Kane, R. L., Johnson, P. E., Town, R. J. and Butler, M. (2004). A structured review of the effect of economic incentives on consumers' preventive behavior. *American Journal of Preventive Medicine* **27**, 327–352.

Marteau, T. M., Ashcroft, R. E. and Oliver, A. (2009). Using financial incentives to achieve health behaviour. *British Medical Journal* **b1415**.

Paul-Ebhohimhen, V. and Avenell, A. (2007). Systematic review of the use of financial incentives in treatments for obesity and overweight. *Obesity Reviews* **9**, 355–367.

Popay, J. (2008). Should disadvantaged people be paid to take care of their health? No. *British Medical Journal* **337**, a594.

Sandel, M. J. (2012). What money can't buy: The moral limits of markets. New York: Farrar, Straus and Giroux.

Seal, K. H., Kral, A. H., Lorvick, J., et al. (2003). A randomized controlled trial of monetary incentives versus outreach to enhance adherence to the hepatitis B vaccine series among injection drug users. *Drug and Alcohol Dependence* **71**, 127–131.

Shaw, J. (2007). Is it acceptable for people to be paid to adhere to medication? No. *British Medical Journal* **335**, 233.

Slater, J. S., Henly, G. A., Ha, C. N., et al. (2005). Effect of direct mail as a population-based strategy to increase mammography use among low-income underinsured women ages 40 to 64 years. *Cancer Epidemiology, Biomarkers and Prevention* **14**, 2346–2352.

Sutherland, K., Christianson, J. B. and Leatherman, S. (2008). Impact of targeted financial incentives on personal health behavior. *Medical Care Research and Review* **65**, 36S–78S.

Volpp, K. G., Troxel, A. B., Pauly, M. V., et al. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *The New England Journal of Medicine* **360**, 699–709.

Relevant Websites

http://www.cabinetoffice.gov.uk/resource-library/applying-behavioural-in sight-health
Cabinet Office.

http://www.kcl.ac.uk/iop/research/centres/csihealth/index.aspx
King's College London.

http://www.stickk.com/
Stickk.com.

http://chibe.upenn.edu/
University of Pennsylvania.

Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs

G Miller, Stanford University, Stanford, CA, USA, and National Bureau of Economic Research, Cambridge, MA, USA
KS Babiarz, Stanford University, Stanford, CA, USA

Glossary

Contracting-out A government practice of contracting with private organizations for health service delivery.
Extrinsic motivation Incentives or motives that originate from without (e.g., financial rewards offered by others).
Intrinsic motivation Incentives or motives that originate from within (deriving enjoyment or personal satisfaction from completing a task, for example).
Multitasking A reduction in effort devoted to noncontracted outcomes when agents are responsible for multiple tasks or multidimensional tasks, but only rewarded for performance on a subset of them.

Introduction

Poor performance of health care providers plagues the delivery of health services in many low- and middle-income countries. The underlying reasons are complex and incompletely understood, but poor performance is not simply due to inadequate training or deficiencies in provider knowledge.

Instead, a growing body of evidence documents substantial deficits in provider effort. One striking example is the high absenteeism rates (as high as 75%) among health professionals documented in a number of studies. When providers are present, a sizeable 'know-do' gap (or failure to do in practice what a provider knows to do in principle) also contributes to low-quality medical care. Provider effort may also not focus directly on improving health – for example, health professionals may provide unnecessary services that are not medically appropriate (e.g., intravenous glucose drips to create the illusion of therapeutic effectiveness). Moreover, even when providers exert appropriate effort during a clinical encounter, they may do little to promote the health of their patients outside of the encounter (e.g., through prevention and outreach activities).

One might expect that given weaker market incentives, these problems would be more prevalent in public sector health service delivery. However, suboptimal provider effort can be sustained in equilibrium in all sectors, including private practice, due to well-known market failures. For example, a well-established literature demonstrates that asymmetric information limits the ability of patients and the lay public to observe provider effort or judge medical care quality. As a result, patients are unable to penalize underperforming providers through their choices. These problems are compounded by market conditions and rigidities common in low- and middle-income countries, including inadequate regulatory processes and a relatively large government role in financing and delivering health services (e.g., given the more prevalent infectious diseases and larger positive externalities in service delivery).

To better align provider incentives with patient and population welfare (or health – one argument of welfare), 'pay-for-performance' schemes have become increasingly common in developing country health service delivery. In principle, the idea is straightforward: drawing on the logic of performance pay in human resource management, this approach rewards providers directly for achieving prespecified performance targets related to health. Use of performance incentives in wealthy countries began in earnest during the 1990s with programs that rewarded both process indicators and measures of clinical quality. Examples of performance targets include immunization rates; disease screening; adherence to clinical guidelines; and the adoption of case management processes, physician reminder systems, and disease registry systems. The UK went further with the National Health Service's Quality and Outcomes Framework, tying physician practice bonuses to a comprehensive range of quality indicators. Performance pay in low- and middle-income country health programs emerged in the late 1990s, and its use has grown rapidly since then.

In practice, pay-for-performance contracts are complex and fraught both with difficult tradeoffs and with the possibility of 'multitasking' and other unintended consequences. This article outlines the key conceptual issues in the design of pay-for-performance contracts and summarizes the existing empirical evidence related to each. In doing so, it focuses on four key conceptual issues: (1) what to reward, (2) who to reward, (3) how to reward, and (4) what perverse incentives might performance rewards create. The article concludes by highlighting important areas for future research and by noting the overall lack of evidence on many key aspects of incentive design in the health sector.

What to Reward

If 'you get what you pay for,' then it presumably follows that one should pay for what one ultimately wants. If a health program's primary objective is good patient or population health outcomes, it would seem natural for performance incentives to reward good health or health improvement directly rather than the use of health services or other health inputs. Rewarding health outcomes rather than health input use not only creates strong incentives for providers to exert effort, but

it can also create incentives for providers to innovate in developing new, context-appropriate delivery strategies. Put differently, rather than tying rewards to prescriptive algorithms for service provision (often developed by those unfamiliar with local conditions), rewarding good health outcomes encourages providers to use their local knowledge creatively in designing new delivery approaches to maximize contracted health outcomes.

In practice, however, very few pay-for-performance schemes have rewarded good health. At the time of writing this review, the authors are aware of only two: performance incentives for primary school principals in rural China to reduce student anemia and incentives for Indian day care workers in urban slums to improve anthropometric indicators of malnutrition among enrolled children. In the Chinese study, researchers measured student hemoglobin concentrations at the beginning of an academic year, issued incentive contracts rewarding anemia reduction to principals shortly afterwards, and measured student hemoglobin concentrations again at the end of the school year. School principals responded creatively, persuading parents to change their children's diets at home as well as providing micronutrient supplementation at school; and anemia prevalence fell by approximately 25%. In the Indian study, researchers measured child anthropometrics at day care facilities, issued incentive contracts rewarding providers for each child with an improved malnutrition score, and repeated anthropometric measurement 3 months later. In response, day care workers visited mothers' homes and promoted the use of nutritious recipe booklets; malnutrition indicators declined by approximately 6%.

The fact that so few pay-for-performance programs reward health outcomes may reflect important limitations to doing so that arise in practice. Instead, performance incentives generally reward the use of prespecified health inputs. In the following sections, shortcomings and tradeoffs inherent in incentive contracts that reward health outcomes will be discussed.

Share of Variation in Contracted Outcome under Provider Control

One drawback to rewarding good health is that even when exerting optimal effort, a relatively small amount of variation in health outcomes may be under the control of providers. For illustration, consider the case of neonatal survival. Maternal health behaviors during pregnancy are key determinants of birth weight, and low birth weight is a leading risk factor for neonatal death. Although rewarding maternity care providers for neonatal survival could in theory motivate them to engage expectant mothers from an early stage of their pregnancy, providers may be unlikely to succeed or may believe a priori that they will be unable to change maternal health behaviors. Patients or community members may also respond to changes in provider behavior caused by performance incentives; in some cases, these responses may undermine provider actions. For example, if educators or health providers take direct action to improve nutrition because of performance incentives, parents may respond by reducing children's dietary quality at home. In both cases, if providers do not believe that their effort will ultimately be rewarded, they may simply not respond to the performance incentive.

A clear (and common) alternative is to reward the use of health services and inputs, particularly those that are relatively sensitive to provider effort. Providers generally have a greater influence on service use than health outcomes, and even more so on quality of clinical care. In several rigorous studies of pay-for-performance incentives in Rwanda, providers were rewarded for prenatal care visits, immunizations among children and pregnant women, institutional deliveries, HIV testing, and a wide range of service quality indicators. Incentive payments were offered for each service provided and were weighted using overall quality scores. This set of incentives motivated providers both to increase delivery of contracted services and to raise the overall quality of care. Researchers observed that the program increased institutional delivery rates by 23% and preventive service use among children under the age of 4 years by 25–50%, and it also reduced the 'know-do' gap by 20%. Although not directly contracted, infant weight-for-age and child (2–4 year old) height-for-age rose by approximately 0.5 and .25 standard deviations, respectively. Researchers also observed that the incentives led to a 15% increase in the rate of HIV testing and counseling among couples, and an 18% increase in the probability that both partners in HIV-discordant households had been tested for HIV at least once.

Interactions with Provider Skill/Human Capital Base

A second limitation to rewarding good health outcomes is that providers may not possess adequate ability to innovate if they lack the necessary skills and human capital. These skills can be both technical and interpersonal. Scholars suggest that providers may be unsuccessful in responding to performance incentives when success requires changing the patient behavior (which requires skills beyond clinical ability). In the Rwandan program, providers were unsuccessful in increasing contraceptive use and in persuading patients to complete the contracted sequence of four prenatal care visits in part because of local patient preferences (superstitions about acknowledging pregnancies at an early stage). To address this shortcoming, one program in the Democratic Republic of Congo paired performance incentives with consulting services for community outreach and business planning. Health facility managers were encouraged to submit quarterly business plans detailing their strategies to achieve incentivized targets, and consultants provided them with custom-tailored advice.

Measurement of Contracted Outcomes

A third obstacle to rewarding good health is that health outcomes can be more difficult and expensive to measure than health service or input use, particularly when physiological health indicators must be measured directly. For example, all else equal, the expense of measuring hemoglobin concentrations would potentially be an important barrier to scaling up the China performance pay program described above. Alternatively, incentives tied to service and input use have successfully relied on combinations of self-reporting and random

audits to measure contracted outcomes on a larger scale. Examples of contracted health inputs measured this way include well baby visits and adherence to clinical protocols during medical visits. Finally, measurement of contracted outcomes (either health input use or health outcomes) among patients in clinical settings may pose fewer measurement challenges than in community wide settings section (Perverse incentives and unintended consequences considers tradeoffs between rewards for outcomes among patients vs. population members).

Who to Reward

Another important issue in designing performance incentives is deciding who to reward. Which agents at what organizational level will be most efficient and effective in improving on contracted outcomes? This section describes conceptual issues in contracting at the macro- (or organizational) and the micro- (individual) level.

Macrolevel Incentives: Organizations and Local Government

At the macrolevel, international organizations (like the Global Alliance for Vaccines and Immunization and the Millennium Challenge Corporation) increasingly use 'results-based financing' in providing aid. Central governments in low- and middle-income countries also frequently contract with private organizations or transfer resources to local government to deliver health services – and performance incentives are often included in these schemes. As performance pay shifts risk to incentivized agents, the more risk averse the agent, the greater the expected compensation must be (all else equal). One advantage of organizational-level incentives is that collectively, organizational agents may effectively be less risk averse than individual employees. This is because idiosyncratic risk that effort will not result in good performance – and thus not be rewarded – is pooled across individuals within organizations. As a result, overall program costs may be lower when contracting at the organizational level (all else equal). However, contracted organizations must then solve their own internal principal–agent problems, and they may pass the costs of doing so on to the principals contracting with them.

There are many circumstances under which central governments 'contract-out' health service delivery to private organizations (typically non-governmental organizations (NGOs)). One is settings in which public sector facilities are largely absent – for example, regions of postconflict Afghanistan and Haiti. Under these conditions, governments and international organizations have contracted with NGOs to open facilities, recruit and train providers, and manage all aspects of service delivery. In the context of Afghanistan and Haiti, achieving performance targets was rewarded with operating budget transfers of up to 10% of the base contract amounts (paid by the World Bank and the US Agency for International Development (USAID), respectively). In Afghanistan, studies found that these contracting strategies were associated with improvements in service availability (measured as the ratio of facilities to population, which increased by approximately 30%, and the share of facilities providing antenatal care, which rose by 45–75%) and institutional delivery rates (which roughly doubled). In Haiti, research suggested that performance pay was associated with 13–24% point increases in full childhood immunization coverage and 17–27% point increases in institutional delivery rates.

Contracting-out also occurs when public sector facilities exist but perform poorly. For example, in 1999, the Cambodian Government began contracting with NGOs to manage health service delivery in five randomly selected districts (eight districts were chosen for contracting, but not all districts had suitable quality proposals from NGOs). Contracts rewarded eight explicit performance indicators (immunization rates, vitamin A supplementation, antenatal care use, medical supervision of deliveries, institutional delivery rates, contraceptive use, and use of public vs. private sector health facilities). Researchers found that after 5 years, performance-based contracting led to a 32% point increase in antenatal care use, a 16% point increase in completion of recommended childhood immunizations, and a 17% point increase in vitamin A supplementation. Cambodia's contracting strategy also improved the general facility operations (24 h service availability, staff attendance, managerial supervision, and equipment availability).

Macrolevel performance incentives have also become increasingly common in the public sector. In the 1980s, central governments in many low- and middle-income countries began transferring funds for service provision to local governments, decentralizing authority over policy design and management. One of the rationales for decentralization is that local governments have superior information about local preferences and are therefore better able to satisfy them. However, even if local governments have superior information about local preferences, they do not necessarily have strong incentives to satisfy them. Decentralization can therefore include performance-based incentives. For example, a recent initiative in Indonesia gave block grants to village leaders to provide maternal and child health services and to run schools. In a randomly selected subset of villages, the size of subsequent block grants was tied to performance according to 12 performance measures (8 maternal and child health indicators and both enrollment and attendance in primary and secondary schools). Scholars found that with performance incentives, midwives in treatment villages worked longer hours, increasing the availability of health services – and prenatal care visits rose by 37% points. Local administrators in incentivized districts also used central government funds more efficiently, negotiating savings in education (without any apparent decline in school attendance) and reallocating the savings to the health sector.

Under all of these circumstances, organizational autonomy may be critical for the success of incentive programs. The Cambodian program experimented explicitly with the degree of independence given to contracting NGOs, using both more restrictive 'contracting-in' and more autonomous 'contracting-out' arrangements. Management and facility indicators improved more in contracting-out districts, and there is suggestion that health indicators did as well. Other cases illustrate the breadth of responses to performance incentives enabled by autonomy. For example, hospitals in Sao Paulo, Brazil with

municipal health delivery contracts that rewarded hospital efficiency, patient volume, and service quality developed creative organizational strategies tailored to their own hospital settings. Hospital spending fell and efficiency indicators rose without measurable declines in service quality; researchers estimate that to produce comparable changes in patient discharges absent performance incentives, hospitals would need to increase spending by approximately 60%.

An important limitation of macrolevel incentives is that they may not translate into private rewards for organizational leaders. Although performance incentives could, in principle, be structured this way, to date they have generally been designed as operating budget transfers and eligibility for future contracts (rewards paid as budget transfers vs. private income is discussed in more detail in Section How to Reward). A related drawback is the possibility that organizational policies and regulations limit organizational or local government ability to solve their own internal principle–agent problems (e.g., if managers are not permitted to use budget transfers for employee bonuses). In Cambodia, contracting NGOs increased the use of many (presumably productive) health inputs, but actual health outcomes (the infant mortality rate and diarrhea incidence among children under 5 years) did not improve. NGOs managing hospitals in Afghanistan and Costa Rica (under similar programs) made improvements in facility management and service provision, but there were no measurable gains in health input use (e.g., immunizations).

Microlevel Incentives

At the microlevel, organizations often use performance incentives to solve principal–agent problems with individual employees. These incentives can target upper-level managers and/or rank-and-file providers that they supervise.

An important virtue of rewarding managers for good performance is that they possess greater flexibility for innovation in service delivery. In contrast, lower-level health workers often must follow detailed, highly prescriptive protocols from which they are not allowed to deviate. For example, a recent study shows that Chinese primary school principals (who manage schools) offered performance rewards for reducing student anemia not only supplemented school meals with vitamins, but they also took the initiative to discuss nutrition with parents, persuading them to increase their children,s consumption of iron-rich foods at home. As a result, anemia prevalence among participating children fell by roughly 25%. In Nicaragua, health facility managers were given performance incentives for offering and providing both prenatal care and well child services to a large share (90%) of local CCT program beneficiaries. In response, managers took the initiative to partner with community organizers (*promotoras*), school teachers, and the local media to conduct community outreach campaigns encouraging mothers to bring their children for checkups. These managerial efforts were reportedly successful: nearly all providers were judged to have achieved the performance targets, preventive care use increased by 16% points, and vaccination rates rose by 30% points.

In practice, many pay-for-performance schemes to date have rewarded individual providers rather than their managers for good performance. Although rank-and-file health workers may have less flexibility to innovate in service delivery, their effort may ultimately matter most for organizational performance. Additionally, because they have the most direct contact with target populations, individual providers may also have better knowledge about local conditions. For example, day care workers in the Indian program rewarding reductions in malnutrition made more frequent home visits in addition to providing more nutritious meals at day care facilities. Through these home visits, they encouraged mothers to use nutritious recipe booklets, and malnutrition among children at their day care centers declined by 4.2% over a 3-month period. In Rwanda, individual public sector providers responded to incentives for higher prenatal care and institutional delivery rates by partnering with midwives to identify and refer pregnant women for services. The associated increase in institutional deliveries was 10–25% points.

In addition to lacking flexibility to innovate in service delivery, there can be other limitations to incentivizing individual health workers as well. A potentially important one is that rewarding health workers for their own individual performance may create disincentives for teamwork or cooperation. Alternatively, rewarding providers for group performance creates incentives for free-riding because individual health workers do not bear the full cost of shirking – and may be rewarded for good performance among coworkers.

How to Reward

Using performance incentives to increase provider effort necessarily requires assumptions about what motivates providers. It is reasonable to assume the providers care about both financial compensation and patient welfare to varying degrees. However, human motives are complex, and other factors undoubtedly play a role too – professional recognition and the esteem of colleagues, pride in one's work, opportunities for professional advancement (career concerns), working conditions, and amenities where one lives, for example. From the standpoint of policy or program design, many of these other factors cannot be translated into performance rewards as easily as financial incentives. However, these other motives can interact with financial incentives in important ways.

This section discusses general conceptual issues in the structure of performance incentive contracts.

Balancing Fixed Versus Variable Compensation

As discussed in the literature outside of health (on executive compensation, for example), performance pay should optimally balance fixed (unconditional) and variable (performance-based) pay. On one hand, performance bonuses must be sufficiently large to influence provider behavior, and on the other aligning executive effort with firm interests may require that a large share of total compensation be tied to firm performance through performance pay. Several studies suggest that in health care, performance incentives may be ineffective if they are too small.

However, increasing variable pay as a share of total compensation increases the financial risk borne by providers. As providers are generally risk-averse (to varying degrees), they must be compensated for bearing additional risk inherent in pay-for-performance contracts. Negotiations over a health service delivery contract in Haiti between an NGO (Management Sciences for Health) and USAID illustrates this point. When renegotiating its contract, Management Sciences for Health was only willing to accept the additional risk imposed by performance pay if USAID would increase the total amount that could be earned to exceed contractual payments under the alternative unconditional contract (under the performance pay contract, fixed payments were set to 95% of the unconditional contract amount, and an additional 10% was made conditional on good performance).

The Functional Form of Provider Rewards

A second issue in the structure of performance pay contracts is the functional form mapping incentive payments onto performance indicators. Absent knowing what the contract theory literature suggests is needed for optimal incentive contract design, a simple approach is to offer rewards that are linear in contracted outcomes. Examples include constant incremental rewards per child reduction in malnutrition, per child reduction in anemia, or per infant delivery supervised by a skilled birth attendant.

Other programs have adopted a step-function approach, offering bonuses for surpassing one or more bright-line performance thresholds. Depending on its specific form, this approach can have theoretical grounding and may also be appropriate when thresholds have clinical significance (e.g., vaccination rates at levels that confer herd immunity). Contract theory suggests that optimal incentive contracts are likely to be nonlinear in contracted outcomes, and step functions could provide a reasonable approximation of these nonlinearities. Some scholars also argue that setting bright-line aspirational goals could change institutional culture to be more results or goal oriented. Although there is little evidence among studies of performance pay, bright-line performance incentives may help to focus attention on contracted outcomes when provider attention is scarce as well.

A drawback to the step-function approach can be a greater risk that provider effort will not be rewarded. Specifically, it creates strong incentives in the neighborhood of a threshold, but it may also be a poor motivator for health workers far below (or above) a threshold. The information required for optimal contract design (including the cost of provider effort, the health productivity of provider effort, and the utility functions of both providers and the contracting principal) is also unlikely to be available in practice.

Salary Versus Operating Budget Rewards

Structuring performance rewards as private income or operational budget revenue also requires assumptions about what motivates providers. In one extreme, if providers were purely motivated by private financial considerations, offering rewards as private income would presumably induce them to exert a greater effort. In the other extreme, if providers were purely philanthropic, incentive payments made as operational revenue could be more effective. Given that preferences are mixed in reality (and also include other things such as professional esteem, pride in one's work, career aspirations, etc.), predictions about the relative effectiveness of different types of financial incentives are ambiguous and may be context specific. One study suggests that NGO employees providing health services in Afghanistan responded markedly to performance incentives even though bonuses accrued to facilities and did not result in personal financial gain. In principle, combinations of the two are possible, although one is unaware of schemes that mixed the two. In practice, macrolevel rewards are often paid as operational revenue, whereas microlevel rewards are typically offered as private income.

Nonfinancial Incentives

Although pay-for-performance contracts strengthen extrinsic incentives, intrinsic motivation is commonly thought to be an important determinant of provider effort as well. Although not focused specifically on health care provider behavior, research on intrinsic motivation in psychology suggests that more altruistic individuals work harder to achieve organizational goals. In the health sector, altruistic individuals are more likely to work for health delivery organizations with explicit charity mandates, suggesting that intrinsic motivation may be heterogeneous across types of health facilities. Health care providers with greater intrinsic motivation may also be more responsive to professional recognition among community members or peers. In such cases, nonfinancial rewards as well as other psychological tools (such as priming, task framing, and cognitive dissonance) may be close substitutes (or may even be more effective) than financial incentives.

Qualitative and anecdotal evidence from field studies support the hypothesis that health care providers are intrinsically motivated. Health workers employed by NGOs in postconflict Afghanistan reportedly felt a great sense of pride and accomplishment after meeting contracted performance targets. A program (not formally evaluated) in Myanmar offered new scales for measuring patient weight to providers who met tuberculosis (TB) case identification and registration goals. In townships with these (essentially) nonfinancial incentives, identification of TB cases rose by 30% points relative to informal comparison townships. Anecdotal reports suggest that Zambian health workers participating in an incentive program (rewarding malaria treatment, infant and maternal care, and childhood immunizations) responded more favorably to trophies than to cash incentives. Finally, case studies suggest that health providers rewarded for good performance with t-shirts, badges and certificates, and recognition photographs may have been successful.

One rigorous quantitative study concurs with this qualitative and anecdotal evidence. In studying Zambian hair stylists with financial and nonfinancial incentives to sell condoms to salon clients, researchers found that public recognition outperforms monetary incentives. These results are heterogeneous across stylists and are largely due to strong behavioral responses among stylists believed to be more committed to the cause of HIV prevention.

Perverse Incentives and Unintended Consequences

The use of incentives to improve health program performance is fraught with the possibility of unintended and potentially perverse consequences. This section discusses some of these concerns and describe the empirical literature related to each.

Noncontracted Outcomes

One type of unintended behavioral response to performance incentives has been studied in the theoretical literature on 'multitasking.' When agents are responsible for multiple tasks or multidimensional tasks (some of which are unobservable or noncontractible), rewarding performance on a subset of contractible tasks or outcomes can lead to a reduction in effort devoted to noncontracted outcomes. The degree to which this occurs may depend in part on the extent to which noncontracted outcomes share inputs with contracted outcomes.

Empirically, some studies of performance incentives have found evidence of such behavioral distortions. A Kenyan school meal program rewarding improved pupil malnutrition rates found that subsidized meal preparation crowded out teaching time by 15%. Similarly, providing incentives to Chinese primary school principals for reductions in student anemia may also have displaced teaching effort, leading to lower test scores in some cases. Findings across empirical studies of performance incentives are heterogeneous, however. Several rigorous studies also report no clear evidence of distortionary or detrimental reallocation of effort or other resources in response to performance incentives.

Beyond the standard multitasking framework, performance incentives may lead to other closely related behavioral distortions. For example, although not studied empirically (to the best of our knowledge), performance incentives could lead to reallocation across multiple substitute activities related to the same disease or health outcome – or even the purposeful neglect of one to earn higher rewards for another (rewarding the successful treatment of a disease would undermine incentives to prevent it). Given the growing emphasis on 'impact evaluation,' another related example would be distortionary reallocation of effort and resources toward an evaluation's primary outcomes (and away from outcomes not emphasized by the evaluation). As demonstrating 'impact' can lead to new or continued funding, the evaluation process itself may therefore create important behavioral distortions (depending on the beliefs of the evaluated organization).

Heterogeneity in the Return to Effort across Contracted Outcomes

Among contracted outcomes, providers may also allocate effort to those that yield the largest (net) marginal return. In Rwanda, researchers found that rewards for good performance were most effective in improving outcomes that appear to have the highest marginal return or require the least effort. For example, performance incentives were more effective in increasing institutional delivery rates among pregnant women already in contact with community health workers (a relatively easy task because new patient relationships did not have to be created) than they were in initiating the use of early prenatal care (which the researchers suggest to be a relatively difficult task because doing so requires early identification of pregnant women not yet in contact with the health care system). Moreover, the incremental payment for institutional deliveries was relatively high (US$4.59), whereas the incremental payment for completion of quarterly prenatal care visits was relatively low (US$0.09). Ultimately, institutional delivery rates rose by more than 20% points, but there were no increases in the share of women completing all quarterly prenatal care visits.

Patient/Subpopulation Selection

In addition to altering how providers choose among tasks, performance incentives may also influence how providers allocate effort among patients or community members. Although not the focus of our review, incentives for patient selection are a ubiquitous concern with the use of high-powered incentives that emphasize cost containment (e.g., capitated contracts under managed care in wealthy countries). With performance incentives for good patient outcomes, selection against the sickest or most remote patients ('cherry picking') may occur if producing contracted outcomes among them is relatively difficult or costly.

Performance could alternatively be linked to population rather than patient outcomes, but providers could then be discouraged from providing services to individuals outside of the predefined population. Similarly, they may simply focus on the easiest to treat subpopulations within their defined service area. Some pay-for-performance schemes have tried to limit perverse incentive like these by offering larger rewards for services provided in more difficult or remote areas. Although such design features may reduce incentives for selection, eliminating them is a nearly impossible task (as the literature on risk adjustment suggests).

Erosion of Intrinsic Motivation

Finally, pay-for-performance incentives may have unintended consequences for the institutional culture of health care organizations and for the intrinsic motivation of individual providers. One study develops a model in which effort in the presence of rewards is a function of intrinsic motivation (operationalized as altruism, but which could also include pride in one's work, etc.), extrinsic motivation (material self-interest), and 'reputational' motivation (related to social- or self-image). In the model, monetary rewards undermine 'reputational' motivation and can therefore crowd-out effort by changing the perceived meaning of one's actions (an 'image-spoiling' effect). Both laboratory and field evidence lend some empirical support to this prediction. In one experiment asking students to perform an altruistic task (collecting charitable donations), evidence suggests that the net effect of small monetary incentives on prosocial effort is negative – students put more effort into the task when they were not compensated than they did when offered a small incentive.

Table 1 Partial list of pay-for-performance programs that have not been formally evaluated

Country	*Year*	*Who to Reward*	*What to Reward*	*How to Reward*	
Latin America					
Argentina		Provincial governments	Volume of poor women and children enrolled in health insurance; performance on 10 health indicators	60% of per-enrollee funding is fixed, 40% linked to performance on 10 targets.	[n]
Belize	2001	Public and private facilities		30% of total capitated service payments are paid monthly with deductions for failure to meet efficiency, quality, and administrative process indicators.	[u]
Costa Rica	1994	Public hospitals	Clinical performance (low delivery complications, low reinfection rates)	Budgetary bonuses	[h]
Honduas		Private hospitals	Health input quality indicators	Payment for each indicator given according to the extent to which the indicator is met (70% performance on a target translates to 70% funding for that indicator)	[n]
Europe/Asia					
Armenia	2008	Primary care providers	Unclear	Unclear	[o]
Bangladesh	2010	Primary care facilities	Infant and maternal care use, postpartum contraception	Unclear if institutional bonuses or provider-level bonuses	[k]
Indonesia	2007	Village governing bodies	12 Health and education indicators	20% of annual block grants determined by village performance on each of 12 contracted indicators	[i]
Nepal	2005	Health workers in public health facilities	Attended deliveries (home or institutional)	US$4.70 for each delivery attended	[l]
Africa					
Benin	2012	Public and private nonprofit health facilities	Maternal and infant health, malaria service use	Salary bonuses	[d]
Burundi	2006	Public health centers and hospitals	24 Specific services	Payment for each contracted service provided; payments weighted (up to 25% additional) for quality; payments up to 80% higher in poor and remote areas	[c]
Cameroon					
Central African Republic	2012	Private providers	Maternal and child health services, technical and capacity building indicators	Quarterly payments to facilities directly, used partly for worker bonuses and general operating budget	[t]
Egypt	2006	Public and private service providers in district provider Organizations	Family planning, immunization	Salary supplements to public and private service providers (up to 275% of base salary)	[e]
Ethiopia	2009	Commuity health workers	Peer and community based health education and outreach	nonfinancial incentives and recognition	[a]
Ghana	early-mid 2000s	NGO sector health workers	Varies across NGO provider	Varies across NGO provider	[f]
Liberia	2008	NGO health systems managers	Six administrative and managerial indicators and 12 targeted services	Operating budget bonuses	[d]
Malawi	2012	Primary care facilities	Quality as measured by a standards-based management and recognition tool		[b]

(*Continued*)

Table 1 Continued

Country	*Year*	*Who to Reward*	*What to Reward*	*How to Reward*	
Mali	2012	Primary care providers	Essential obstetric and newborn care service use	Unclear	[q]
Mozambique	2011	Community health workers	Institutional deliveries, vaccination completion rates, combination of input and output based indicators	Unclear	[r]
Senegal	2012	Public sector hospitals, health management teams, and health centers	Increased care use and quality indicators	Unclear	[p]
Somaliland	2009	Nurses and traditional birth attendants	Institutional deliveries	Nurses received bonuses for each attended delivery; traditional midwives received an incentive for each referral	[j]
South Sudan	2009	NGO health systems managers	Vaccination rates, Vitamin A supplimentation, insecticide treated bed net use, underweight children, staffing, sufficient drug supply, clinical vignette performance	Less than 80% of targets yields 95% contract payment; 80–99% leads to 100% payment; 100% of targets leads to 106% of contract payments	[j]
Tanzania	2011	Public health centers, nonprofit hospitals and dispensaries	Unspecified indicators contracted (set of indicators specified for each type of facility)	Operating budget bonuses	[d]
Uganda	2004	Private nonprofit health facilities	Increased patient volume, prenatal care visits, attended deliveries, immunization rates, contraception use, malaria treatment	Operating budget bonuses	[g]
Zambia	2004	Public providers	Malaria and sexually transmitted infection incidence; prenatal care, attended deliveries, postnatal care, patient satisfaction, immunization rates	Salary bonuses or nonfinancial awards (trophies)	[m]
Zimbabwe	2011	Provincial and district health executives, district hospitals, and rural health centers	Infant and maternal health indicators	Service payments for each service provided; payments weighted by score on quality indicator tool; payments upweighted for delivery of services in remote areas	[s]

[a]Amare, Y. (2011). Non-financial incentives for voluntary community health workers: A qualitative study. *L10K Working Paper No 2.* Addis Ababa, Ethiopia: The Last Ten Kilometers Project, JSI Research & Training Institute.

[b]The Broadbranch Initiative (2012). *Improving maternal and neonatal health in Malawi.* Available at: http://broadbranch.org/BBA/Partners_Projects/Entries/2011/5/19_Improving_Maternal_and_Neonatal_Health_in_Malawi.html (accessed 03.12.12).

[c]Busogoro, J. F. and Beith, A. (2010). *Pay for performance for improved health in Burundi.* Washington, DC: US Agency for International Development.

[d]Ergo, A., and Paina, L. (2012). *Verification in performance based incentive schemes.* Washington, DC: US Agency for International Development.

[e]Huntington, D., Zaky, H. H. M., Shawky, S., and Fattah, F. A. (2009). *Impact of provider incentive payments on reproductive health services in Egypt.* Geneva: World Health Organization.

[f]Lievens, T., Serneels, P, Garabino, S., et al. (2011). Creating incentives to work in Ghana: Results from a qualitative health worker study. *Health, Nutrition and Population Discussion Paper.* Washington, DC: The World Bank.

[g]Lundberg, M. (2008). *Client satisfaction and the perceived quality of primary health care in Uganda.* In: Amin, S. (ed.). *Are you being served? New tools for measuring service delivery.* Washington, DC: World Bank Publications.

[h]McNamara, P. (2005). Quality-based payment: Six case examples. *International Journal for Quality in Health Care*, **17**(4), 357–362.
[i]Morgan, L., Brinkerhoff, D., and Najib, M. (2012). *Community engagement and performance-based incentives: The view from Indonesia*. Washington, DC: US Agency for International Development.
[j]Morgan, L. and Eichler, R. (2011). *Performance-based incentives in Africa: Experiences, challenges and lessons*. Washington, DC: US Agency for International Development.
[k]Population Council. (2010). *Pay for performance (P4P) operations research study*. New York: Population Council.
[l]Powell-Jackson, T. and Hanson, K. (2012). Financial incentives for maternal health: Impact of a national programme in Nepal. *Journal of Health Economics*, **31**: 271–284.
[m]The US Agency for International Development (2006). Zambia pilot study of performacne-based incentives. *Quality Assurance Project, Operations Research Results*. Washington, DC: US Agency for International Development
[n]The US Agency for International Development (2010). *Performance based incentives primer for USAID missions*. Washington, DC: US Agency for International Development.
[o]The US Agency for International Development (2010). *Armenia primary health care reform project*. Washington, DC: US Agency for International Development.
[p]The US Agency for International Development (2012). *Better health systems: Strategies that work*. Washington, DC: US Agency for International Development.
[q]The US Agency for International Development (2012). Performance-based Incentives and quality of maternal-newborn health care in low-resource settings: Opportuities and challenges for performance measurement reesarch. *Meeting Report*. Washington, DC: US Agency for International Development.
[r]The US Global Health Initiative (2011). *Mozambique strategy 2011–2015*. Washington, DC: US Global Health Initiative. Available at: http://www.ghi.gov/documents/organization/175133.pdf
[s]The World Bank (2012). Project information document, appraisal stage: Zimbabwe Health Results Based Financing. *Report no.: AB6635*. Washington, DC: The World Bank.
[t]The World Bank (2012). Results-based financing for health. Washington, DC: The World Bank. Available at: http://www.rbfhealth.org/rbfhealth/content/central-african-republic-car
[u]Vanzie, M., Hsi, N., Beith, A. and Eichler, R. (2010). *Using supply-side pay for performance to strengthen health prevention activities and improve efficiency: The case of Belize*. Washington, DC: US Agency for International Development.

In low- and middle-income countries, there is similar concern that the use of financial incentives may lead to demoralization (due to perceptions of 'bureaucratization'), reductions in intrinsic motivation, and less trust between patients and providers. Over time, the quality of individuals entering the public health workforce could also decline if the use of financial incentives selects against intrinsically motivated health care workers.

Even if extrinsic incentives appear to work in the short run, the erosion of intrinsic motivation can still be a longer-run concern. Psychology experiments suggest that individuals offered monetary incentives to perform an otherwise intrinisically rewarding task put substantially less effort into the task (compared with control groups) when the incentives were removed. This has been attributed to the effect of extrinsic rewards on individuals' perception of themselves, on the value of the rewarded task, and on social perceptions of the task. Although not yet studied in low- and middle-income country health programs, one study of performance pay in the US (at Kaiser Permanente Hospitals) supports these findings.

Conclusion

This article summarizes important conceptual issues in the design of pay-for-performance incentive schemes. These include choice of contracted outcomes, the organizational level at which to offer incentives, the structure of incentive contracts, and what the unintended consequences of performance pay might be. In doing so, the existing peer-reviewed evidence related (in varying degrees) to each was also surveyed. The authors highlight that despite the growing body of research on performance incentives, very little of it has studied the underlying conceptual issues that are outlines (which is critical for the design of better performance incentives). It is also noted that evaluation has not kept pace with growth in the use of performance pay: **Table 1** lists the programs that have not been studied (or studied rigorously) to the best of our knowledge. Strategically selected empirical research on these unstudied programs may provide a low-cost way of strengthening the body of evidence on foundational issues inherent in the design of performance incentives. In the conclusion, we also raise additional issues about which little is known.

The first is that there is substantial heterogeneity in responses to performance pay both across and within programs. The authors therefore caution against direct comparison of pay-for-performance schemes across different organizational, social, and institutional environments. However, it is also noted that understanding the underlying sources of this heterogeneity may provide insight into the circumstances under which performance pay is more or less effective (or socially desirable) too. For example, lack of autonomy among providers or health care organizations may be a critical obstacle to the effective use of performance pay in the public sector (because it restricts the range of behavioral responses that are possible).

Performance incentives may also interact with the preexisting incentives and social norms in important ways. In one study, the impact of performance pay varied across incentivized agents by a factor of three or more (and the underlying source of heterogeneity was not strongly correlated with demographic and socioeconomic characteristics). Another found that provider responses to performance pay varied significantly by baseline provider quality indicators. More generally, adequate bureaucratic capacity to enforce contracts, collect data, and verify performance is presumably necessary for pay-for-performance schemes to succeed. Analysis of heterogeneous responses to performance incentives is an important area for future research.

Second, pay-for-performance schemes may have important equity implications. Given that the net return to provider effort will undoubtedly vary across activities and subpopulations, performance pay may lead providers to focus on individuals with varying socioeconomic or health characteristics. Pay-for-performance contracts offered to village governments in Indonesia attempted to address this concern by allocating equal performance pay budgets across geographic regions with varying socioeconomic characteristics (to prevent some regions from benefitting disproportionately from the performance scheme). Competition among villages for performance rewards therefore occurred within, but not across, regions.

Finally, there has been surprisingly little rigorous empirical evaluation of the full welfare consequences of performance pay. The necessary building blocks for a cost–benefit analysis include a full understanding of the behavioral responses to performance pay and their magnitudes (including unintended ones) and a method for valuing each in common (typically monetary) units. Such evaluations are critical for understanding the ultimate social desirability of pay-for-performance schemes.

Acknowledgment

We are grateful to the National Institutes of Health/National Heart, Lung, and Blood Institute Grant Number R01HL106023 for support.

See also: Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Cost Function Estimates. Development Assistance in Health, Economics of. Economic Evaluation of Public Health Interventions: Methodological Challenges. Efficiency in Health Care, Concepts of. Global Health Initiatives and Financing for Health. Health and Health Care, Macroeconomics of. Health Labor Markets in Developing Countries. Health Services in Low- and Middle-Income Countries: Financing, Payment, and Provision. Health Status in the Developing World, Determinants of. Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity. Markets in Health Care. Pay for Prevention. Physician-Induced Demand. Physician Management of Demand at the Point of Care. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Primary Care, Gatekeeping, and Incentives. Priority Setting in Public Health. Public Health in Resource Poor Settings. Public Health: Overview. Public Health Profession. Resource Allocation Funding Formulae, Efficiency of. Theory of System Level Efficiency in Health Care

Further Reading

Ashraf, N., Bandiera, O. and Jack, K. (2012). No margin, no mission? A field experiment on incentives for pro-social tasks. *Working Paper*. Cambridge: Harvard Business School.

Basinga, P., Gertler, P. J., Binagwaho, A., et al. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: An impact evaluation. *Lancet* **377**(9775), 1421–1428.

Bloom, E., Bhushan, I., Clingingsmith, D., et al. (2006). *Contracting for health: Evidence from Cambodia. Cambridge*: Harvard University.

Das, J. and Hammer, J. (2007). Money for nothing: The dire straits of medical practice in Delhi, India. *Journal of development economics* **83**(1), 1–36.

Eichler, R. and Levine, R. (eds.) (2009). *Performance incentives for global health: Potential and pitfalls*. Washington, DC: Center for Global Development.

Ellis, R. P. and McGuire, T. G. (1993). Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives* **7**(4), 135–151.

Gertler, P. and Vermeersch, C. (2012). Using performance incentives to improve health outcomes. *World Bank Policy Research Working Paper No. 6100*, Impact Evaluation Series No. 60. Washington, DC: The World Bank.

Gneezy, U., Meier, S. and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives* **25**(4), 191–209.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics & Organization* **7**, 24–52.

Loevinsohn, B. and Harding, A. (2005). Buying results? Contracting for health service delivery in developing countries. *Lancet* **366**(9486), 676–681.

Miller, G., Luo, R., Zhang, L., et al. (2012). Effectiveness of provider incentives for anaemia reduction in rural China: A cluster randomised trial. *British Medical Journal* **345**, e4809.

Olken, B. A., Onishi, J. and Wong, S. (2012). Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia. *National Bureau of Economic Research Working Paper, w17892*. Cambridge: National Bureau of Economic Research.

Singh, P. (2011). Performance pay and information: Reducing child malnutrition in urban slums. *MPRA Working Paper*. Munich: Munich Personal RePEc Archive.

Soeters, R., Peerenboom, P. B., Mushagalusa, P. and Kimanuka, C. (2011). Performance-based financing experiment improved health care in the democratic republic of Congo. *Health Affairs* **30**(8), 1518–1527.

Witter, S., Fretheim, A., Kessy, F. L. and Lindahl, A. K. (2012). Paying for performance to improve the delivery of health interventions in low-and middle-income countries. Cochrane Database of Systematic Reviews, Issue 2 Art. No.: CD007899. DOI: 10.1002/14651858.CD007899.pub2.

Peer Effects in Health Behaviors

JM Fletcher, Yale School of Public Health, New Haven, CT, USA

Introduction

Health economists have long been interested in examining the determinants of, and potential policies for, reducing unhealthy behaviors in the population. Although a main focus in this area has historically been on issues of policy involving taxation, access restrictions, advertising, etc., a shift toward evaluating the basic social or nonmarket determinants of unhealthy behaviors has occurred in the literature. This is perhaps most obvious in the research regarding children and adolescent behaviors, wherein peer pressure is often thought to play a substantial role in determining choices such as smoking decisions. Indeed, there is now a large and growing literature in health economics that asks variants of the question, "Do peers influence an individual's health behavior decisions?" The areas of interest range from substance use to eating behaviors and weight outcomes whereas the peer group definitions range from best friends to classmates, residential neighbors, and beyond. Indeed, although there have been several recent reviews of the literature examining social effects on health behaviors (Fletcher, 2010a, 2011a), these papers are being updated because of the rapid expansion in research in this area. This new research is based on previous work; moreover, quasi-experimental methods as well as new identification strategies are utilized in the research.

One reason for the increasing interest in achieving these research milestones is the policy implications of the existence of peer effects in health behaviors. Specifically, peer effects often imply a 'social multiplier' for interventions – if the health of one individual is increased, the effect of the intervention may be multiplied through peers. This type of social effect is seen as an 'endogenous social effect' in the literature of economics. In contrast, peer effects that operate through the characteristics of peers are labeled 'exogenous social effects or contextual effects' (see Manski, 1993). The presence of such endogenous social effects could increase the potential benefits of intervention without increasing the costs. In contrast to these benefits, the presence of peer effects could also work to spread unhealthy behaviors (such as smoking). The awareness of a social multiplier operating in determining health can also help to inform whether targeted (e.g., based on influential individuals within networks) or broad-based policy is more effective. Also, peer effects imply that the composition of a person's neighborhood and/or school could affect his/her health behavior; because many policies can reorganize peer groups, such as school ability grouping (tracking), busing, school grade-span configuration, and residential zoning, there are a host of potentially important policy domains that plays a role in reducing poor health behavior in the presence of peer effects.

Although research regarding health decisions of socially connected individuals may continue to expand along with the multiple growth of interaction within social ties, the empirical hurdles to credibly estimating peer influence remain relatively unchanged and difficult to overcome. This article discusses some of the general empirical issues with their brief history, current controversies, and future directions.

Empirical Issues

Just as the policy and health importance of peer effects is likely to be substantial, so too is the empirical difficulty of credibly estimating causal effects. There are (at least) four standard primary empirical issues that researchers face. In many empirical settings, some are generic problems of measurement and omitted variables, whereas others are somewhat specific to peer effects research.

First, researchers must define a relevant peer group. This step seems simple, but data limitations typically force researchers to define peers on the basis of convenience rather than on theory. This has created peer group definitions that range from state-based groups to nominated best friends, and everything in between. For example, Harding (2003) uses census tracts, Evans *et al.* (1992) use metropolitan level data, Case and Katz (1991) use city block level data, Fletcher (2010a) uses school grades, Fletcher (2010b) uses school classrooms, Mayer and Puller (2008) use 'Facebook Friends,' and Sacerdote (2001) uses roommates to create relevant reference groups for the outcomes to be examined. Although there are several data sets that include nominated friends and peers, the vast majority do not. New data sets may reduce this issue over time, particularly those collecting online social network data, but this will raise the issue of whether online social contacts represent an important and relevant peer group for the determination of health decisions, and if they do, what types of health decisions are relevant when considering online peers.

A second empirical difficulty is the endogeneity of peer groups. Does a person smoke because his friend smokes or did he choose his friend for the sake of smoking? Because individuals typically have some degree of choice over their interaction with others (schoolmates, neighbors, friends, etc.), separating peer selection from peer influence is a particularly difficult empirical problem, and peer selection effects would typically inflate standard estimates of 'peer effects.' In fact, there seems to be a 'relevance-endogeneity' trade-off between the first and the second empirical difficulties (Fletcher, 2010a). As the researcher broadens the definition of the peer group (such as pertaining to the state level), the endogeneity of the peer group probably diminishes, but the relevance of the peer group may weaken. In contrast, best friends are probably a relevant definition of a peer group for many health behaviors but the endogeneity of best friend is magnified.

A third empirical difficulty in peer effect research lies in its potential nature for omitted variable bias through shared influences. For example, smoking bans may reduce tobacco use in all members of a school-based or community-based

peer group. These shared factors can lead to inflated estimates of peer effects if sufficient control variables are not included.

A fourth empirical difficulty in peer effects research is the reflection problem (Manski, 1993), where the researcher may be unable to distinguish between whether Bill influences Ted or Ted influences Bill. Although it is not essential to disentangle these two influences in order to establish whether there is any social effect for determining health behaviors, it can be useful to separate these effects in order to understand the importance of the initial causal effect as against the feedback effects to further understand the processes of health spillovers. Although most researchers explicitly acknowledge each of these difficulties, they often adopt different approaches in attempting to overcome them.

Indeed, there is a two-decade-old history examining peer effects in many health behaviors, which can provide some examples of the difficulty with this research topic while outlining ways that other researchers have attempted to circumvent the empirical issues as outlined above. Typically, researchers have used neighborhood or school-based definition of 'peers' when examining health behaviors such as tobacco, alcohol, and drug use.

A Brief History of Empirical Approaches

Case and Katz (1991) provide a seminal look at the effects of neighborhood peers on risky behaviors and other outcomes, although they are unable to tackle many of the aforementioned empirical issues. In particular, the authors acknowledge that they are unable to control for all environmental confounders and their self-selection into neighborhoods. The authors use what has become a typical empirical framework in the literature:

$$Y_{ig} = X_{ig}B + \overline{X}_{-ig}\delta + W_g\theta + \alpha\overline{Y}_{-ig} + \varepsilon_{ig} \quad [1]$$

where, Y_{ig} is the health behavior choice of individual i in peer group g (e.g., neighborhood), individual and family characteristics are contained in a vector X, and peer characteristics are measured as group-level averages of the X vector excluding the individual, labeled $\overline{X}_{-ig}$. Unobserved factors are contained in the vector, W_g. Finally, $\overline{Y}_{-ig}$ is the group-level average outcome excluding the individual (e.g., the proportion of individuals in the same neighborhood who report smoking). The main coefficient of interest is the endogenous effect α, which indicates the extent to which individuals are influenced by their peers' choices. If α is positive, interventions that change the behavior of individuals (or subsets of individuals) within a reference group would be predicted to spillover on nontreated individuals in the same reference group. In addition to acknowledging the potential for omitted group-level variables as well as self-selection (where ε_{ig} and $\overline{Y}_{-ig}$ are correlated), the authors are also unable to resolve the simultaneity bias (this issue was not fully discussed until Manski, 1993). The authors find evidence of substantial correlation between own and neighborhood peer substance use, crime, and other behaviors.

Norton *et al.* (1998) focus on schoolmate peer effects in alcohol and tobacco use of teenagers, and they use an instrumental variables strategy to address the endogeneity of peer groups (see also Evans *et al.* (1992) for an analysis of teenage pregnancy). Although the focus on endogeneity is important, there is little scope to control for the shared environment due to both data limitations and the instruments (such as neighborhood drug availability and safety) being potentially invalid – in fact, the results have suggested that noninstrumented results are preferable for extremely large peer effects. The general approach of using schoolmates or grademates has been used by many subsequent studies (e.g., Gaviria and Raphael, 2001), wherein too the quality of the instruments are uncertain; specifically, all contextual effects are often assumed to not exist in order to use these variables as instruments.

More recently, Fletcher (2010a) has suggested this approach to be inappropriate and instead proposes a combined instrumental variables/fixed effects design with conceptually appealing diagnostic tests (following Bifulco *et al.*, 2011; Lavy and Schlosser, 2007) in order to validate a preferred instrument set, although the validity of the instruments is still widely questioned. Specifically, Fletcher argues that the increasing proportion of the smoking grademates is due to smoking status of individuals in their households (which can be empirically demonstrated), which does not directly affect respondent smoking even when school-fixed effects are controlled (which is a maintained, untestable assumption). Although Fletcher shows the evidence that exposure of smoking grademates from households of smokers is conditionally random within school, there are ways by which this instrument could be invalidated because, for example, if mothers of grademates are smokers, it simply implies that there is access to tobacco for the respondent. See also Fletcher (2011b) for an examination of peer influences in alcohol consumption.

There have been several alternatives to the instrumental variable approach in the literature. Clark and Loheac (2007) use panel data and a lagged measure of peer behaviors that is combined with school-fixed effects in order to adjust for endogeneity, a large portion of the shared environment, and the reflection problem:

$$Y_{igt} = X_{ig}B + \overline{X}_{-ig}\delta + W_g\theta + \alpha\overline{Y}_{-igt-1} + \varepsilon_{igt} \quad [2]$$

The reflection problem is eliminated because current smoking decisions cannot affect past schoolmate smoking decisions. Although school-fixed effects reduce the issue of contextual effects, a maintained assumption is that, within schools, students choose friends randomly. A second weakness of this design is the need to assume a specific time structure where individual decision making and social influence processes are concerned (e.g., 1 day, 1 week, 1 month, 1 year, 2 years, etc.) (Manski, 1995). Specifically, Manski (1995, p. 136) states, "Of course, one cannot simply specify a dynamic model and claim that the problem of inference on social effects has been resolved. Dynamic analysis is meaningful only if one has reason to believe that the transmission of social effects follows the assumed temporal pattern."

An alternative to implementing a lag structure research design or an instrumental variables strategy is to focus on estimating contextual social effects instead of endogenous social effects. The most convincing work in this area uses random assignment of peers. For example, Kremer and Levy

(2008) use data from a university that randomly assigns freshmen to shared dormitory rooms:

$$Y_{ig} = X_{ig}B + \overline{X}_{-ig}\delta + W_g\theta + \varepsilon_{igt} \quad [3]$$

where, in this case, $\overline{X}_{-ig}$ could be thought of as a lagged endogenous social effect examined in some studies or roommate's precollege alcohol consumption. What allows the estimate to produce a contextual effect rather than an endogenous one is that the individual is not exposed to the actual drinking behavior, but rather is being exposed to having a roommate who has the characteristic of being a past drinker. Additionally, the random assignment of roommates eliminates the concerns regarding the endogeneity of the peer group. Kremer and Levy show that a fresh student who is randomly assigned a roommate with alcoholic past during high school has lower college performance than the student who is assigned a nondrinking roommate. The focus on the roommate's predetermined high school drinking behavior as the peer effect of interest also eliminates issues of simultaneity bias.

Because not all data sets are able to leverage the random assignment of 'friends,' several studies attempt to leverage quasi-random variation in observational data. For example, Bifulco *et al.* (2011) use a cross-cohort, within-school design to link the outcomes of students to their (quasi-randomly assigned) classmates' characteristics:

$$Y_{ig} = X_{ig}B + \overline{X}_{-ig}\delta + W_g\theta + \varepsilon_{igt} \quad [4]$$

That is, the authors examine the 'peer effects' of having a higher share of grademates with educated mothers or a higher share of grademates who are racial/ethnic minorities. This focus on contextual effects sidelines the need for a solution to the reflection problem because student smoking cannot affect grademate race, but some of the important policy issues that are tied to a social multiplier through endogenous peer effects cannot be evaluated directly.

Newer Approaches and Extension of Outcomes

Although the more traditional literature examining peer effects and health behaviors has focused primarily on substance use outcomes and has used a range of empirical approaches, the more recent literature in this area has broadened research designs and has dramatically expanded the range of health outcomes under study – especially, weight and mental health outcomes.

Apart from the literature in health economics, the set of studies that has received the most media attention is from Nicholas Christakis, James Fowler, and a set of coauthors. Their first study has brought a new outcome of interest to the literature by examining whether obesity is 'socially contagious.' Specifically, the authors have found that the chances of an individual becoming obese increased by more than 50% when his/her friend has become obese. The authors have used the Framingham Heart Study data, which contain up to 32 years of longitudinal measures of BMI for individuals in one area of Massachusetts. To these data, the authors have matched information from the original respondents' records, on which respondents have individually been asked to name a person who can be contacted in case the survey team does not reach them directly at follow up; this contact person is treated by Christakis and Fowler (2007) as a 'friend.' Thus, the first issue with this research is whether the contact person is truly a peer. The authors estimate regressions using the following parsimonious empirical model:

$$\text{health}_{it} = \delta\,\text{health}_{jt} + \beta_1\,\text{health}_{it-1} + \beta_2\,\text{health}_{jt-1} + \beta_3 X_{it} + \varepsilon_{it} \quad [5]$$

where, the health (obesity) of person i is linked to person j and δ is the coefficient of interest – the endogenous social effect or 'social multiplier.' A positive estimate on δ suggests that an intervention which reduces the chances of an individual becoming obese will also reduce the chances of obesity in his/her peer.

To overcome endogeneity, Christakis and Fowler (2007) have assumed that lagged health outcomes for the friend (health_{jt-1}) is a sufficient control, that is, after controlling for lagged obese status of a friend, they have assumed that there is no additional issue of friendship selection. Unfortunately, to the extent that this control variable does not completely eliminate selection effects, the estimated coefficient of interest (δ) will probably be upwardly biased. The authors have controlled for own-lagged health in order to control for aspects of the individual's genetic disposition or other time-invariant characteristics. The second issue is confounding due to shared influences. Without explicitly controlling for shared environmental factors, the authors have appealed to a comparison between mutually nominated friends and nonmutual friends (with unreciprocated nomination), arguing that directionality of nominations does not matter if environmental confounding is the primary explanation. Finally, the authors neither discussed nor attempted to overcome the empirical complications from the reflection problem. Unfortunately, each of those empirical issues listed above would probably lead to upwardly biased estimates of peer effects. So, what proportion of the 50% estimated peer effect is due to bias and what proportion is an actual peer effect? To address these empirical concerns, Cohen-Cole and Fletcher (2008a) have provided an examination focusing on one of the empirical issues in peer effects models – shared environmental factors that may bias upward the estimates. The authors have used the National Longitudinal Study of Adolescent Health (Add Health), which includes the nationwide longitudinal data on adolescents in the US over approximately seven years. Although the Framingham study has a much longer time horizon and focuses on adults, the Add Health data contain information on actual 'best friends' who are named by the respondent; this is arguably a more appropriate peer than the contact person in the Framingham data. Cohen-Cole and Fletcher have first estimated eqn [5] on the basis of the Add Health data in order to replicate the baseline findings of Christakis and Fowler (2007) that are based on the Framingham data. Interestingly, both papers, using different data of different age groups, arrive at point estimates for δ from eqn [1] for the 'peer effect' of BMI of 0.05, meaning that a one unit increase in a friend's BMI over time is correlated with a 0.05 unit increase in one's own BMI. However, when

Cohen-Cole and Fletcher controlled for shared environmental factors such as school-fixed effects, the coefficient fell by approximately 40%, no longer being statistically significant. Thus, current evidence suggests that the empirical problems described above are problematic enough to reduce confidence in any peer effects in obesity resulting from this specific model.

In an attempt to further explore the potential upward bias in the Christakis/Fowler empirical model, Cohen-Cole and Fletcher (2008b) took an alternative approach. The authors asked the question: 'Is the empirical model [5] so weak that it would produce estimates of peer effects in behaviors where the true peer effect should be zero?', that is, the authors conducted a falsification test of the empirical model by showing that estimating eqn [5] with the Add Health data would also produce results suggesting 'social contagion' in outcomes that are unlikely to be contagious: acne, headaches, and height. Indeed, the estimates for peer effects in these health behaviors are in some case larger than the Christakis/Fowler estimates of peer effects in obesity. The results of the falsification exercise strongly suggest that the model is insufficiently specific to distinguish between true social effects and the alternative hypotheses as discussed above (e.g., endogeneity of friendships and exposure to shared environmental factors). As in previous work, Cohen-Cole and Fletcher (2008b) have shown that the magnitudes of the fictional social network effects are reduced and when shared environmental influences are controlled, these effects largely disappear.

Based on part on these findings, obesity and weight-related behaviors have been studied in several additional papers. Trogdon *et al.* (2008) use several empirical strategies to examine peer effects. They examine both grade-level peers, similar to the cross-cohort designs already discussed, as well as nominated friends. To control for shared environmental factors, the authors control for school-fixed effects. To address friendship selection and simultaneity bias, the authors use an instrumental variables strategy, where friend's birthweight, weight of parents of friend, and other measures are used as instruments. The limitation with this approach is that it is unclear whether these variables are good instruments for friendship selection. It appears that the instruments have been mainly employed to reduce the importance of the simultaneity issue, though the instruments still need to be excludable from the equation determining one's own weight. In addition to controlling for shared environmental influences, the authors use school-fixed effects to partially control for friendship selection. The implicit assumption with school-fixed effects is that within schools, friendships form randomly.

Like Trogdon *et al.* (2008), Renna *et al.* (2008) also use a single cross-section of the Add Health data to examine the correlation between own and friend's weight outcomes; however, these two papers use different subsamples and Renna *et al.* focus only on nominated friends. Renna *et al.* use school-level fixed effects to control for shared environmental factors and also attempt to reduce the simultaneity issue with an instrumental variables approach. The authors also use the obesity status of parents of friends as instruments. To control for selection of friends, the authors include additional control variables and acknowledge that the estimates are likely to be biased upward. The authors find evidence for peer effects for both genders in the baseline models, whereas only females in the IV models, although the point estimates are very similar. Overall, these papers are suggestive of peer effects but are unable to control for the empirical issues necessary to make the evidence more conclusive.

However, three recent papers have attempted to overcome the methodological issues with the above papers by pursuing alternative research designs. Yakusheva *et al.* (2011) use the roommate design described above with females from a private Midwestern university. The authors show negative correlations between having a heavy roommate and own weight outcomes. Carrell *et al.* (2011) stretch the literature further into the outcome of physical fitness by using random assignment to squadrons in the US Air Force Academy in order to show that squadronmates' level of physical fitness is highly correlated with one's own fitness. Finally, using a new instrumental variable strategy (characteristics of friend of peer), a so-called 'friend of friend' instrument pioneered by Bramoullé *et al.* (2009) and Fortin and Yazbeck (2011) show some evidence of peer effects in fast-food consumption. Although these papers have considerably strengthened the research designs from past work and have extended the set of health behaviors, additional work is needed to further understand the potential for whether obesity is indeed 'socially contagious.' This work requires different (and hopefully more representative) samples and further replication.

In addition to weight-related outcomes, the literature examining peer influences in health outcomes has also begun to examine the realm of mental health. Although some older papers have attempted to examine social influences on suicidal behaviors, this literature is yet to incorporate newer and more rigorous research designs. Hence, the existence of peer effects is still uncertain. However, other measures of mental health have been explored recently. Eisenberg *et al.* (2011) have applied the roommate design to a variety of anxiety and depressive symptoms using a sample of freshman college students from two universities. The authors find no evidence of peer influence in measures of happiness. However, symptoms of anxiety appear to be correlated between roommates and there is some suggestive evidence of depressive symptoms being correlated between male roommates. See also Fletcher (2010c) for evidence that classmate mental health may reduce school performance.

Considering that research has expanded the domain of health behavior under study, new directions have been adopted in empirical methods on the basis of nonexperimental data. For example, a new direction in the study of social networks with implications for the study of health is the analysis of interdependent duration decisions. Because many health outcomes and behaviors have important time components such as smoking and drinking histories, utilizing new methods in this area could prove useful. The current state of the art includes the theoretical framework as outlined in Brock and Durlauf (2008) as well as the empirical applications of de Paula (2009) and de Paula and Honore (2010). Likewise, Fletcher and Ross (2012) have attempted to combine a control function approach with a cross-cohort design (as outlined above) to estimate the effects of best friend's smoking and drinking behaviors on individual health choices. Work by Yves Zenou and colleagues have accumulated a set of papers that build a game theoretic model of network formation with interesting empirical implications (e.g., Calvó-Armengol *et al.*, 2009).

In addition to these new research methods, there are also new data opportunities as well as research design opportunities emerging. Mayer and Puller (2008) leverage data from the social networking website Facebook.com in order to examine correlations between friends' health behaviors, but they do not focus on causal inference. Mapping friendship networks through the use of cell phone usage information may also transform our ability to construct and track social networks in the future (Eagle *et al.*, 2009). However, these new data sources will not alleviate the need to confront the difficulties of estimating empirical models of social influence.

Conclusions

This article briefly outlines some of the history, empirical challenges, current research and controversies, and directions for the future of this research area. Indeed, it is important to point out that this article being necessarily unexhaustive, does not cover important research areas that share many of the same issues described herein. Perhaps most obvious are the exclusion of the neighborhood effects literature that focuses on health outcomes and the emerging literature that examines potential peer effects in other health related areas such as doctors' prescribing patterns and medical technology adoption. These important areas are beyond the scope of this article, which focuses only on the peer effects in health behaviors.

Research in the economics literature examining the effects of peers on health behaviors is now more than two decades old. There has been impressive progress as well as a stable set of challenges still not fully resolved. There has been a broadening of the set of behaviors and outcomes under consideration including weight and mental health besides the use of quasi-experimental research designs for additional outcomes of interest. In contrast, the growing volume of data on peer influence, especially online, has not been met by new research designs and methodologies that can produce entirely convincing results. This will be an important challenge to the researchers' work of expanding the literature on peer effects in health behaviors.

Acknowledgments

This article extends similar summaries of current work on peer effects/social network effects in health behaviors found in Fletcher (2010a, 2011a). The author thanks the Robert Wood Johnson Foundation Health & Society Scholars program for its financial support.

See also: Smoking, Economics of

References

Bifulco, R., Fletcher, J. M. and Ross, S. L. (2011). The effect of classmate characteristics on individual outcomes: Evidence from the Add Health. *American Economic Journal: Economic Policy* **3**(1), 25–53.

Bramoullé, Y., Djebbari, H. and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics* **150**(1), 41–55.

Brock, A. and Steven Durlauf (2008). *Adoption curves and social interactions.* NBER Working Paper 15065.

Calvó-Armengol, A., Patacchini, E. and Zenou., Y. (2009). Peer effects and social networks in education. *Review of Economic Studies* **76**, 1239–1267.

Carrell, S., Hoekstra, M. and West, J. (2011). Is poor fitness contagious? Evidence from randomly assigned friends. *Journal of Public Economics* **95**(7–8), 657–663.

Case A. and L. Katz (1991). *The company you keep: The effects of family and neighborhood on disadvantaged youth.* NBER Working Paper 3705.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **357**, 370–379.

Clark, A. and Loheac, Y. (2007). It wasn't me, it was them! social influence in risky behavior by adolescents. *Journal of Health Economics* **26**(4), 763–784.

Cohen-Cole, E. and Fletcher, J. M. (2008a). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics* **27**(5), 1382–1387.

Cohen-Cole, E. and Fletcher, J. M. (2008b). Detecting implausible social network effects in acne, height, and headaches longitudinal analysis. British Medical Journal **337**, a2533.

Eagle, N. A., Pentland and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* **106**(36), 15274–15278.

Eisenberg, Daniel, Golberstein, Ezra, Whitlock, L. and Downs., F. (2011). *Social contagion of mental health: Evidence from college roommates.* University of Michigan Working Paper.

Evans, W., Oates, W. and Schwab., R. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy* **100**, 966–991.

Fletcher, J. M. (2010a). Social networks and health. *New palgrave dictionary of economics.* Available at: http://www.dictionaryofeconomics.com/article?id=pde2010_S0005&edition=current&q=fletcher&topicid=&result_number=1 (accessed 15.07.13).

Fletcher, J. M. (2010b). Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Economics* **19**(4), 466–484.

Fletcher, J. M. (2010c). Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school. *Journal of Policy Analysis and Management* **29**(1), 69–83.

Fletcher, J. M. (2011a). Peer effects in obesity. In Cawley, J. (ed.) *Handbook of the social science of obesity.* Available at: http://www.oup.com/us/catalog/general/subject/Economics/Health/?view=usa&ci=9780199736362 (accessed 15.07.13).

Fletcher, J. M. (2011b). Peer influences on adolescent alcohol consumption: Evidence using an instrumental variables/fixed effect approach. *Journal of Population Economics* **25**(4), 1265–1286.

Fletcher, M. and Ross, L. (2012). *Estimating the effects of friendship networks on health behaviors of adolescents.* NBER Working Paper 18253.

Fortin, B. and Yazbeck M. (2011). *Peer effect, fast food consumption, and adolescent weight gain.* SSRN Working Paper. Available at: http://papers.ssrn.com/sol3/papers.cfmabstract_id=1759978 (accessed 01.03.10).

Gaviria, A. and Raphael, S. (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics* **83**(2), 257–268.

Harding, D. J. (2003). Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology* **109**(3), 676–719.

Kremer, M. and Levy, D. M. (2008). Peer effects and alcohol use among college students. *Journal of Economic Perspectives* **22**(3), 189–206.

Lavy V. and Schlosser, A. (2007). *Mechanisms and impacts of gender peer effects at school.* NBER Working Paper 13292

Manski, C. (1995). *Identification problems in the social sciences.* Cambridge, MA: Harvard University Press.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* **60**(3), 531–542.

Mayer, A. and Puller, S. (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics* **92**, 329–347.

Norton, E. C., Lindrooth, R. C. and Ennett, S. T. (1998). Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use. *Health Economics* **7**(5), 439–453.

de Paula, A. (2009). Inference in a synchronization game with social interactions. *Journal of Econometrics* **148**(1), 56–71.

de Paula, A. and Honore, B. (2010). Interdependent durations. *Review of Economic Studies* **77**(3), 1138–1163.

Renna, F., Grafova, I. B. and Thakur., N. (2008). The effect of friends on adolescent body weight. *Economics and Human Biology* **6**, 377–387.
Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *Quarterly Journal of Economics* **116**(2), 681–704.
Trogdon, J., Nonnemaker, J. and Pais, J. (2008). Peer effects in adolescent overweight. *Journal of Health Economics* **27**(5), 1388–1399.
Yakusheva, O., Kapinos, K. and Weiss, M. (2011). Peer effects and the freshman 15: Evidence from a natural experiment. *Economics & Human Biology* **9**(2), 119–132.

Further Reading

Cohen-Cole, E. and Fletcher J. M. (2008). *Estimating peer effects in health outcomes: Replies and corrections to fowler and christakis*. SSRN Working Paper. Available at: http://papers.ssrn.com/sol3/papers.cfmabstract_id=1262249 (accessed 01.03.10).
Halliday, T. J. and Kwak, S. (2009). Weight gain in adolescents and their peers. *Economics and Human Biology* **7**(2), 181–190.
Lundborg, P. (2006). Having the wrong friends? Peer effects in adolescent substance use. *Journal of Health Economics* **25**(2), 214–233.
Powell, L., Taurus, J. and Ross, H. (2005). The importance of peer effects, cigarette prices, and tobacco control policies for youth smoking behavior. *Journal of Health Economics* **24**, 950–968.

Peer Effects, Social Networks, and Healthcare Demand

JN Rosenquist, Harvard Medical School, Boston, MA, USA
SF Lehrer, Queen's University, Kingston, ON, Canada

Glossary

Actor (or node) An individual actor or agent in the network.
Homophily The tendency of individuals with similar characteristics to associate with one another. Also referred to as selection.
Peer Effects Actions, preferences, and norms of an individual's peer group that may influence an individual's behavior.
Social network A set of actors and relationships (or ties) linking the actors together. Social networks can be used to study the structure of a social organization and how this structure influences the behavior of individual actors.
Tie A tie connects 2 nodes. In social network analysis, a tie is determined through a self-reported and/or observed social tie.

Overview

The notion that social influences are important in both the development of numerous health outcomes and decisions related to healthcare use has great intuitive appeal. Health economists have modeled many potential mechanisms through which social influences affect outcomes ranging from practice patterns among providers to the choice to engage in risky behaviors. Effectively understanding the nature and scope of these effects is critical; if such influences are ignored estimates of the impact of policy interventions will in many cases be biased because they neglect the indirect pathway that occurs due to spillovers or what is known as the social multiplier effects. There have been many recent theoretical and empirical developments in this area on this topic. For example, studies on animal populations have shown that there is an association between social status and increased odds of specific diseases due to biochemical responses to low status affecting a creature's immune system. More abstractly, economic theorists are increasingly directly modeling concerns for social status in the specification of utility function because this is as important an aspect of human decision-making potentially including the choice of medicine as one's occupation.

Research on social interactions first appeared in the economics literature with Veblen's analysis of conspicuous consumption, where consumption levels are used to signal wealth. Formal analysis of the influence of social groups in economic models developed following Schelling (1971) who demonstrated how the existence of these interactions may result in the formation of ghettos and segregation of individuals across neighborhoods, even in situations where most individuals prefer living in an integrated neighborhood. Formal models generally include social interactions by allowing for strategic complementarities, which occur when the marginal utility to one person of undertaking an action is increasing with the average amount of action taken by one's peers. Although the initial developments in this literature were primarily made by theorists, there has been both a growing body of empirical work and policies proposed to take advantage of social interactions in health-related behaviors. This research is summarized in the following text.

Definitions of Peer Effects, Social Learning, and Social Network Effects

As with many topics which claim interdisciplinary roots, the concept of social network effects as it relates to behavior can fall prey to semantic differences that may confuse many readers. Here, first the concepts of peer effects and social learning are defined which, in addition to being common topics in the economic literature, are the key foundation of social network models and effects. This is then followed by the definitions of the concept of social network models, which represent social ties through which peer effects and social learning occur across defined communities.

Peer Effects and Social Learning

Peer effects are commonly studied in economics and studies take a broad view of what constitutes peer influence. For example, one study examined decisions by school children in Kenya of whether to take drugs that kill intestinal worms already in the body. These drugs directly helped the individual who takes them, and generate positive externalities by breaking the transmission cycle, a pathway commonly referred to as the social multiplier. Researchers often try to make a distinction of whether the peer's behavior versus their peer's characteristics influenced one's decision because only the former pathway would lead to a social multiplier.

For policy purposes, a key issue is to understand the channels through which peer effects operate. In this example, do children take these drugs because of information sharing from communicating with those who took the drug earlier, social learning by observing how others' behavior and subsequent outcomes, reduced stigma or identity/image concerns may be lessened because others have taken the drug, or is it simply imitation? Although understanding the pathway the

social interactions operate is important as discussed in the next subsection, even identifying these effects is challenging.

Part of the challenge researchers must face is how to properly define an individual's peer group. In many papers researchers make ad hoc assumptions on the structure of peer groups. For example, only people who work in the same department within an organization or all individuals in a certain geographic area are considered. Researchers thus rely on the use of aggregates – such as the average characteristics or lagged behaviors of all classmates – to proxy for the social network. Such a setup is constraining: It means that individuals within a group must interact with each other, rather than with individuals outside the group. This can be a strong assumption if, for example, groups are formed by the researcher at the grade level, which suggests that students only interact with kids in the same grade but not with kids in different grades at the same school.

Social Network Models

If the concept of a peer group is not defined correctly, measurement error can be introduced by potentially omitting relevant peers, and when those neglected channels are not considered the underestimation of actual information flows can occur. As a result, more recent research is trying to directly model the formation of social networks and utilizing methods from graph theory to consider the microstructure of interaction among individuals within a community. As **Figure 1** outlines, while direct peer effects (panel (a)) and social norms shared throughout a community (panel (b)), each capture potential avenues of social learning and influence, there is a third way of conceptualizing this pathway, where the structure of individual social ties throughout a community can capture these community effects (panel (c)). Generally, social networks are defined as a set of actors and relationships (or ties) linking the actors. Networks can be egocentric in which case the network is built out from a subset of a population, or sociocentric, where a full population of individuals is included. Most social networks measured are egocentric due to the challenges related to data collection, though sociocentric networks are more empirically appealing to study as there are no assumptions required about missing data from unobserved individuals and social ties biasing results. Social network analysis can be used to study the structure of a social organization and how this structure influences the behavior of individual actors. As such, this kind of analysis extends the study of peer effects and social learning beyond a given actor's social ties.

Social Network Effects and the Challenge of Identification

Both intuition and previous work suggest that social influences and, by extension, social networks, are important when driving economic behavior. Assuming that these factors are significant, there remains the difficult question of assessing the magnitude of such an effect. There are three main identification challenges facing an empiricist. The first was termed the 'reflection problem' by Manski (1993) and is an issue that mimics a simultaneity problem. For example, in studying cigarette smoking a reflection problem arises when student and peer smoking are determined simultaneously, which inherently convolutes the measure of peers' influence.

The next challenge for the analyst is more complicated because individuals generally choose their friends in part based on the characteristics they favor. This second challenge is a form of selection bias and leads to a correlated unobservables problem. Social networks are not created in a vacuum, as the result of a random stochastic shock. Rather, they are formed based on the preferences of individual actors (nodes) in the vast majority of cases. Therefore, it is entirely conceivable that sorting into networks (also known as homophily) may occur based on traits and behaviors that are linked with and/or signals of future preferences. Thus, one usually

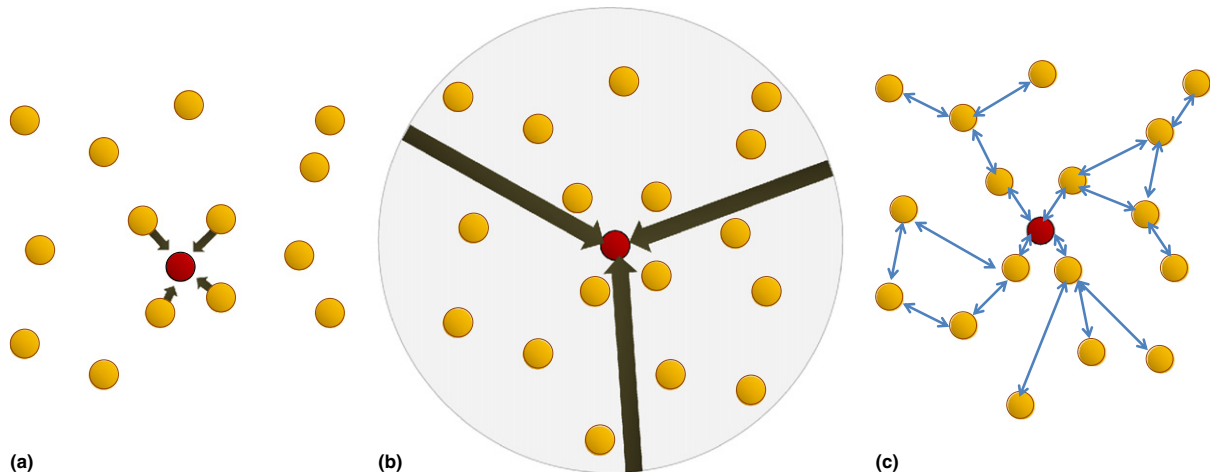


Figure 1 (a) Individual-level peer effects: Influence on an actor in a given community is exerted through direct social contacts. (b) Social norms: Influence on an actor is exerted only through the entire community, regardless of the relative strength of (or existence of a clear) social tie. (c) Social network effects: In social network models, the entire community can exert influence (even in the absence of a direct social tie between the actor and other community members), however, the relative influence of other members is measurable and related to the distance from the actor in Euclidean (or network) space.

does not know with certainty if an individual may be influenced to choose a specific diet or health plan because of the influence of their friend, or they chose their friend because of the friend's revealed preferences in the first place. In the latter case, it may be possible to observe what appear to be examples of social influence or contagion within a network, but are actually, primarily, artifacts of prior selection.

The third and final challenge involves the possible presence of unobserved group-level characteristics that affect both individuals and their peers. That is, some third confounding factor is responsible for the observed association between one's behavior and that of their peers. Taken together, if an individual is a member of some group, can the analyst distinguish a role for the characteristics and behaviors of others in the group in influencing that individual's choices?

More formally, much of the empirical literature on social economics has involved variations of a general linear model, dubbed by Manski the linear-in-means model.

$$Y_{igt} = \beta_0 + \beta_1 X_{igt} + \beta_2 Y_{-igt} + \beta_3 X_{-igt} + \beta_4 G_g + \varepsilon_{igt}$$

where Y_i is the health outcome under study for individual i who is a member of peer group g at time t, X_{igt} is a vector of individual exogenous controls, X_{-igt} is a vector of contextual controls common to all members of group g since the $-i$ notation indicates everyone but person i, Y_{-igt} is the mean peer choice in group g, G_g reflects common environmental influences affecting all members of group g, and ε_{igt} is a random error term with mean 0. β_1 expresses individual effects, β_3 contextual effects, β_4 correlated effects, and β_2 endogenous social effects. The linear-in-means model thereby provides a formal expression to three hypotheses often advanced to explain the common observation that individuals belonging to the same group tend to behave similarly.

The 'reflection problem' occurs if the peer variable measures peer group members at time t, which is obtained at the same time one's own health outcome is measured. It is called the reflection problem because it is similar to the problem of interpreting the almost simultaneous movements of a person and her reflection in a mirror. To overcome this challenge and identify the endogenous social effect, researchers generally ensure that all the regressors are known (predetermined) at the time of regression, which in theory avoids simultaneity problems. That is, the peer variable is constructed using earlier behaviors that were hopefully measured immediately before any interactions among the group g. Manski notes that if the transmission of peer effects really follows this temporal pattern, the identification problem is alleviated.

Empirical Approaches in the Estimation of Peer Effects in a Social Network

Given the challenges outlined earlier, how can a researcher infer true parameter estimates for social influence along networks? Although this is a thorny (some might argue intractable) challenge, there are in fact a number of experimental and analytical approaches that can be used to control for selection bias among other identification issues that economists are quite familiar with.

Overcoming Selection Bias

Researchers have attempted to overcome the selection bias in one of three main ways. Studies have used insights from randomized experiments to induce credible exogenous variation into aspects of social networks in an effort to identify their impacts. This research design can be seen in a number of papers, including the work of Duflo and Saez (2003) who explored not only the existence but also the mechanism underlying peer effects in the context of demand for benefits at the workplace. Specifically, they examined the role of social learning on the choice of employer-sponsored retirement plans, using individual data on employees of a large university a random sample of employees and focused on the question of whether people are influenced by the decisions of other employees in the same department. In a subset of departments some individuals were encouraged by an offer of a financial incentive to attend a benefits information fair organized by the employer. Not all departments were treated and the authors compared both benefits fair attendance and retirement plan enrollment decisions across departments and also looked within departments comparing outcomes of those who receive the treatment with their untreated coworkers. Receiving the letter led to a large increase in the likelihood of attendance and untreated individuals within departments where some individuals treated also had higher odds of attending the fair and plan enrollment 5 and 11 months after the fair. This presents convincing evidence that peer effects likely influence demand for benefit plan decisions.

Another method to identify the impacts of social factors on health outcomes is the use of instrumental variables to mitigate the correlation between unobservables and social network variables. This approach has been used in a large number of studies that have examined the role of peers on health behaviors such as cigarette smoking and obesity. However, these studies are frequently critiqued because the statistical properties and economic validity of these instruments are of debate. For example, some used their friends' birth weight as an instrument for whether their friend is currently obese in influencing whether one is overweight themselves. However, peer birth weight may influence other peer outcomes besides simply weight that may also directly affect one's own health outcomes.

Finally, several studies have attempted to use very rich data to control for unobserved confounders to identify the effects of social networks on health outcomes and show that those with greater levels of contacts with friends and neighbors have a reduced likelihood of enrolling in a Medicare-managed care plan relative to purchasing a medigap policy or having coverage through Medicare alone. Although the authors do account for a large set of unobserved confounders it remains possible that more sociable households are more risk tolerant or more optimistic than less sociable households, thus making these households more open to purchasing newer or 'riskier' insurance products, such as Medicare Health Maintenance Organizations (HMOs). This strategy is also used in international datasets. It has been found that social network effects are large and that both temporal and spatial proximity among household heads is the mechanism underlying this effect. However, one always can be concerned that those who lead

opinions in rural villages may also have certain characteristics favoring health plan adoption decisions.

Despite these limitations, it is worth noting that in certain contexts in the educational setting an explicit rule determines assignment to different peer groups. A sharp regression discontinuity design can be utilized when there is an explicit cutoff and individuals cannot change their behavior ex ante in an effort to sort to a specific side of the cutoff. This situation mimics a randomized experiment and one can simply compare individuals who just lie on either side of the cutoff. In an education setting, several countries share competitive admissions policies to secondary school leading to very different peer groups. For example, in China and Romania the secondary school system differs markedly from that of the USA which enabled researchers take advantage of the features and institutional structure of these systems. Specifically, students compete for positions in the higher ranked secondary schools by writing a high school entrance examination at the completion of junior middle school. Administrators at each senior high school grant admission to students whose exam performance is above a cutoff score. Thus, students who just get into a higher ranked school (perhaps by scoring only a point above the cutoff) have access to much stronger peers as measured by performance on the entrance examination, than students who scored just below the cutoff and now must attend a lower ranked school. One can imagine situations in healthcare settings where individuals are assigned to different treatment centers or nursing homes on the basis of health statistics creating an opportunity to examine how peers in different centers/homes affects one's health.

In addition to the examples listed so far, a number of other methodological approaches have been advanced in the statistical community so as to mitigate the strength of assumptions in social network models. One model is to assign all subjects into two separate groups randomly, and within them assess whether there appears to be a contagion effect from nonneighbors by using time series results in the first bin to predict results in the second. Nonzero results suggest social influence has traveled along the network over time. Another, simpler approach, relates to setting parameter bounds. In this approach, bounded parameters greater than zero would show that the entirety of the observed effect could not be due to selection (homophily) alone, though the magnitude of the effect may be significantly decreased.

Selected Application of Social Network Models to Health and Healthcare

A number of data sources such as the Add Health Study that collects survey data on self-reported friendship networks, and a growing number of social network data from electronic sources such as Facebook, LinkedIn, and others present researchers with new opportunities to study the impact of social learning and peer effects within networks on health behaviors over time. An earlier version of this approach is provided in Christakis and Fowler (2007) who used 32 years of data on 12 000 people from the Framingham Heart Study. This study garnered substantial attention from the popular press with a conclusion that obesity appears to spread through social ties. Fortunately, these spirited academic debates have led to a multitude of methodological improvements that should shed new and more convincing light on the role of social networks in the spread of obesity.

In contrast to risky health behaviors, there have been relatively few studies that have performed true social network analyses in the context of demand for healthcare and health insurance. There have, however, been a number that have looked broadly at peer effects as they relate to purchasing decisions in health or in areas (such as insurance) closely related to health. The most convincing evidence on the role of social networks relates to health plan decisions addressing the question of whether the information one receives from their peers affects their choices; even when product quality is difficult to ascertain. This information may come from direct communication with peers who have already purchased a particular health insurance bundle. Alternatively, it may arise from the observation of peers' purchasing decisions. This phenomenon as noted earlier is often referred to as social learning.

Whereas social learning has been extensively studied in theory, the empirical evidence is limited because social learning is difficult to identify in practice. Empirical analyses on the effects of social networks have difficultly providing direct evidence of causal relationships from consumption decisions and/or the product satisfaction of other members of one's social networks on individual decisions related to consumption of health products due to various conceptual and data problems including selection bias. Selection bias arises because people tend to associate with others based in part on some group characteristics they favor that are unobserved by the researcher. Thus, observing that individuals in the same group make similar consumption decisions may simply reflect shared preferences and not informational spillovers.

Exploring social effects in universities is a popular research design. Sorensen (2006) subsequently examined that health insurance selections are correlated across employees within the same department to multiple campuses in the University of California system. Nearly all full-time and some part-time employees are eligible to enroll in one of the health plans offered through the benefits program. He uses statistical models to examine whether individual- or department-level factors influence the decision to choose specific health plans. His empirical evidence provides convincing evidence that social effects (i.e., decisions of coworkers) play a role on individuals' choices of employer-sponsored health plans that is as large as many individual factors including age, income, and family status. The strength of the effect depends on factors such as the department's size or the employee's demographic distance from his coworkers. His research results have large policy implications because if all of one's coworkers in a specific department have chosen the same plan, then the social influence may overpower any individual incentives to switch plans when the provider raises the premium.

Another study uses data from a field experiment that changed the size of work-related social networks for those who were randomly assigned the intervention. These changes in the size of social networks are not due to choices by the individuals themselves and are free from selection bias, thereby providing the authors a unique opportunity to

estimate the causal impact of social networks on self-assessed measures of health. The effect of social networks on different health measures depends crucially on whether one holds a job. Those assigned to have a larger social networks report greater satisfaction with their mental and physical health when they are employed and the authors show that this effect is not due to the income channel.

In summary, the estimates (only a subset of which are discussed in the preceding paragraphs) indicate that social networks are an important determinant of the health insurance choices, health behaviors, and health as well. There is a large literature demonstrating positive associations between network size and mental health outcomes whereas another literature interestingly, finds that peers play small roles on actors in the medical system such as doctors in terms of their treatment choice and decisions to specialize. Future study of physician networks and social influence within them will benefit from new datasets being created by research teams.

Not only has there been studies investigating whether social influences exist but one may also wonder whether policies that aim to influence group dynamics subsequently shape individual health outcomes. Indeed social influences have been incorporated in many recent areas of health policy including efforts to reduce obesity. For example, weight-loss support groups have been in place for many years and use social influences in a manner similar to Alcoholics Anonymous to shape health behavior. In 2007, the academic journal *Obesity* devoted an entire issue to the evaluation of workplace interventions to reduce obesity. Last, several states including Arkansas have built policies around the mechanism of stigmatization as the channel of endogenous social effects, by providing students with weight report cards that provide information on their rank in the body mass index (BMI) distribution. Whereas each of these policies is built around specific social mechanisms that are hypothesized to alter obesity, their design does not appear to be based on a large body of evidence. Further, most seem to not have undergone a rigorous ex post policy evaluation of their effectiveness. Designing and evaluating policies that aim to take advantage of social multipliers are clearly areas for further research.

Final Comment

There is substantial evidence that individuals' beliefs, actions, and choices in the health sector are impacted by beliefs, actions, and choices of their peers. Insights from behavioral economics on framing and social influences are increasingly being used by the healthcare industry to influence individual choice in regards to specific products. Although substantial progress has been made on the econometrics of identification of social interactions, more careful work is needed to understand the mechanisms driving these social effects as well as whether there are moderators. The authors believe that this can be accomplished using field experiments and credible research designs. Further, in the education literature there is growing evidence that peer impacts can be distributed unevenly and more work in health economics is needed to understand the consequence of heterogeneity among peers. In the sciences, research conducted with animal populations demonstrates that social influences operating within the environment an individual engages in also directly affect gene expression, suggesting biological mechanisms underlying the social interaction effects. Findings from further empirical studies on the form and mechanisms of peer interactions are needed to guide further theoretical work. Naturally, the converse holds and despite a burgeoning literature over the past 15 years, an incomplete understanding of such effects remains. When studies are performed and results are interpreted with an appropriate amount of care and caution, one believes that there can be a great deal inferred from peer and social network models of influence to researchers in the health economics community and both health policymakers and the healthcare industry.

See also: Education and Health. Instrumental Variables: Informing Policy. Peer Effects in Health Behaviors

References

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **357**, 370–379.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Quarterly Journal of Economics* **118**(3), 815–842.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* **60**(3), 531–542.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology* **1**, 143–186.

Sorensen, A. T. (2006). Social learning and health plan choice. *RAND Journal of Economics* **37**(4), 929–945.

Further Reading

Beisetov, E., Kubik, J. and Moran, J. (2004). *Social interaction and health insurance choice of elderly: Evidence from health and retirement survey.* Mimeo: Syracuse University, Maxwell School for Citizenship and Public Affairs, Center for Policy Research, Working Paper.

Blume, L., Brock, W., Durlauf, S. and Ioannides, Y. (2010). Identification of social interactions. In Benhabib, J., Bisin, A. and Jackson, M. (eds.) *Handbook of social economics*, pp. 853–964. Amsterdam: Elsevier.

Ding, W. and Lehrer, S. F. (2007). Do peers affect student achievement in China's secondary schools? *Review of Economics and Statistics* **89**(2), 300–312.

Evans, W., Oates, W. and Schwab, R. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy* **100**(5), 84–117.

Fernald, R. D. and Maruska, K. P. (2012). Social information changes the brain. *Proceedings of the National Academy of Sciences of the USA* **109**(supplement 2), 17194–17199.

Ferschtman, C., Murphy, K. M. and Weiss, Y. (1996). Social status, education, and growth. *Journal of Political Economy* **104**, 108–132.

Jackson, M. O. (2008). *Social and economic networks*. Princeton, NJ: Princeton University Press.

Jackson, M. O. (2011). An overview of social networks and economic applications. In Benhabib, J., Bisin, A. and Jackson, M. O. (eds.) *The handbook of social economics*, vol. 1A, pp. 511–579. Amsterdam: Elsevier Science.

Manski, C. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives* **14**(3), 115–136.

McClellan, M., McNeil, B. J. and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**(11), 859–866.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**, 159–217.

Moffitt, R. (2001). Policy interventions, low-level equilibria, and social interactions. In Durlauf, S. and Young, P. (eds.) *Social dynamics*, pp. 45–82. Cambridge: Brookings Institution Press and MIT Press.
Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
Soetevent, A. (2006). Empirics of the identification of social interactions; An evaluation of the approaches and their results. *Journal of Economic Surveys* **20**(2), 193–228.
Trogdon, J., Nonnemaker, J. and Pais, J. (2008). Peer effects in adolescent overweight. *Journal of Health Economics* **27**(5), 1388–1399.
Veblen, T. B. (1934). *Essays on our changing order* (edited by Leon Ardzrooni), pp. xviii, 472. New York: The Viking Press.

Performance of Private Health Insurers in the Commercial Market

J Abraham and P Karaca-Mandic, University of Minnesota, Minneapolis, MN, USA

Introduction

Private health insurers play a large role in providing financial protection against the high cost of medical care in the United States. In 2010, approximately 64% of the overall US population had some form of private health insurance. The Centers for Medicare and Medicaid Services' National Health Expenditure Accounts (NHEA) project that aggregate private health insurance premiums across all market segments will reach $888.6 billion in 2012. This article examines the performance of private insurers operating in the fully insured, US employer group and individual markets. Although many private insurers cover Medicare and Medicaid beneficiaries, and serve as third-party administrators for self-insured employers, this article does not consider these aspects of production.

Later in this article, the authors present a conceptual framework to introduce and link the health insurance concepts of premiums, profits, administrative costs, loading fees, and medical loss ratios (MLRs). Following this, it summarizes the empirical evidence on insurers' performance with respect to premiums as well as loading fees and MLRs.

The types of private insurance demanded by individuals vary by their age and employment status as well as other factors. For individuals under the age of 65 years, the dominant form of coverage is employer-sponsored insurance (ESI). According to the Kaiser Family Foundation/HRET Survey, ESI premiums averaged $5 429 for single policies and $15 073 for family policies in 2011. ESI premiums have grown rapidly over the past decade and have contributed to the erosion of ESI, particularly among small employers. From the Medical Expenditure Panel Survey – Insurance Component, only 35.7% of private sector employers with fewer than 50 employees offered coverage in 2011, down from 47.2% in 2000.

Individuals who lack ESI may purchase coverage directly from an insurer. Nationally, the individual market is small relative to ESI, with an estimated 18.9 million (7%) nonelderly individuals reporting coverage purchased directly from an insurer based on the Annual Social and Economic Supplement to the 2010 Current Population Survey. Unlike ESI, persons demanding individual coverage often seek to bridge short-term coverage gaps, including the time period between two jobs, between school and a job, or between retirement and Medicare eligibility.

For the individual market, the average premium for single coverage was $2985, whereas it was $6328 for family coverage, based on a 2009 survey by America's Health Insurance Plans, the national association representing the health insurance industry. Although reported premiums are often lower for individual policies relative to ESI, they typically have lower actuarial values, defined as the percentage of total average costs for benefits that the plan will cover. Actuarial values are directly affected through a plan's cost-sharing provisions, including deductibles, coinsurance rates, and out-of-pocket maximum spending limits.

During the legislative debate over the Patient Protection and Affordable Care Act (ACA) of 2010, policymakers raised several concerns about the functioning of private health insurance markets and insurer behavior, particularly within the individual and small employer group markets. Central to this discussion was whether individuals and small businesses were getting poor value for their premiums because of insurers' high administrative costs and excessive profits. With passage of the ACA, dramatic changes to US health insurance markets are expected. This article provides important baseline information and evidence regarding the performance of US health insurers in the ESI and individual markets.

Conceptual Framework for the Production of Health Insurance

The premium charged by an insurer for a given level of coverage includes three basic components: expected claims, administrative expenses, and profit. Expected claims represent the amount of money that an insurer expects to pay hospitals, physicians, and other providers during the coverage period for the services incurred by a policyholder. Expected claims depend on both the negotiated prices between an insurer and providers as well as the types and quantities of medical services demanded. Expected claims are also related to the policy's benefit design, including the size of the deductible if there is one, the level of coinsurance and/or copayments, and out-of-pocket maximum limits as well as the quantity and types of medical services covered by the policy.

Insurers incur expenses for several different types of administrative functions. These functions may include general administration, information technology, product development and provider network management (e.g., contracting), sales (e.g., marketing, agent or broker commissions, underwriting, enrollment, and member services), medical management (e.g., utilization review or case management for high-cost enrollees), claims adjudication, and regulatory compliance. Insurers also incur expenses for premium taxes and fees. Finally, insurers are expected to incorporate some level of (normal) profit into the premium.

An insurer's production can be summarized more formally. In a differentiated goods industry, like the market for health insurance, with similar but not identical products, one can assume, for simplicity, each firm f of F firms sells one product. The profits π_f of firm f can be written as

$$\pi_f = (p_f - mc_f)Ms_f(p) - C_f \qquad [1]$$

where p_f is the price of firm f product, mc_f is the marginal cost of production, M is the exogenously determined size of the market, $s_f(p)$ is the share of firm f product, where p is the vector of all firms' prices, and C_f is the fixed cost of production. Average profits per policy can be expressed by dividing eqn [1]

 doi:10.1016/B978-0-12-375678-7.00911-1

with the term $Ms_f(p)$:

$$\text{avgprofit}_f = (p_f - mc_f) - \left(C_f / Ms_f(p)\right) \quad [2]$$

For health insurance products, p_f is the premium of the policy (prem_f) and mc_f includes expected claims paid out by the insurer (claims_f) and administrative expenses (admin_f) per policy such that

$$\text{avgprofit}_f = \text{prem}_f - \text{claims}_f - \text{admin}_f - (C_f / Ms_f(p)) \quad [3]$$

Assuming for simplicity that the fixed costs are sunk, such that $C_f = 0$, eqn [3] can be rewritten to obtain the standard health economics textbook expression for health insurance premiums:

$$\text{prem}_f = \text{claims}_f + \text{admin}_f + \text{avgprofit}_f \quad [4]$$

The loading fee, L_f, represents the portion of the total premium above and beyond the actuarially fair value or expected claims to be received from the policy during the coverage period. Typically, the loading fee is modeled as a multiplier to expected claims:

$$\text{prem}_f = (1 + L_f)\text{claims}_f \quad [5]$$

For example, if the premium is \$125 and expected claims total \$100, the loading fee is 0.25 or 25%.

A closely related concept to the loading fee is the MLR. Before the passage of federal health reform, the MLR has been defined as the ratio of expected claims paid by the insurer to the premium. Expressing the loading fee as a multiplier of expected claims, the MLR can be written as follows:

$$\text{MLR}_f = 1/(1 + L_f) \quad [6]$$

A closely related concept to the loading fee is the MLR.

Several factors influence the performance of health insurers. Insurers typically sell multiple products and each product is defined by a set of attributes. Thus, insurance products are often differentiated within and across firms. Common ways in which insurance products differ include their actuarial value and benefit design, the breadth and reputation of their provider network, and the level of customer service provided. These factors contribute to insurers' expected claims and loading fees. Although detailed information on product attributes may be easily observed for a single insurer, no comprehensive data source exists to facilitate analyses for the current population of US health insurers and their products.

Competition is another important factor affecting insurers' performance. Economic theory predicts that stronger insurance market competition among homogenous products should lead to lower premiums, *ceteris paribus*. Also, to the extent that purchasers of insurance have greater market power, this may lead to lower premiums. The structure of upstream markets is important too. Because insurers negotiate with local hospitals and physician practices over reimbursement rates, more competitive provider markets may improve insurers' leverage in negotiations and lead to lower input prices. Furthermore, for the individual and small employer market segments, brokers and agents play an important role in facilitating coverage purchases. Insurers' commission schedules with brokers and agents may be affected by the extent of competition among them as well.

Over time, insurers' premiums have followed a cyclic pattern. Called the underwriting cycle, this pattern reflects fluctuations in premiums and insurer profitability generated by decisions of firms to trade off profits for expanded market share. As noted by Grossman and Ginsburg (2004), several factors have contributed to the historical underwriting cycle, including the timing of forecasted cost trends relative to premium setting, the degree of insurance market competition, and the presence of not-for-profit insurers in the market.

Finally, the regulatory environment can affect insurers' performance. Following a failed attempt to pass federal healthcare reform in 1993–94, many states passed legislation to improve the functioning of the small employer and individual markets for insurance. Prevalent forms of regulation include guaranteed issue, guaranteed renewability, and premium rating limitations (e.g., rate bands), as well as mandated benefits (e.g., mammography screening). Differences in state regulations of insurers have led to variation in many insurance market outcomes, including premiums, administrative expenses, and coverage.

Empirical Evidence on Insurer Performance

Within the empirical literature on health insurance production, many studies have investigated factors that help to explain variation in premiums, although not necessarily focusing on specific components. A smaller body of research has focused on estimating the size of insurers' loading fees and/or MLRs. There is heightened awareness around the latter following implementation of federal minimum MLR regulation in 2011 as part of the ACA. Below, this article summarizes the empirical evidence on the factors that influence premiums as well as insurers' loading fees and/or MLRs for the employer group and individual markets.

What Factors Influence Premiums?

Market structure

Insurance market structure can be defined in a number of ways. Three of the most common measures include the total number of insurers operating in the relevant geographic and product market; a four-firm concentration ratio that provides the percentage of market share captured by the largest four firms in the market; and the Herfindahl–Hirschman Index (HHI), which is the sum of the squared market shares of the firms operating within the market, measured in percentage terms. The HHI has an upper bound of 10 000, corresponding to a monopoly.

Two studies have examined the influence of insurance market structure on premiums. Using a national sample of health maintenance organizations (HMOs) for the period 1988–91, Wholey *et al.* (1995) examined the relationship between premiums and the number of HMOs in the market. They found evidence that premiums were lower as the number of firms increased, providing support for advocates of managed competition. Also examining the effects of changes in

market structure, Dafny *et al.* (2012) investigated whether consolidation in the US health insurance industry, driven by a large merger, led to higher ESI premiums. They found that real premiums have grown by two percentage points in a typical market due to an average market-level change in HHI of 698 points.

Karaca-Mandic *et al.* (2013a) recently investigated the effects of competition in the market for insurance agents and brokers on ESI premiums for small employers. Using the Medical Expenditure Panel Survey – Insurance Component and data from the National Association of Health Underwriters, the authors provide evidence that premiums of policies offered by small employers are lower in markets with stronger competition among insurance agents and brokers.

Effects of regulation

Several actuarial and econometric studies have examined the effects on premiums of state-level insurance regulations, including benefit mandates (e.g., coverage for mammograms, *in vitro* fertilization, or mental health services) and rating regulations. Monheit and Rizzo (2007) summarize this evidence. Overall, there is mixed evidence with respect to the effects of benefit mandates on premiums. Some studies suggest these mandates are associated with modest premium increases, whereas others find no relationship. Work by Kowalski *et al.* (2008) focused on the relationship between state regulations implemented during the 1990s and premiums, including benefit mandates, guaranteed issue, community rating, guaranteed renewability, any-willing-provider laws, and individual market premiums. Using data from eHealthinsurance and Golden Rule, they found evidence that community rating regulations raised premiums and that the rate of increase was substantially higher if guaranteed issue regulations accompanied community rating regulations.

Other factors

Within the literature, a number of studies have documented variation in premiums by state, firm size, and plan type. Some studies have also examined premium variation over time within the context of the underwriting cycle and found evidence that insurers respond to a higher prior-year MLR (higher fraction of claims relative to premiums) by raising premiums in the next year (Born and Santerre, 2008). Using a proprietary data set of large employers, Dafny (2010) examine ESI premium setting and found that insurers engage in 'direct' price discrimination, charging higher premiums to firms with deeper pockets, as measured by operating profits. Finally, Karaca-Mandic *et al.* (2013a,b) use panel data for 2001–09 from the National Association of Insurance Commissioners (NAIC) to analyze the factors explaining variation in premiums per member month (PPMM) of coverage in the individual market. They found that insurers operating in other market segments, including the group and Medigap markets, had higher PPMM, whereas those operating in the Medicaid market had lower PPMM. They also documented differences in premiums by whether the insurer is local, regional, or national; whether the insurer is an HMO; and the size of the insurer as measured by the total number of member months of coverage.

What Is the Evidence on Loading Fees, Medical Loss Ratios, and Their Components?

The health insurance loading fee represents the portion of the premium above the expected claims paid by the insurance company. The loading fee includes general and claims-related administrative expenses, profits, broker commissions, and other sales-related expenses, and corresponds to the cost of transferring part of the risk-bearing from the individual to the insurer. According to the conventional theory of insurance, a risk-averse individual is willing to pay for a premium above the expected claims, and thus the size of the load reflects the value of a certain amount of income over an uncertain one with the same expected magnitude. Therefore, under the conventional theory, one can think of the loading fee as the 'relevant price' of the health insurance policy (Phelps, 2010), and therefore, a key factor when considering the value of coverage. In Nyman's theory of insurance, which does not rely on risk aversion to explain why individuals purchase insurance, premium is a more relevant price of insurance as it embodies the amount of non-health-related goods an individual gives up when healthy to receive an income transfer when sick (Nyman, 2003).

As a crude approximation to capture the extent of administrative costs and other fees associated with private insurance provision, the Centers for Medicare and Medicaid Services' Office of the Actuary estimates the net cost of private health insurance as the difference between the private health insurance premiums paid and benefits received in the NHEA. **Table 1** documents the net cost of private insurance as

Table 1 Private health insurance premiums paid and benefits received (in billions of dollars)

	1997	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*
Private health insurance premiums	362.50	388.10	419.90	459.60	503.00	560.50	614.50	658.90	702.90	740.20	776.20	807.60	828.80	848.70
Private health insurance benefits	322.60	347.40	373.70	407.10	445.80	488.50	527.50	566.90	607.70	640.60	673.50	707.50	734.00	746.00
Net cost	39.9	40.7	46.2	52.5	57.2	72.0	87.0	92.0	95.2	99.6	102.7	100.1	94.8	102.7
Net cost as percentage of benefits	12.4	11.7	12.4	12.9	12.8	14.7	16.5	16.2	15.7	15.5	15.2	14.1	12.9	13.8

Source: http://www.cms.hhs.gov/NationalHealthExpendData/downloads/tables.pdf (Table 15, accessed 27.08.12).

a percentage of private insurance benefits paid. As one can see, this percentage has remained relatively stable in the 12%–16.5% range during the 1997–2010 period. These statistics represent an average measure, without distinguishing among ESI (overall or across employer group sizes), individual market coverage, or supplemental Medicare insurance products.

The NAIC is the organization of insurance regulators from the 50 states, the District of Columbia, and the five US territories. It has one of the world's largest insurance industry databases, which are regularly used by industry leaders to determine market share, conduct market research, and monitor industry trends. The NAIC also publishes aggregate annual data separately by market segment (individual, group, supplemental Medicare, etc). **Table 2** presents data from the Statistical Compilation of Annual Statement Information for Health Insurance Companies published by the NAIC for the aggregate US group market. These statistics also do not distinguish by employer group size, but indicate that overall loading fees for the group market ranged between 11% and 20% for the 1997–2008 period. The MLR, defined as the percentage of premiums spent on clinical services, ranged between 83% and 90% during the same time period.

As noted in the section 'Introduction,' small employers are much less likely to offer health insurance relative to large employers. One explanation for this pattern is that employer groups face different loading fees. Thorpe (1992) posits several reasons why this might occur. Differences in transaction costs are one possible reason. Given the fixed costs of marketing and underwriting (in states where it is allowed), small employer groups are more expensive on a per-employee basis relative to larger ones. Furthermore, because small employers exhibit greater price sensitivity in their demand for insurance, they may be more likely to switch insurers as prices change, leading to additional expenses for marketing and underwriting over time. Third, given the voluntary nature of the market, insurers often express concern about the potential for adverse selection among small employers, and in turn, this may lead to higher risk premia.

The most commonly reported set of loading fee estimates by firm size dates back to two decades when an actuarial study was prepared by the Hay/Huggins Company for the US Congress' House Committee on Education and Labor in 1988. These estimates reflected the underwriting practices of major insurers, and suggested loading fees of approximately 40% for the very smallest firms (1–4 employees), 25% for those slightly larger (20–49 employees), and 18% for those with 50–99 employees. Hay/Huggins also reported that fees decline to 16% for those with 100–499 employees and to 12% for those up to 2500 employees. These estimates from the 1980s are still referenced frequently in the literature, including current health economics and health insurance texts.

In recent work, Karaca-Mandic *et al*. (2011) generated new estimates on the size of loading fees and how they differ across the firm size distribution using data from the confidential MEPS Household Component–Insurance Component Linked File. Overall, they found that firms of up to 100 employees face similar loading fees of approximately 34%. Loads decline with firm size and are estimated to be on average 15% for firms with between 101 and 10 000 employees and 4% for firms with more than 10 000 employees.

Focusing specifically on the individual market, Pauly and Nichols (2002) used NAIC data for the period 1988–99 and reported that expenses related to administration, sales, and risk-bearing represented between 30% and 40% of the premiums for individual market insurance. More recently, Abraham and Karaca-Mandic (2011) examined variation in MLRs among US health insurers in the individual market using NAIC data from 2002, 2005, and 2009. The authors documented large variation in MLRs by state, with enrollment-

Table 2 Health insurance industry aggregates, annual statement data from the NAIC for total US group market

Year	*Premiums earned (thousands)*	*Amount incurred for provision of health-care services (thousands)*	*Loading fee*	*Medical loss ratio (%)*
1997	44 559 067	39 972 761	0.11	90
1998	50 231 225	44 330 794	0.13	88
1999	114 735 686	101 529 936	0.13	88
2000	126 648 010	111 698 125	0.13	88
2001	111 019 585	96 970 954	0.14	87
2002	122 195 391	105 307 633	0.16	86
2003	131 529 713	113 158 119	0.16	86
2004	141 303 514	119 484 906	0.18	85
2005	157 094 448	131 132 072	0.2	83
2006	161 129 572	133 750 007	0.2	83
2007	169 768 926	145 427 131	0.17	86
2008	173 578 348	149 730 464	0.16	86
2009	174 888 283	152 544 111	0.15	87

Notes:

1. 1997, 1998, 1999, and 2000 represent only HMOs.
2. Medicare supplement, Federal Employees Health Benefit Plan, Title XVIII Medicare, and Title XIX Medicaid are not included in statistics reported.
3. Loading fee is calculated as (Premiums Earned/Claims Incurred) − 1.
4. Medical loss ratio is calculated as 100 (Claims Incurred/Premiums Earned).

Source: Reproduced from National Association of Insurance Commissioners (various years). Exhibit of premiums, enrollment and utilization of the *Statistical Compilation of Annual Statement Information for Health Insurance Companies*.

weighted average MLRs ranging from 0.629 in New Hampshire to values greater than 1 in Alabama, Massachusetts, Michigan, and North Dakota in 2009. Additionally, they estimated that 29% of insurer-state observations in the individual market would have MLRs (based on the historical definition) below the 80% minimum threshold imposed by the new ACA regulations, corresponding to 32% of individual market enrollment.

In 2011, a study by the General Accounting Office (GAO, 2011) analyzed insurers' MLRs in the individual and group markets using 2010 data and employing ACA MLR standards, which include adjustments for quality improvement expenses, federal and state taxes and licensing/regulatory fees, and life-years of enrollment. For the individual market, they found wide variation in MLRs, with only 43% of credible insurers and 48% of covered lives at or above the 2011 standard. For the small and group markets, these percentages were notably higher.

Although these studies provide valuable descriptive information, there is little empirical research to understand the factors that explain variation in insurers' loading fees or MLRs. A study by Karaca-Mandic *et al.* (2013b) investigates the determinants of MLR variation in the individual market over the 2001–09 time period. They evaluated how MLRs are influenced by changes in the composition of insurer and provider markets, the employer size distribution, and the demographic and health status of the population. Results suggest that insurance market structure is inversely related to MLRs. Insurers in markets in which they are the only credible insurer have lower MLRs, on average. This is consistent with such firms having higher market power. Here the classification of being a 'credible' insurer refers to having at least 1000 member-years of coverage as stipulated by federal regulations. Although the predicted average MLR is 77% for an insurer that is the only credible firm in the insurance market, it is 82% for an insurer with 2–4 other credible firms in the market, and 83% for an insurer with 5 or more credible firms in the market.

Conclusion

This article summarizes current evidence on US health insurers' performance in the ESI and individual markets, providing important baseline information about premiums as well as loading fees and MLRs. Dramatic changes to US health insurance markets and insurers' performance are expected as a result of the Patient Protection and ACA of 2010. In 2014, insurance exchanges will be implemented and will serve as organized marketplaces through which individuals and small employers can buy coverage. Exchanges will also be a primary mechanism through which coverage will be expanded to millions of uninsured, lower-income Americans who lack access to affordable ESI options. Finally, insurers are adjusting to a very different regulatory environment created by the ACA, including minimum MLR regulation and premium rate review enacted in 2011, as well as several major changes to benefit designs, the adoption of modified community rating, and the requirement that most individuals obtain health insurance in the United States beginning in 2014.

See also: Health-Insurer Market Power: Theory and Evidence. Managed Care. Private Insurance System Concerns

References

Abraham, J. and Karaca-Mandic, P. (2011). Regulating the medical loss ratio: Implications for the individual market. *American Journal of Managed Care* **17**, 211–224.

Born, P. and Santerre, R. (2008). Unraveling the health insurance underwriting cycle. *Journal of Insurance Regulation* 66–84. Spring.

Dafny, L. (2010). Are health insurance markets competitive? *American Economic Review* **100**(4), 1399–1431.

Dafny, L., Duggan, M. and Ramanarayanan, S. (2012). Paying a premium on your premium? Consolidation in the U.S. health insurance industry. *American Economic Review* **102**(2), 1161–1185.

GAO (2011). Private health insurance early experiences implementing new medical loss ratio requirements. Available at: http://www.gao.gov/new.items/d11711.pdf (accessed 26.07.13).

Grossman, J. and Ginsburg, P. (2004). As the health insurance underwriting cycle turns: What next? *Health Affairs* **23**(6), 91–102.

Karaca-Mandic, P., Abraham, J. and Phelps, C. (2011). How do health insurance loading fees vary by group size? Implications for healthcare reform. *International Journal of Health Care Finance and Economics* **11**(3), 181–207.

Karaca-Mandic, P., Abraham, J. and Simon, K. (2013b). Is the medical loss ratio a good proxy for market power in the individual market for health insurance? *University of Minnesota Working Paper.*

Karaca-Mandic, P., Feldman, R. and Graven, P. (2013a). The role of agents and brokers in the market for health insurance. *University of Minnesota Working Paper.*

Kowalski, A. E., Congdon, W. J. and Showalter, M. H. (2008). State health insurance regulations and the price of high-deductible policies. *Forum for Health Economics and Policy* **11**, 8.

Monheit, A. C. and Rizzo, J. (2007). Mandated health insurance benefits: A critical review of the literature. *Department of Human Services. Technical Report.* New Jersey. Available at: http://www.cshp.rutgers.edu/Downloads/7130.pdf (accessed 26.07.13).

Nyman, J. (2003). *The theory of demand for insurance.* Stanford, CA: Stanford University Press.

Pauly, M. and Nichols, L. M. (2002). The non-group health insurance market: Short on facts, long on opinions and policy disputes. *Health Affairs* Web Exclusive

Phelps, C. (2010). *Health economics.* Reading, MA: Addison-Wesley. Available at: http://content.healthaffairs.org/content/early/2002/10/23/hlthaff.w2.325.full.pdf (accessed 26.07.13).

Thorpe, K. E. (1992). Inside the black box of administrative costs. *Health Affairs* **11**(2), 41–55.

Wholey, D., Feldman, R. and Christianson, J. (1995). The effect of market structure on HMO premiums. *Journal of Health Economics* **14**, 81–105.

Relevant Websites

http://www.naic.org/
National Association of Insurance Commissioners.

http://www.cms.hhs.gov/NationalHealthExpendData/downloads/tables.pdf
National Health Expenditure Accounts.

http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/Proj2011PDF.pdf
National Health Expenditure Projections for the United States.

Personalized Medicine: Pricing and Reimbursement Policies as a Potential Barrier to Development and Adoption, Economics of

LP Garrison, University of Washington, Seattle, WA, USA
A Towse, Office of Health Economics, London, UK

Background: Barriers to Personalized Medicine

More than 10 years have passed since the completion of the first sequencing of the human genome in 2001. Since then, there has been continued and growing interest in the potential to use this new genetic information to better predict patient response to therapy. A variety of terms have been coined to describe this potential: personalized medicine (PM), stratified medicine, tailored medicine, and individualized therapy, among others. This article will adopt PM as the descriptor, more because of its currency and popularity than because of its more accuracy than other terms. PM has been defined many a times as "providing the right treatment to the right patient at the right time." Arguably, that is the aim of all medical therapy: the important difference in PM is the use of a new biomarker-based diagnostic test to further define and identify a subgroup of patients (called 'stratification') for whom the treatment performs better – in terms of either cost-effectiveness or benefit–risk balance. Incentives for the development and use of PM raise a number of interesting economic issues.

In the year 2000, Francis Collins, the current head of the US National Institutes of Health, said: "In the next five to seven years, we should identify the genetic susceptibility factors for virtually all common diseases – cancer, diabetes, heart disease, the major mental illnesses – on down that list." Clearly, this has not come to pass. In 2005, a more guarded assessment in the report 'Personalized Medicine: Hopes and Realities' from The Royal Society cautioned: "Pharmacogenetics is unlikely to revolutionize or personalize medical practice in the immediate future."

Before considering potential reasons for this lack of expected progress, it is useful to define some relevant biological, epidemiological, and clinical concepts and terms:

- Pharmacogenetics versus pharmacogenomics – the former is the study of how people's genetic makeup affects their response to medicines, whereas the latter is the application of genomic concepts to the development and clinical application of pharmaceuticals.
- Genotype versus phenotype – the former represents a person's genetic makeup, as reflected by his or her deoxyribonucleic acid (DNA) sequence, whereas the latter is an observable trait or characteristic of an organism (that may or may not be inherited).
- Germline versus somatic mutations – the former are heritable variations, whereas the latter are acquired (e.g., in cancer).
- Biomarkers – they are a broad array of biological indicators, including genetic variants, proteins, and endogenous metabolites, among others.
- Predictive versus prognostic biomarker – a prognostic biomarker is used to project patient health (e.g., life expectancy) based on patient characteristics, whereas a predictive biomarker predicts response to an intervention (e.g., a drug).

The slower-than-expected progress over the past decade could be for a number of reasons – scientific, regulatory, and economic. It does seem clear that the science is more difficult than first hoped. First, it is always important to remember that the science behind drug development is complex and uncertain. Almost 9 out of 10 new medicines under development fail between Phase 1 and Phase 3. Despite increasing amounts spent on drug development, industry productivity, as measured by approved new molecular entities, has been stagnant in recent years. The most recent estimate from the UK Office of Health Economics is that the average cost of drug development has grown to approximately $1.5 billion US$2011 per new compound. Clearly, the scientific unknowns and challenges are substantial.

Adding the parallel development of a predictive, biomarker-based test may reduce one scientific challenge but adds another that involves its own uncertainties. Furthermore, prediction based on genetic makeup is generally imprecise. Although a small number of single mutations (monogenic diseases) lead to specific health conditions (e.g., Huntington's disease), most complex diseases (such as diabetes) are affected by a large number of genes. Although an entire genome can be sequenced, it is only the beginning of understanding the biological function of most genes. Although some traits, such as height, are highly heritable, others are not: indeed, twin studies indicate that genes account for only approximately a quarter of the variation in lifespan – perhaps the ultimate measure of health. Clearly, gene–environment interaction is a very important influence. Regulation of drugs and devices may also be a barrier: the approval pathway for new pharmacogenomic tests has not been defined until recently and the standards of evidence for clinical utility for combination products (i.e., drug and test) are still being debated. Furthermore, there is not a level playing field between *in vitro* diagnostic tests (IVDs), which need regulatory approval for marketing, and laboratory-developed tests (LDTs) also called in-house tests (IHTs). The evidentiary requirements and quality controls for IVDs are much greater. In Europe, public health providers face a different regulatory regime that does not require each test to be approved. Yet, these different tests and providers compete in some PM applications.

But there are also some potential economic barriers that could increasingly be a factor given growing efforts to control medical spending in most developed countries. The incentive issue for existing marketed drugs is fairly obvious. Once a new, patented medicine is on the market with prices and reimbursement established around the world, the manufacturer has a very limited economic incentive to discover the subset of

 doi:10.1016/B978-0-12-375678-7.01210-4

patients in whom the drug works best if reimbursement will then be restricted to the subset. The problem is that reimbursement prices for drugs – especially outside the US – tend to be rigid and inflexible during the period of the patent. Thus, even if most of the aggregate health benefit – and hence economic value – is concentrated in a subset of patients, the manufacturer will receive or capture a much smaller share of that aggregate value if reimbursement is restricted to that subset but the reimbursed price does not rise to reflect their higher average health gain. From a longer term dynamic perspective, if price inflexibility holds and there is a strong possibility that the targeting test would be forthcoming, then drug manufacturers may have much less incentive to develop the drug in the first place. This has long-term, dynamic implications for the incentives for R&D investments in PMs. However, some targeting using tests might actually facilitate R&D and the demonstration of efficacy and safety, although this is difficult to predict *a priori* unless the mechanism of action of the drug is closely linked to a biomarker, for example, as was the case for imatinib (Gleevec®, Novartis) for chronic myelogenous leukemia (CML).

The rigidity of diagnostic reimbursement could be an even larger economic barrier – especially for companion diagnostics for drugs that are already on the market. In the US and most EU health-care systems, reimbursement for diagnostics is based on an administered pricing system linked imperfectly to the expected marginal cost of production and distribution, meaning that some diagnostics might garner profits, whereas others might just break even or even lose money (but persist for other business reasons). Thus, there is a limited incentive to incur the substantial fixed costs of evidence generation that would be necessary to demonstrate the clinical utility of the biomarker-based test in combination with the drug.

This article focuses on economic issues related to pricing and reimbursement policies as a potential barrier to the development and adoption of new, innovative PM technology. There are other important economic issues in PM, such as the impact on drug development costs, the relation to physician incentives to test and to follow test results, and issues related to targeting as a strategy for late entrants into a drug class. The next section summarizes the key theoretical issues. This is followed by a discussion of some key examples of PMs that are in use. The relevance of literature on the cost-effectiveness of PMs is discussed next. The article ends by identifying six major policy challenges. The general conclusion is that pricing and reimbursement systems will need to implement more flexible value-based rewards for PMs if the appropriate amount of R&D and evidence generation is going to be supported.

Economic Incentives for Personalized Medicine Development: A Framework

To this point, only 10–15 PM tests enjoy a significant volume of use, and the amount of evidence about their health and economic impact is limited. But many have been the subject of a cost-effectiveness analysis (CEA) (Wong *et al.*, 2010), and most are considered cost-effective by usual standards – or they would not be in use in health systems. For the purpose of this article, the more interesting and relevant work is the theoretical analysis of economic incentives and their policy implications for PM R&D.

This work has addressed two major questions. First, is it likely that current market structures for innovative PMs and their companion diagnostics will produce the optimal amount of PM development and use? Second, if the current situation is likely to be suboptimal, what could be done to improve it? It is important to recognize that these two questions include not only short-term questions about static efficiency but also long-term questions about dynamic efficiency that affect the sustainability of both PM and the patented medicines industry as a whole, especially because PM is often cited as the 'new paradigm' for drug development.

In an article published in 2002, Danzon and Towse developed a formal model of pharmacogenetic testing to examine the conditions under which the development of these targeted interventions was likely to be socially optimal. The model uses the common assumption of a societal willingness to pay for health gains (i.e., a threshold amount), as is often used in pharmacoeconomic CEA. However, they let the share of the gain captured by the drug manufacturer vary. The essential elements of their model are intuitively straightforward. Imagine that a diagnostic test can stratify a patient population that has previously received a drug into responders (R), those who benefit from the drug, and nonresponders (N), those who do not benefit. Before the test is available, both responders and nonresponders received the drug at price *P*. In Danzon and Towse's model, testing provides social value in several ways, including (1) avoiding drug spending on the nonresponders and (2) avoiding the costs and adverse health effects of adverse events in nonresponders. Of course, the value of the product among responders exists in either scenario. The key drivers of the social value of testing are the averted costs of adverse events times the share of N, and the cost of testing both R and N groups. Testing will generally be socially beneficial if the aggregate cost of testing is less than costs savings from not using the drug plus avoiding adverse events in N.

Their model identifies the key determinants of the incentives for drug manufacturers and payers to embrace pharmacogenetic testing. They also note implications for test developers and drug development more generally. Clearly, it will be critical for drug manufacturers to be able to capture a large share of the aggregate value created after the test is introduced. This generally means that the price paid for responders in the testing scenario must rise roughly in proportion to the rise in average patient benefit (due to eliminating adverse events and/or nonresponders), compared with the no-testing scenario, and also the price of test must be modest, relative to the cost of treatment. Although payers would generally prefer to pay less, in theory, they would be willing to accept a situation where the total amount paid out is the same if the net health benefits (i.e., among the responders) are the same as long as the costs of testing and the savings from avoiding adverse events in nonresponders are factored in. Danzon and Towse conclude: "The willingness of payers to award higher prices for targeted benefits ... will be essential to retaining neutrality in investment incentives" (p. 10).

Danzon and Towse (2002) also note that it is possible that the testing scenario could produce an eventual market size

that is so small (in terms of total revenues to the drug manufacturer) that it will not be sufficient to justify the costs of drug development if that cost is fixed (that is, independent of the size of the patient population). In practice, however, the US Food and Drug Administration (FDA) permits fewer and much smaller trials for orphan drugs. And government subsidies and other incentives may be needed in these situations for socially optimal investment. They also suggest that with free entry into the business of developing tests, manufacturers – with this new genetic knowledge – will have an incentive to incorporate testing into the drug development paradigm, which might also reduce the costs of that development. This is because in most markets (with the notable exception of the US), it is not possible to increase prices once a drug is launched because of rigid government price regulation. Thus, *ex post* targeting leads to lower volume but does not guarantee an increased price to reflect the greater health gain per patient in the smaller population. To date, although it is clear that manufacturers are increasingly considering testing as part of drug development the aggregate impact has not been large over the past 10 years.

Danzon and Towse assume the price and availability of the test is a given. They focus on the social benefits of health gains and of any reductions in cost. Considering many of these same issues in a less formal analysis, Garrison and Austin (2007) build on this analytical approach by analyzing the incentives for both the diagnostic company and the drug company for the codevelopment of companion diagnostic test (Dx) and a first-in-class drug therapy (Tx). They also explore an additional potential benefit of PM – the 'value of knowing.' They consider six scenarios that represent combinations of the following four factors:

1. whether Tx and Dx pricing reimbursement are value based or cost based, and whether they are flexible over time;
2. timing – whether Tx is already on the market, i.e., *ex post* versus *ex ante*;
3. whether intellectual property protection – to prevent copycats – is a barrier to entry;
4. the competitiveness of insurance market over short versus long term.

Although the overall result in terms of the importance of flexible drug pricing and reimbursement is the same, Garrison and Austin extend the model by adding some value creation for the 'value of knowing,' i.e., that the test–drug combination will be of higher value to the responders as they will know they will benefit. Assuming that they are risk averse, this reduction in uncertainty should give them greater peace of mind. This makes the total pool of value created larger: how this aggregate value is divided among patients, payer/insurers, drug manufacturers, and test developers is an important question. They also emphasize that although pricing and reimbursement for new drugs could be considered 'value-based' in the US and the key markets in the EU during the patent life, reimbursement for diagnostic tests tends to be a more rigid, administered pricing system, which could be called 'cost based.'

Their illustrative model considers a case where 20% of the users are responders and 80% are nonresponders. In the absence of a test, the value created in the 20% is essentially spread over the 100% (subject to adjustment for adverse events in the nonresponders) by setting the price based on the average benefit (including the nonresponders). In their base case scenario, with no Dx available, the (patented) Tx captures all of the value created (given the willingness-to-pay threshold). The five other scenarios explore variations on the four assumptions listed in this article while recognizing that the total value created is now greater due to the value of knowing. It does, then, become *ex post* a different zero-sum game, i.e., after the diagnostic becomes available – there is now a slightly bigger pie to divide up. The implications of the six scenarios are intuitive and straightforward.

If the price of the Tx is fixed, and the Dx enters the market *ex post*, the revenues going to the Tx manufacturer would fall dramatically. The Dx may capture this or not, depending on whether its pricing is regulated and at what level. The Tx manufacturer would suffer a severe revenue loss and has a very limited incentive to encourage Dx entry – unless the value could be recaptured by the Tx manufacturer by owning the Dx. Indeed, this is the circumstance for most drugs already on the market: neither Tx manufacturers nor Dx developers have a strong incentive to develop a test that identifies responders. Oncotype Dx for predicting breast cancer recurrence, which is discussed below, is one exception: the manufacturer avoided the cost-based reimbursement system in the US and charged approximately $3500 for the test. It is true that some panels of diagnostic tests might be lucrative under this system, but this would seem to be less likely true for the novel complex diagnostics needed for PM if their payment is based on the summation of the expected costs of the analytic steps as opposed to a measure of the incremental health gain.

If the Tx manufacturer's drug comes to market in combination with a companion test, then there are several different possibilities. If the Dx is subject to marginal cost-based reimbursement, then the manufacturer will want to capture as much value as possible through the drug price, which has much more flexibility at launch in most countries. If the Tx manufacturer also owns the Dx, then, in theory, the value capture could be split arbitrarily between the two. But, in practice, given administered pricing for the Dx and strong intellectual property protection for the drug, the incentive probably remains to capture as much value through the drug price as possible.

These scenarios illustrate several key points about the economic barriers that companion Dx–Tx products face. First and foremost is that inflexible pricing and reimbursement, which does not adjust to reflect value created, could undermine the rewards for developing PMs. Second, manufacturers would ideally enter the market with a combination product that has been clinically tested and validated *ex ante* as a combination. Flexible, value-based pricing and reimbursement would appear to be a necessary condition for encouraging optimal investments to produce more PM; of course, it cannot guarantee a large volume of PM development because scientific, regulatory, and drug development realities represent constraints.

Personalized Medicine Products

Although some have been disappointed by the slow progress in PM over the past 12 years, there has been a gradual accumulation of PM products, so that now between 10 and 20

Table 1 Companion diagnostic testing in PM

Technology	*Economic and testing features*
HER2 testing for breast cancer	A low-cost immunohistochemistry (IHC) (approx. $100–$200) test for human epidermal growth factor receptor 2 (HER2)-positivity was used in the initial clinical trial program and was provided by diagnostic companies. Subsequently, a higher cost and more accurate test ($300–$500) was developed (called 'FISH') and is in use. Approximately 80% of initial testing is done with IHC, with FISH retesting for patient with equivocal results. The drug manufacturer receives nearly all of the economic value created by the combination from the drug trastuzumab (Herceptin®, Roche)
BCR-ABL testing for chronic myelogenous leukemia (CML)	An example of an *ex ante* test (breakpoint cluster region-Abelson (BCR-ABL) gene) closely tied to the development of the drug: large majority of value capture by the drug imatinib (Gleevec®, Novartis). A second, BCR-ABL test is used to monitor for resistance and assignment to second-line therapies
Oncotype Dx® (Genomic Health) for breast cancer recurrence	An example of a relatively high-cost, value-capturing test aimed at avoiding unproductive chemotherapy
EGFR mutation testing in nonsmall-cell lung cancer (NSCLC)	An example where the stratifying mutation (epidermal growth factor receptor (EGFR)) was identified in trials that also included test-negative patients
HLA-B∗5701 allele testing for abacavir in HIV	Example of a test to identify patients who are more likely to suffer a severe adverse reaction to the HIV drug abacavir (Ziagen®, ViiV Healthcare)
KRAS testing in colorectal cancer	The KRAS mutation predicts which patients will not respond to two different monoclonal antibody treatments for colorectal cancer. The biomarker was identified after the products were on the market
PreDx® (Tethys Biosciences) diabetes risk test	This multimarker test identifies which prediabetic patients are at high risk of progressing to Type 2 diabetes: it indicates whether to begin prophylactic treatment with metformin
ALK mutation testing in NSCLC	Example of the drug crizotonib (Xalkori®, Pfizer) that targets a small subset (approximately 4%) of patients in disease condition with significant unmet medical need. It offers substantial survival gains in the subset, but with high testing cost per identified responder that must be factored in

notable PM products are available. Many of the PM products have been in oncology where it has been possible to link somatic mutations to chemotherapeutic response. **Table 1** presents some examples of PM products that illustrate the range of issues that arise in combining companion diagnostics with drugs to achieve PM. More details on three of these examples are provided in the next three paragraphs.

Trastuzumab (Herceptin®, Roche), a biological compound for the treatment of human epidermal growth factor receptor 2 (HER2)-positive breast cancer, has been called the first 'poster child' for PM. It was first approved in the US for the treatment of metastatic breast cancer (MBC) in 1998. Longer term (3-year follow-up) trials were initiated in early-stage breast cancer (EBC) and marketing approval was received in 2004. It provides a number of useful lessons as a case study in PM. First, these combination products can have a long gestation and life cycle. The potential of the HER mechanism was discovered in the early 1980s, but it took approximately 15 years to yield a viable compound combined with a predictive test. Second, the EBC indication, which was approved 6 years later than the MBC indication, produced much larger per-patient health gains and benefited many more women in the aggregate. Because the initial price of trastuzumab for treatment of MBC was set closer to the implicit willingness-to-pay threshold for cost-effectiveness (>US$100 000 per quality-adjusted life-year (QALY)) in the US, treatment in the EBC indication was much more cost-effective (<US$30 000 per QALY) (Garrison *et al.*, 2007). So what appeared to be high-cost medicine for its initial approved indication can be reasonably cost-effective over the entire product life cycle. The company might well have priced it higher if the favorable results in EBC would have been anticipated. Another complexity of the trastuzumab story is the persistence of basic issues about testing strategy. Two general types of tests are in use – a cheaper immunohistochemistry (IHC) test and a more expensive and more accurate fluorescence *in situ* hybridization test (FISH). IHC is the more common test and is used in 80% of all initial tests in the US. Even after all these years of experience, the optimal companion testing strategy is still under debate due to uncertainty about real-world test performance: approximately 80% of women with breast cancer receive the IHC test initially and some are retested with FISH.

Imatinib (Gleevec®, Novartis) is the unusual example of a PM combination product developed through rational drug design. It has been heralded as a virtual cure for many patients with CML. It has the distinction of having one of the fastest approval times by the FDA. The breakpoint cluster region-Abelson (BCR-ABL) enzyme that promotes cancer cell development appears only in cancer cells, can be identified by a test, and can be blocked. Once again, the large majority of the value created is captured by the drug, which has been estimated to be cost-effective in the US with a cost per QALY of approximately $50 000.

Oncotype Dx® for predicting breast cancer recurrence could be considered the poster child for a value-capturing PM test at approximately $3500 per test in the US. It is based on a '21-gene signature' that was constructed through retrospective analysis of historical tumor samples to generate a patented index to predict the likelihood of recurrence. The major economic benefit is that it avoids chemotherapy costs and side effects (including the risk of death) in women with EBC. The manufacturer was able to circumvent usual coding and pricing practices in obtaining the US Medicare reimbursement. The alternative would have been to use 'code stacking' of analytic steps, such as RNA extraction, reverse transcription, gene amplification, and interpretation and report. But this would have resulted in a payment level of only approximately US$540 (Gustavsen *et al.*, 2010).

Personalized Medicine, Cost-Effectiveness Analysis, and Pharmacoeconomics

In general, the methods of economic evaluations in PM are no different than standard CEA. In the usual case in pharmacoeconomics, the cost-effectiveness of a new medicine is assessed in a population with a particular disease (such as diabetes or rheumatoid arthritis) for the subset of patients who use the drug. A standard pharmaceutical CEA compares the impact of the use of a new medicine on health outcomes (usually measured in terms of QALYs gained) and on medical costs. The impact of PM further segments this subpopulation through the use of the biomarker-based test. The CEA, therefore, changes to the question of the economic impact of the use of the combination of the test and the treatment versus using the drug (or another treatment) without testing in the full population. The test is often called a companion diagnostic, and the pair has been called a 'codependent technology' (e.g., by regulators in Australia). The overall standard analysis in the field of pharmacoeconomics remains 'cost-utility analysis' with the primary metric being the incremental cost-effectiveness ratio (ICER) in terms of cost per QALY gained, although it is not used by payers in all jurisdictions, for example, Germany.

Given the cost of the test and the cost of the treatment, the usual principles of CEA apply, except that the cost of the testing and outcomes for all test recipients must be included in the analysis. It has been argued that the QALY may not capture the utility gain that patients may receive from the value of knowing, i.e., the reduction in uncertainty in terms of whether the medicine will work well for them. From a modeling approach, this reduction amounts to less uncertainty about the value of the drug (assuming the test has high accuracy). If this were to be included, some add-on or adjustment to the QALY gain would be needed: there are various preference elicitation techniques in economics that can be used to elicit willingness to pay for such product attributes.

Besides creating value by reducing uncertainty, a targeted Dx–Tx combination could create further value in a population beyond what is captured in the QALY and ICER in, at least, two other ways. First, suppose that some responders were noncompliant because they were not sure whether or not they were responding. A companion Dx could increase their confidence about response and hence compliance with Tx, producing better outcomes in those patients, and thus creating more value at the aggregate level. Second, in the situation where the Dx is not available, some patients who would be responders might not actually choose to go to the doctor for the broader diagnosis. With an available Dx, however, they might seek this care. Thus, overall uptake would increase. Measuring this aggregate value gain would require a somewhat more sophisticated model than is typically used in CEA for a typical patient: a population-level uptake and use model would be needed, more akin to a budget impact model. But it is feasible to model and estimate these additional sources of value.

Regulatory and Policy Issues and Implications

There are several major scientific challenges in the development of PM, including the low probability that a new mechanism of action will work, the challenge of parallel development of a new companion diagnostic, and the uncertainty of genetic prediction of complex traits. Clearly, the scientific unknowns and challenges are substantial. Nonetheless, it is important to ask what can be done to optimize the economic incentives. The following six potential challenges need to be researched, and, if appropriate, policy changes made.

Flexible Value-Based Pricing for Drugs

Economic theory would suggest that payers and pricing and reimbursement authorities should allow for flexible pricing (both up and down) for drugs at launch and postlaunch if the evidence suggests that they can be targeted in a narrower patient group or used in a number of different indications or subgroups of different value. The UK 2009 Pharmaceutical Price Regulation Scheme, for example, included (1) provisions for 'flexible pricing' that allowed drug developers to seek approval from the National Institute for Health and Clinical Excellence for higher prices when evidence of value increased and (2) provisions for new indications to be launched at different prices (higher or lower) than existing indications. However, neither of these provisions has been used to date in the UK. This may reflect uncertainty on the part of companies as to how the policy would be applied or the linkage of the UK market via parallel trade and reference pricing to other EU markets where such provisions do not apply.

Regulatory Flexibility in Drug–Test Combination Development

Given the evidence on drug development failure rates, pharmaceutical companies can be expected to be resistant to increases in development costs caused by adding a test into the development program (notwithstanding the potential advantages of *ex ante* vs. *ex post* stratification for the ability of the pharmaceutical company to extract rents). In the case of the drug–test companion development, one might, therefore, expect that it is important initially that the delivery of a prototype assay for use in Phase 3 development does not call for significant investment in advance of being in position to recognize the efficacy or otherwise of the drug itself in Phase 2. To achieve this would require flexibility in the regulatory hurdles for tests as part of drug development. (And this links to the importance for achieving socially optimal outcomes of having an environment in which better quality follow-on tests are incentivized to enter the market.) This flexibility in the drug development process may require the payer's pricing and reimbursement arrangements for drug–test combinations to concentrate on the evidence from the Phase 3 drug development trial and not require a patient randomization to use the test (e.g., the 'double randomized trial' proposed in the Australian government's guidelines for codependent technologies.) As the model changes over time, drug developers may not have to codevelop new tests and biomarkers but can draw on existing ones. In this context, the key issue will not be for the drug developer: instead, it will be necessary to ensure that diagnostic developers have an incentive to bring improved

tests to the market and will obtain value-based remuneration for them.

Flexible Value-Based Reimbursement for Diagnostics

Testing that enables better choice of treatment or improved disease management may generate downstream health effects of extended life and/or improved quality of life. In addition, a diagnostic test can enhance the level of information about a specific clinical condition or health state, and so reduce or eliminate uncertainty. This may have value independent of any ability to improve treatment choice – although insurers' willingness to pay for this may be unclear. Historically, pricing and reimbursement systems for diagnostics have been focused on the expected cost of making and conducting the test (which may depend on the technology platform used) and not the value delivered. This has meant that the price of a new diagnostic is often based on the price of existing tests ('cross-walking') with similar clinical use or with similar characteristics or on production cost based on analytic steps. For example, in the US, a number of diagnostics are reimbursed via code stacking by reporting a combination of reimbursement Current Procedural Terminology codes describing laboratory protocol stages. Theory would suggest that both diagnostic-dedicated and drug health technology assessment (HTA) processes should use a common, consistent, and comprehensive approach to assessing value to enable the development of pricing and reimbursement arrangements capturing the full incremental value offered by those technologies. Policymakers also need to consider the incentive for companies to invest in evidence collection to raise the standard of clinical utility data available to support the case for using a companion test. Such incentives might come from rewarding competing tests that supply evidence that they bring improved health gain and information through greater accuracy.

Dividing the Combination Value between the Drug and the Diagnostic

Accepting for now the notion that flexible value-based pricing is desirable for both an innovative PM and an innovative companion diagnostic raises the question, how can the problem of synergistic/codependent technologies noted above be best resolved? As a starting point, it may be easier to start with the case of the *ex post*, 'standalone' test. Recall that the average development cost of a new drug is $1.5 billion US$2011, and the productivity of drug R&D has been declining over time. In contrast, the cost to develop a new biomarker-based assay is in the order of $10 million to $50 million – at the high end if additional clinical trials are needed to validate the biomarker-based test. Consider a hypothetical case where an existing biopharmaceutical is on the market and earning $1 billion in annual revenues despite having only a 50% response rate. Suppose a reliable biomarker-based test is invented that can predict the responders. In theory, the payer would be indifferent to giving them $500 million and perhaps a premium for the value of knowing (i.e., reduced uncertainty for both the payer and the patient). Clearly, if the drug manufacturer thought in advance that this might happen, the prospect of lower revenues would have reduced the likelihood of them making the initial investment. One could argue, however, that it was the drug manufacturer's invention that created the health gain in the responders. Then, the extra value created by the Dx manufacturer consists of (1) avoiding any adverse event treatment costs and any related health losses in the nonresponders plus (2) the 'value-of-knowing premium.' (Note that if the premium were, say, 5% of total health gain, it would be worth $50 million: an amount that could support substantial evidence generation.) With flexible value-based pricing, this logic would argue that the drug innovator's price would be roughly doubled (i.e., to $1 billion in revenues) and the Dx innovator would receive a reward in relation to cost savings and QALY gains in the nonresponders plus a premium for the value of knowing (i.e., $50 million). This split is arbitrary in a static sense, but it can be argued that it reinforces dynamic efficiency by considering the relative size of the investments needed to develop a drug versus a diagnostic. Clearly, this issue deserves further research.

'Follower' Diagnostic Tests 'Piggyback' on Clinical Utility Evidence

Evidence on drug effectiveness and value is generated by the drug developer in order to obtain regulatory and pricing approval. The drug company has a patent (no one can copy the product) and also regulatory exclusivity on the data they generate (no one can reference the data to get regulatory approval by claiming that their product does the same). The situation is more complicated in diagnostics: patents may be less robust, regulatory hurdles are lower, and so it is more difficult to prevent others appropriating the benefits of evidence from studies the diagnostic manufacturer has paid for. As a consequence, there may be underinvestment in evidence generation. Data exclusivity could be given to evidence for diagnostics. This would require 'follow-on' tests to replicate the evidence generated by the 'first-in-class' test (as is the case for drugs). However, under current diagnostic regulations in the US and EU countries, LDTs (and in the EU any tests provided by public health providers) have to meet lower regulatory hurdles and so could not be prevented from piggybacking on clinical utility evidence. In addition to the data exclusivity requirement, there may, therefore, need to be an expectation that payers will pay more for the test with the stronger evidence base (i.e., that they will ignore tests without an evidence base when making HTA). Of course, a balance has to be struck, as with patenting. The objective is not to delay competition to provide innovative tests but to ensure initial providers have the potential to earn a return if they have evidence to support claims of value. Research is clearly needed on both effective incentives to prevent piggybacking for a least a period of time and how long they should be put in place.

Flexibility in Diagnostic Test Evidence for Reimbursement

Given the US and EU regulatory environments, a pragmatic approach could be taken to collecting evidence on the clinical utility of diagnostics, for example, using small randomized

studies (if not built into Phase 3 of drug development), which lead to conditional reimbursement approval and evidence-based reimbursement rate to incentivize manufacturers plus real-world, postlaunch data collection. To facilitate real-world data collection, increased investment in national e-health records would, certainly, help, although the overall investment case for such records is much greater than simply facilitating assessment of the benefits (or otherwise) of developing and validating diagnostic tests.

Conclusion

Although there is a general perception that the growth of PM has been slower than hoped, it is not clear how much of this perceived shortfall is due to scientific and regulatory constraints versus economic incentives. Many of the PM products have been in oncology, where it has been possible to link somatic mutations to drug response. The science is more difficult outside of oncology. There are, clearly, problems with economic incentives in this area as well. However, the empirical evidence on this is limited by the challenge of constructing the counterfactual. These may reflect the scientific challenge: not enough drug-diagnostic combinations have received regulatory approval to allow exploration of the issues around commercial success. However, the relatively small number of regulatory approvals may also reflect perceived commercial limitations as well as scientific challenges.

Over the past dozen years, the thinking about PM predictive tests has evolved from thinking primarily about using genetic mutations or combinations as predictors of clinical response to a wider range of 'biomarkers,' and even to changing the name of the whole area to 'molecular diagnostics.' Owing to the rapidly falling cost of sequencing an individual's entire genome (e.g., projected to be less than $1000 in the near future), a new set of economic issues are arising around the question of how to use this plethora of data from whole genome scans. These new economic issues have not been addressed here but will need to be better understood in future discussions of the economics of PM. In the extreme, the emerging pharmacogenomics diagnostic industry will change drastically, making it possible for a whole genome test to be performed at birth on an individual. Still, testing for somatic mutations would be needed and incentives would be required to invest in providing evidence that a particular set of biomarkers (including genes) is predictive of disease.

See also: Adoption of New Technologies, Using Economic Evaluation. Biopharmaceutical and Medical Equipment Industries, Economics of. Cost-Effectiveness Modeling Using Health State Utility Values. Economic Evaluation, Uncertainty in. Information Analysis, Value of. Patents and Other Incentives for Pharmaceutical Innovation. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Research and Development Costs and Productivity in Biopharmaceuticals. Value of Information Methods to Prioritize Research

References

Danzon, P. and Towse, A. (2002). The economics of gene therapy and of pharmacogenetics. *Value Health* **5**(1), 5–13.

Garrison, L. P. and Austin, M. J. F. (2007). The economics of personalized medicine: A model of incentives for value creation and capture. *Drug Information Journal* **41**, 501–509.

Garrison, L. P., Veenstra, D. L., Carlson, R. J., Carlson, J. J. and Meckley, L. M. (2007). Backgrounder on pharmacogenomics for the pharmaceutical and biotechnology industries: Basic science, future scenarios, policy directions. *Report of the Pharmaceutical Outcomes Research Policy and Program.* Department of Pharmacy, University of Washington. Available at: http://sop.washington.edu/porpp/general/reports.html (accessed 08.10.13).

Gustavsen, G., Phillips, K. and Pothier, K. (2010). *The reimbursement landscape for novel diagnostics*. Weston, MA: Health Advances. Available at: http://www.healthadvances.com/pdf/novel_diag_reimbursement.pdf (accessed 08.10.13).

Wong, W. B., Carlson, J. J., Thariani, R. and Veenstra, D. L. (2010). Cost effectiveness of pharmacogenomics: A critical and systematic review. *Pharmacoeconomics* **28**(11), 1001–1013.

Further Reading

Blair, E. D., Stratton, E. K. and Kaufmann, M. (2012). The economic value of companion diagnostics and stratified medicines. *Expert Review of Molecular Diagnostics* **12**(8), 791–794.

Davis, J. C., Furstenthal, L., Desai, A. A., et al. (2009). The microeconomics of personalized medicine: Today's challenge and tomorrow's promise. *Nature Reviews Drug Discovery* **8**(4), 279–286.

Faulkner, E., Annemans, L., Garrison, L., et al. Personalized Medicine Development and Reimbursement Working Group (2012). Challenges in the development and reimbursement of personalized medicine – payer and manufacturer perspectives and implications for health economics and outcomes research: A report of the ISPOR Personalized Medicine Special Interest Group. *Value Health* **15**(8), 1162–1171.

Garrison, L. P. and Austin, M. J. F. (2006). Linking pharmacogenetics-based diagnostics and pharmaceuticals for personalized medicine: Scientific and economic challenges. *Health Affairs* **25**(5), 1281–1290.

Keeling, P., Roth, M. and Zietlow, T. (2012). The economics of personalized medicine: Commercialization as a driver of return on investment. *New Biotechnology* **29**(6), 720–731.

Meckley, L. M. and Neumann, P. J. (2010). Personalized medicine: Factors influencing reimbursement. *Health Policy* **94**(2), 91–100.

Ramsey, S. D., Veenstra, D. L., Garrison, L. P., et al. (2006). Toward evidence-based assessment of laboratory-based diagnostics and genetic test for health plan reimbursement. *American Journal of Managed Care* **12**, 197–202.

Trusheim, M. R., Berndt, E. R. and Douglas, F. L. (2007). Stratified medicine: Strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews Drug Discovery* **6**(4), 287–293.

Veenstra, D. L. (2009). The economic value of innovative treatments over the product life cycle: The case of targeted trastuzumab chemotherapy for breast cancer. *Value in Health* **12**(8), 1118–1123.

ENCYCLOPEDIA OF
HEALTH ECONOMICS

ENCYCLOPEDIA OF
HEALTH ECONOMICS

EDITOR-IN-CHIEF
Anthony J Culyer
University of Toronto, Toronto, Canada
University of York, Heslington, York, UK

ELSEVIER

Amsterdam • Boston • Heildelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

First edition 2014

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the Library of Congress.

ISBN 978-0-12-375678-7

Printed and bound in the United States of America

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

Project Manager: Gemma Taft
Associate Project Manager: Joanne Williams

EDITORIAL BOARD

CONTRIBUTORS TO VOLUME 3

AJ Atherly
University of Colorado Anschutz Medical Campus, Aurora, CO, USA

R Baker
Glasgow Caledonian University, Glasgow, UK

ER Berndt
Massachusetts Institute of Technology, Cambridge, MA, USA

J-R Borrell
University of Barcelona, Barcelona, Spain

F Breyer
University of Konstanz, Konstanz, Germany

AH Briggs
University of Glasgow, Glasgow, Scotland, UK

C Cassó
University of Barcelona, Barcelona, Spain, and Northeastern University, Boston, MA, USA

ME Chernew
Harvard Medical School, Boston, MA, USA

JP Cohen
University of Hartford, West Hartford, CT, USA

R Conti
University of Chicago, Chicago, IL, USA

R Cookson
Centre for Health Economics, University of York, York, UK

PM Danzon
University of Pennsylvania, Philadelphia, PA, USA

DM Dave
Bentley University, Waltham, MA, USA

G Dionne
HEC Montréal, Montreal, QC, Canada

C Donaldson
Glasgow Caledonian University, Glasgow, UK

C Donaldson
Glasgow Caledonian University, Glasgow, Scotland

JA Doshi
University of Pennsylvania, Philadelphia, PA, USA

P Dupas
Stanford University, Stanford, CA, USA

RP Ellis
Boston University, Boston, MA, USA

H Fang
University of Colorado Denver, Denver, CO, USA

R Faria
University of York, York, UK

AM Fendrick
Harvard Medical School, Boston, MA, USA, and University of Michigan, Ann Arbor, MI, USA

M Fleurbaey
Princeton University, Princeton, NJ, USA

J Glazer
Boston University, Boston, MA, USA, and Tel Aviv University, Tel Aviv, Israel

P González
Universidad Pablo de Olavide, Sevilla, Spain

J Graff Zivin
University of California, San Diego, La Jolla, CA, USA

G Gupte
Boston University, Boston, MA, USA

H Haji Ali Afzali
The University of Adelaide, Adelaide, SA, Australia

K Hauck
Centre for Health Policy, Imperial College Business School, London, UK

N Hawkins
Icon PLC, Dublin, Ireland and University of Glasgow, Glasgow, Scotland

J Hurley
McMaster University, Hamilton, ON, Canada

I Jelovac
University of Lyon, Lyon, France, and CNRS, Ecully, France

EM Johnson
Massachusetts Institute of Technology, Cambridge, MA, USA

M Jones-Lee
Newcastle University Business School, Newcastle upon Tyne, UK

B Kachniarz
Harvard Medical School, Boston, MA, USA

P Kanavos
London School of Economics, London, UK

J Karnon
The University of Adelaide, Adelaide, SA, Australia

LA Karoly
RAND Corporation, Arlington, VA, USA

JT Kolstad
University of Pennsylvania, Philadelphia, PA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

J Koola
Harvard University Kennedy School of Government

K Lamiraud
ESSEC Business School, Cergy, France

K Lawson
University of Glasgow, Glasgow, Scotland

TJ Layton
Boston University, Boston, MA, USA

PT Léger
HEC Montréal, Montreal, QC, Canada

M Lewis
Georgetown University, Washington, DC, USA

X Martinez-Giralt
Universitat Autònoma de Barcelona and MOVE, Barcelona, Spain

H Mason
Glasgow Caledonian University, Glasgow, Scotland

C McCabe
University of Alberta, Edmonton, AB, Canada

S McElligott
University of Pennsylvania, Philadelphia, PA, USA

TG McGuire
Harvard Medical School, Boston, MA, USA

E McIntosh
University of Glasgow, Glasgow, Scotland

D Meltzer
University of Chicago, Chicago, IL, USA

PT Menzel
Pacific Lutheran University, Tacoma, WA, USA

A Mills
London School of Hygiene and Tropical Medicine, London, UK

MA Morrisey
University of Alabama at Birmingham, Birmingham, AL, USA

F Moscone
Brunel University, Uxbridge, UK

M Neidell
Columbia University, New York, NY, USA

E Nord
Norwegian Institute of Public Health, Oslo, Norway

MK Olson
Tulane University, New Orleans, LA, USA

S Paisley
University of Sheffield, Sheffield, South Yorkshire, UK

A Panjamapirom
The Advisory Board Company, Washington, DC, USA

I Papanicolas
London School of Economics, London, UK

M Paulden
University of Toronto, Toronto, ON, Canada

JA Rizzo
Stony Brook University, Stony Brook, NY, USA

CG Rothschild
Wellesley College, Wellesley, MA, USA

JA Salomon
Harvard School of Public Health, Boston, MA, USA

G Scally
University of the West of England, Bristol, UK

FM Scherer
Harvard University, Cambridge, MA, USA

E Schokkaert
KU Leuven, Leuven, Belgium

M Shah
University of California, Los Angeles, CA, USA

SP Shah
John Hopkins University School of Medicine, Baltimore, MD, USA

I Shemilt
Campbell and Cochrane Economic Methods Group, and University of Cambridge, Cambridge, UK

L Siciliani
University of York, York, UK

L Siciliani
University of York, York, UK, and Centre for Economic Policy Research, London, UK

K Simon
Indiana University, Bloomington, IN, USA, and National Bureau of Economic Research, Cambridge, MA, USA

AD Sinaiko
Harvard School of Public Health, Boston, MA, USA

FA Sloan
Duke University, Durham, NC, USA

L Smith
University of Michigan, Ann Arbor, MI, USA

PC Smith
Imperial College, London, UK

M Stabile
University of Toronto, Toronto, ON, Canada, and National Bureau of Economic Research, Cambridge, MA, USA

T Stargardt
University of Hamburg, Hamburg, Germany

E Strumpf
McGill University, Montreal, QC, Canada

M Suhrcke
Norwich Medical School, University of East Anglia, Norwich, UK, and UKCRC Centre for Diet and Activity Research (CEDAR), Cambridge, UK

M Tai-Seale
Palo Alto Medical Foundation Research Institute, Palo Alto, CA, USA

P Tappenden
University of Sheffield, Sheffield, UK

JV Terza
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

JL Tobias
Purdue University, West Lafayette, IN, USA

E Tosetti
Brunel University, Uxbridge, UK

M Townsend
University of Toronto, Toronto, ON, Canada

A Towse
Office of Health Economics, London, UK

L Vale
Campbell and Cochrane Economic Methods Group, and Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK

B Van den Berg
University of York, York, UK

S Vandoros
London School of Economics, London, UK

WPMM van de Ven
Erasmus University Rotterdam, Rotterdam, The Netherlands

H Weatherly
University of York, York, UK

W Whittaker
University of Manchester, Manchester, UK

RL Williams
RTI International, Raleigh, NC, USA

E Wilson
Campbell and Cochrane Economic Methods Group, and University of East Anglia Norwich Research Park, Norwich, Norfolk, UK

O Wouters
London School of Economics, London, UK

DJ Wright
University of Sydney, Sydney, NSW, Australia

P Yadav
University of Michigan, Ann Arbor, MI, USA

AP Zwane
Bill & Melinda Gates Foundation

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you're interested in: an alphabetical contents list, cross-references to other relevant articles within each article, and a full subject index.

1 Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

2 Cross-References

Most of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an entry as a See also list, serve four different functions:

i. To draw the reader's attention to related material in other entries.
ii. To indicate material that broadens and extends the scope of the article.
iii. To indicate material that covers a topic in more depth.
iv. To direct readers to other articles by the same author(s).

Example

The following list of cross-references appears at the end of the entry Abortion.

> *See also*: Education and Health in Developing Economies. Fertility and Population in Developing Countries. Global Public Goods and Health. Infectious Disease Externalities. Nutrition, Health, and Economic Performance. Water Supply and Sanitation

3 Index

The index includes page numbers for quick reference to the information you're looking for. The index entries differentiate between references to a whole entry, a part of an entry, and a table or figure.

4 Contributors

At the start of each volume there is list of the authors who contributed to that volume.

SUBJECT CLASSIFICATION

Demand for Health and Health Care

Collective Purchasing of Health Care
Demand Cross Elasticities and 'Offset Effects'
Demand for Insurance That Nudges Demand
Education and Health: Disentangling Causal Relationships from Associations
Health Care Demand, Empirical Determinants of
Medical Decision Making and Demand
Peer Effects, Social Networks, and Healthcare Demand
Physician-Induced Demand
Physician Management of Demand at the Point of Care
Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment
Quality Reporting and Demand
Rationing of Demand

Determinants of Health and Ill-Health

Abortion
Addiction
Advertising as a Determinant of Health in the USA
Aging: Health at Advanced Ages
Alcohol
Education and Health
Illegal Drug Use, Health Effects of
Intergenerational Effects on Health – *In Utero* and Early Life
Macroeconomy and Health
Mental Health, Determinants of
Nutrition, Economics of
Peer Effects in Health Behaviors
Pollution and Health
Sex Work and Risky Sex in Developing Countries
Smoking, Economics of

Economic Evaluation

Adoption of New Technologies, Using Economic Evaluation
Analysing Heterogeneity to Support Decision Making
Budget-Impact Analysis
Cost-Effectiveness Modeling Using Health State Utility Values
Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties
Economic Evaluation, Uncertainty in
Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis
Infectious Disease Modeling
Information Analysis, Value of
Observational Studies in Economic Evaluation
Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes
Problem Structuring for Health Economic Model Development
Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation
Searching and Reviewing Nonclinical Evidence for Economic Evaluation
Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies
Statistical Issues in Economic Evaluations
Synthesizing Clinical Evidence for Economic Evaluation
Value of Information Methods to Prioritize Research
Valuing Informal Care for Economic Evaluation

Efficiency and Equity

Efficiency and Equity in Health: Philosophical Considerations
Efficiency in Health Care, Concepts of
Equality of Opportunity in Health
Evaluating Efficiency of a Health Care System in the Developed World
Health and Health Care, Need for
Impact of Income Inequality on Health
Measuring Equality and Equity in Health and Health Care
Measuring Health Inequalities Using the Concentration Index Approach
Measuring Vertical Inequity in the Delivery of Healthcare
Resource Allocation Funding Formulae, Efficiency of
Theory of System Level Efficiency in Health Care
Welfarism and Extra-Welfarism

Global Health

Education and Health in Developing Economies
Fertility and Population in Developing Countries

Health and Its Value

Health and the Macroeconomy

Health Econometrics

Health Insurance

Human Resources

Markets in Health Care

Pharmaceutical and Medical Equipment Industries

Public Health

Supply of Health Services

Ambulance and Patient Transport Services
Cost Function Estimates
Healthcare Safety Net in the US
Home Health Services, Economics of
Long-Term Care
Production Functions for Medical Services
Understanding Medical Tourism

PREFACE

What Do Health Economists Do?

This encyclopedia gives the reader ample opportunity to read about what it is that health economists do and the ways in which they set about doing it. One may suppose that health economics consist of no more than the application of the discipline of economics (that is, economic theory and economic ways of doing empirical work) to the two topics of health and healthcare. However, although that would usefully uncouple 'economics' from an exclusive association with 'the (monetized) economy,' markets, and prices, it would miss out a great deal of what it is that health economists actually do, irrespective of whether they are being descriptive, theoretical, or applied. One distinctive characteristic of health economics is the way in which there has been a process of absorption into it (and, undoubtedly, from it too); in particular, the absorption of ideas and ways of working from biostatistics, clinical subjects, cognitive psychology, decision theory, demography, epidemiology, ethics, political science, public administration, and other disciplines already associated with 'health services research' (HSR) and, although more narrowly, 'health technology assessment' (HTA). But to identify health economics with HSR or HTA would also miss much else that health economists do.

… And How Do They Do It?

As for the ways in which they do it, in practice, the overwhelming majority of health economists use the familiar theoretical tools of neoclassical economics, although by no means all (possibly not even a majority) are committed to the welfarist (specifically the Paretian) approach usually adopted by mainstream economists when addressing normative issues, which actually turns out to have been a territory in which some of the most innovative ideas of health economics have been generated. Health economists are also more guarded than most other economists in their use of the postulates of soi-disant 'rationality' and in their beliefs about what unregulated markets can achieve. To study healthcare markets is emphatically not, of course, necessarily to advocate their use.

A Schematic of Health Economics

To think of health economics merely in these various restricted ways would be indeed to miss a great deal. The broader span of subject matter may be seen from the plumbing diagram, in which I have attempted to illustrate the entire range of topics in health economics. A version of the current schematic first appeared in Williams (1997, p. 46). The content of the encyclopedia follows, broadly, this same structure. The arrows in the diagram indicate a natural logical and empirical order, beginning with **Box A** (Health and its value) (**Figure 1**).

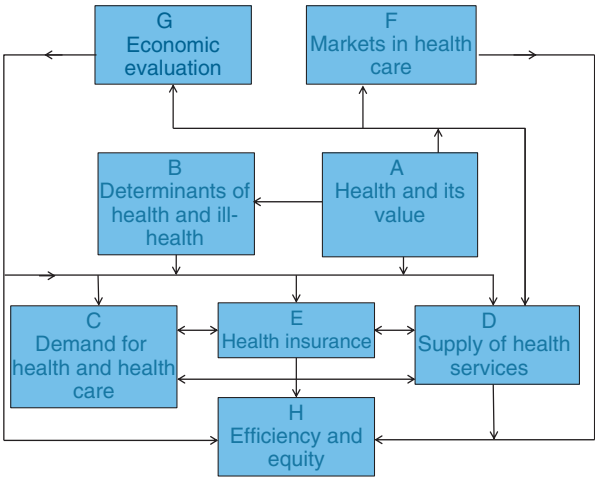


Figure 1 A schematic of health economics.

Box A, in the center-right of the schematic, contains fundamental concepts and measures of population health and health outcomes, along with the normative methods of welfarism and extra-welfarism; measures of utility and health outcomes, including their uses and limitations; and methods of health outcome valuation, such as willingness to pay and experimental methods for revealing such values, and their uses and limitations. It includes macro health economic topics like the global burden of disease, international trade, public and private healthcare expenditures, Gross Domestic Product (GDP) and healthcare expenditure, technological change, and economic growth. Some of the material here is common to epidemiology and bioethics.

Box A Health and its value

Concepts and measures of population health and health outcomes.
Ethical approaches (e.g., welfarism and extrawelfarism).
Measures of utility and the principal health outcome measures, their uses, and limitations.
Health outcome valuation methods, willingness to pay, their uses, and limitations.
Macro health economics: global burdens of disease, international trade, healthcare expenditures, GDP, technological change, and economic growth.

Box B (Determinants of health and ill health) builds on these basics in various 'big-picture' topics, such as the population health perspective for analysis and the determinants of lifetime health, such as genetics, early parenting, and schooling; it embraces occupational health and safety, addiction (especially tobacco, alcohol, and drugs), inequality as a determinant of ill health, poverty and the global burden of disease in low- and middle-income countries, epidemics, prevention, and public health technologies. Here too, much is

Box B Determinants of health and ill health

The population health perspective.
Early determinants of lifetime health (e.g., genetics, parenting, and schooling).
Occupational health and safety.
Addiction: tobacco, alcohol, and drugs.
Inequality as a determinant of ill health.
Poverty and global health (in LMICs).
Epidemics.
Prevention.
Public health technologies.

shared, both empirically and conceptually, with other disciplines.

From this it is a relatively short step into **Box C** (Demand for health and healthcare): here we are concerned with the difference between demand and need; the demand for health as 'human capital'; the demand for healthcare (as compared with health) and its mediation by 'agents' like doctors on behalf of 'principals'; income and price elasticities; information asymmetries (as in the different types of knowledge and understandings by patients and healthcare professionals, respectively) and agency relationships (when one, such as a health professional, acts on behalf of another, such as a patient); externalities or spillovers (when one person's health or behavior directly affects that of another) and publicness (the quality which means that goods or services provided for one are also necessarily provided for others, like proximity to a hospital); and supplier-induced demand (as when a professional recommends and supplies care driven by other interests than the patient's).

Box C Demand for health and healthcare

Demand and need.
The demand for health as human capital.
The demand for healthcare.
Agency relationships in healthcare.
Income and price elasticities.
Information asymmetries and agency relationships.
Externalities and publicness.
Supplier-induced demand.

Then comes **Box D** (Supply of healthcare) covering human resources; the remuneration and behavior of professionals; investment and training of professionals in healthcare; monopoly and competition in healthcare supply; for-profit and nonprofit models of healthcare institutions like hospitals and clinics; health production functions; healthcare cost and production functions that explore the links between 'what goes in' and 'what comes out;' economies of scale and scope; quality of care and service; and the safety of interventions and modes of delivery. It includes the estimation of cost functions and the economics of the pharmaceutical and medical equipment industries. A distinctive difference in this territory from many other areas of application is the need to drop the assumption

Box D Supply of health services

Human resources, remuneration, and the behavior of professionals.
Investment and training of professionals in healthcare.
Monopoly and competition in healthcare supply.
Models of healthcare institutions (for-profit and nonprofit).
Health production functions.
Healthcare cost and production functions.
Economies of scale and scope.
Quality and safety.
The pharmaceutical and medical equipment industries.

of profit-maximizing as a common approach to institutional behavior and to incorporate the idea of 'professionalism' when explaining or predicting the responses of healthcare professionals to changes in their environment.

Supply and demand are mediated (at least in the high-income world) by insurance: the major topic of **Box E** and a large part of health economics as practiced in the US. This covers the demand for insurance; the supply of insurance services and the motivations and regulations of insurance as an industry; moral hazard (the effect of insurance on utilization); adverse selection (the effect of insurance on who is insured); equity and health insurance; private and public systems of insurance; the welfare effects of soi-disant 'excess' insurance; effects of insurance on healthcare providers; and various specific issues in coverage, such as services to be covered in an insured bundle and individual eligibility to receive care. Although the health insurance industry occupies a smaller place in most countries outside the US, the issues invariably crop up in a different guise and require different regulatory and other responses.

Box E Health insurance

The demand for insurance.
The supply of insurance services.
Moral hazard.
Adverse selection.
Equity and health insurance.
Private and public systems.
Welfare effects of 'excess' insurance.
Effects of insurance on healthcare providers.
Issues in coverage: services covered and individual eligibility.
Coverage in LMICs.

Then, in **Box F**, comes a major area of applied health economics: markets in healthcare and the balance between private and public provision, the roles of regulation and subsidy, and the mostly highly politicized topics in health policy. This box includes information and how its absence or distortion corrupts markets; other forms of market failure due to externalities; monopolies and a catalog of practical difficulties both for the market and for more centrally planned systems; labor markets in healthcare (physicians, nurses, managers, and allied professions), internal markets (as when the public sector of healthcare is divided into agencies that commission care on behalf of populations and those that

Box F Markets in healthcare

Information and markets and market failure.
Labor markets in healthcare: physicians, nurses, managers, and allied professions.
Internal markets in the healthcare sector.
Rationing and prioritization.
Welfare economics and system evaluation.
Comparative systems.
Waiting times and lists.
Discrimination.
Public goods and externalities.
Regulation and subsidy.

provide it); rationing and the various forms it can take; welfare economics and system evaluation; waiting times and lists; and discrimination. It is here that many of the features that make healthcare 'different' from other goods and services become prominent.

Box G is about evaluation and healthcare investment, a field in which the applied literature is huge. It includes cost-benefit analysis, cost-utility analysis, cost-effectiveness analysis, and cost-consequences analysis; their application in rich and poor countries; the use of economics in medical decision making (such as the creation of clinical guidelines); discounting and interest rates; sensitivity analysis as a means of testing how dependent one's results are on assumptions; the use of evidence, efficacy, and effectiveness; HTA, study design, and decision process design in agencies with formulary-type decisions to make; the treatment of risk and uncertainty; modeling made necessary by the absence of data generated in trials; and systematic reviews and meta-analyses of existing literature. This territory has burgeoned especially, thanks to the rise of 'evidence-based' decision making and the demand from regulators for decision rules in determining the composition of insured bundles and the setting of pharmaceutical prices.

Box G Economic evaluation

Decision rules in healthcare investment.
Techniques of cost-benefit analysis in health and healthcare.
Techniques of cost-utility analysis and cost-effectiveness analysis in health and healthcare in rich and poor countries.
Techniques of cost-consequences analysis.
Decision theoretical approaches.
Outcome measures and their interpretation.
Discounting.
Sensitivity analysis.
Evidence, efficacy, and effectiveness.
Economics and health technology assessment.
Study design.
Risk and uncertainty.
Modeling.
Systematic reviews and meta-analyses.

The final Box, H, draws on all the preceding theoretical and empirical work: concepts of efficiency, equity, and possible conflicts between them; inequality and the socioeconomic 'gradient;' techniques for measuring equity and inequity; evaluating efficiency at the system level; evaluating equity at system level: financing arrangements; evaluating equity at system level: service access and delivery; institutional arrangements for efficiency and equity; policies against global poverty and for health; universality and comprehensiveness as global objectives of healthcare; and healthcare financing and delivery systems in low- and middle-income countries (LMICs). This is the most overtly 'political' and policy-oriented territory.

Box H Efficiency and equity

Concepts of efficiency, equity, and possible conflicts.
Inequality and the socioeconomic 'gradient.'
Evaluating efficiency: international comparisons.
Techniques for measuring equity and inequity.
Evaluating equity at system level: financing arrangements.
Evaluating equity at system level: service access and delivery.
Institutional arrangements for efficiency and equity.
Global poverty and health.
Universality and comprehensiveness.
Healthcare financing and delivery systems in LMICs.

A Word on Textbooks

The scope of a subject is often revealed by the contents of its textbooks. There are now many textbooks in health economics, having various degrees of sophistication, breadth of coverage, balance of description, theory and application, and political sympathies. They are not reviewed here but I have tried to make the (English language) list in the Further Reading as complete as possible. Because the assumptions that textbook writers make about the preexisting experience of readers and about their professional backgrounds vary, not every text listed here will suit every potential reader. Moreover, a few have the breadth of coverage indicated in the schematic here. Those interested in learning more about the subject to supplement what is to be gleaned from the pages of this encyclopedia are, therefore, urged to sample what is on offer before purchase.

Acknowledgments

My debts of gratitude are owed to many people. I must particularly thank Richard Berryman (Senior Project Manager), at Elsevier, who oversaw the inception of the project, and Gemma Taft (Project Manager) and Joanne Williams (Associate Project Manager), who gave me the most marvelous advice and support throughout. The editorial heavy lifting was done by Billy Jack and Karen Grépin (Global Health); Aki Tsuchiya and John Wildman (Efficiency and Equity); John Cawley and Kosali Simon (Determinants of Health and Ill health); Richard Cookson and Mark Suhrcke (Public Health); Erik Nord (Health and its Value); Richard Smith (Health and the

Macroeconomy); John Mullahy and Anirban Basu (Health Econometrics); Tom McGuire (Demand for Health and Healthcare); John Nyman (Health Insurance); Jim Burgess (Supply of Health Services); Martin Gaynor and Sean Nicholson (Human Resources); Patricia Danzon (Pharmaceutical and Medical Equipment Industries); Pau Olivella and Pedro Pita Barros (Markets in Healthcare); and John Brazier, Mark Sculpher, and Anirban Basu (Economic Evaluation). Finally, my thanks to the Advisory Board: Ron Akehurst, Andy Briggs, Martin Buxton, May Cheng, Mike Drummond, Tom Getzen, Jane Hall, Andrew Jones, Bengt Jonsson, Di McIntyre, David Madden, Jo Mauskopf, Alan Maynard, Anne Mills, the late Gavin Mooney, Jo Newhouse, Carol Propper, Ravindra Rannan-Eliya, Jeff Richardson, Lise Rochaix, Louise Russell, Peter Smith, Adrian Towse, Wynand Van de Ven, Bobbi Wolfe, and Peter Zweifel. Although the Board was not called on for frequent help, their strategic advice and willingness to be available when I needed them was a great comfort.

Anthony J Culyer
Universities of Toronto (Canada) and York (England)

Further Reading

Cullis, J. G. and West, P. A. (1979). *The economics of health: An introduction*. Oxford: Martin Robertson.

Donaldson, C., Gerard, K., Mitton, C., Jan, S. and Wiseman, V. (2005). *Economics of health care financing: The visible hand*. London: Palgrave Macmillan.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. and Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*, 3rd ed. oxford: Oxford University Press.

Evans, R. G. (1984). *Strained mercy: The economics of Canadian health care*. Markham, ON: Butterworths.

Feldstein, P. J. (2005). *Health care economics*, 6th ed. Florence, KY: Delmar Learning.

Folland, S., Goodman, A. C. and Stano, M. (2010). *The economics of health and health care*, 6th ed. Upper Saddle River: Prentice Hall.

Getzen, T. E. (2006). *Health economics: Fundamentals and flow of funds*, 3rd ed. Hoboken, NJ: Wiley.

Getzen, T. E. and Allen, B. H. (2007). *Health care economics*. Chichester: Wiley.

Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (eds.) (1996). *Cost-effectiveness in health and medicine*. New York and Oxford: Oxford University Press.

Henderson, J. W. (2004). *Health economics and policy with economic applications*, 3rd ed. Cincinnati: South-Western Publishers.

Hurley, J. E. (2010). *Health economics*. Toronto: McGraw-Hill Ryerson.

Jack, W. (1999). *Principles of health economics for developing countries*. Washington, DC: World Bank.

Jacobs, P. and Rapoport, J. (2004). *The economics of health and medical care*, 5th ed. Sudbury, MA: Jones & Bartlett.

Johnson-Lans, S. (2006). *A health economics primer*. Boston: Addison Wesley/ Pearson.

McGuire, A., Henderson, J. and Mooney, G. (1992). *The economics of health care*. Abingdon: Routledge.

McPake, B., Normand, C. and Smith, S. (2013). *Health economics: An international perspective*, 3rd ed. Abingdon: Routledge.

Mooney, G. H. (2003). *Economics, medicine, and health care*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

Morris, S., Devlin, N. and Parkin, D. (2007). *Economic analysis in health care*. Chichester: Wiley.

Palmer, G. and Ho, M. T. (2008). *Health economics: A critical and global analysis*. Basingstoke: Palgrave Macmillan.

Phelps, C. E. (2012). *Health economics*, 5th (international) ed. Boston: Pearson Education.

Phillips, C. J. (2005). *Health economics: An introduction for health professionals*. Chichester: Wiley (BMJ Books).

Rice, T. H. and Unruh, L. (2009). *The economics of health reconsidered*, 3rd ed. Chicago: Health Administration Press.

Santerre, R. and Neun, S. P. (2007). *Health economics: Theories, insights and industry*, 4th ed. Cincinnati: South-Western Publishing Company.

Sorkin, A. L. (1992). *Health economics – An introduction*. New York: Lexington Books.

Walley, T., Haycox, A. and Boland, A. (2004). *Pharmacoeconomics*. London: Elsevier.

Williams, A. (1997). Being reasonable about the economics of health: Selected essays by Alan Williams (edited by Culyer, A. J. and Maynard, A.). Cheltenham: Edward Elgar.

Witter, S. and Ensor, T. (eds.) (1997). *An introduction to health economics for eastern Europe and the Former Soviet Union*. Chichester: Wiley.

Witter, S., Ensor, T., Jowett, M. and Thompson, R. (2000). *Health economics for developing countries. A practical guide*. London: Macmillan Education.

Wonderling, D., Gruen, R. and Black, N. (2005). *Introduction to health economics*. Maidenhead: Open University Press.

Zweifel, P., Breyer, F. H. J. and Kifmann, M. (2009). *Health economics*, 2nd ed. Oxford: Oxford University Press.

CONTENTS OF ALL VOLUMES

VOLUME 1

VOLUME 2

VOLUME 3

Pharmaceutical Company Strategies and Distribution Systems in Emerging Markets

P Yadav and L Smith, University of Michigan, Ann Arbor, MI, USA

Background

Decline in research productivity, slow growth in mature markets, a massive number of patent expirations, and pressures for cost containment from major payers together are forcing leading pharmaceutical companies to rethink their strategies for growth in emerging markets. According to Intercontinental Marketing Services (IMS) Health, by 2014 50% of market growth in pharmaceutical industry will come from 17 countries in emerging markets. With emerging economies representing close to 55% of the world's gross domestic product, pharmaceutical companies have identified new opportunities and markets for investment in these contexts. Rapid economic growth combined with continued population increases in countries like India, China, and Brazil offer new business opportunities for pharmaceutical companies. Additionally, improved technological capabilities and industrial development as a result of economic growth create improved methods for engaging in these developing markets.

In low- and middle-income countries, national growth requires health distribution systems that respond to the population's diverse range of health conditions including assured access to appropriate products, medicines, and vaccines to treat and manage those conditions. More specifically, many developing countries require health products, medicines, and vaccines that can address a double burden of disease; one of both communicable diseases (CDs), infectious diseases as well as noncommunicable diseases (NCDs), chronic diseases. In 2002, approximately 46% of the global burden of disease was attributable to NCDs (Young *et al.*, 2009). The resulting interaction between NCDs and CDs may increase susceptibility of individual health. A synergistic negative interaction of disease types also appears prevalent among individuals in low- and middle-income countries. For example, individuals with diabetes may struggle to manage their health with regard to exposure to infection, particularly to the eyes and feet. Likewise, close to a quarter of cancers in developing countries may be attributed to infectious agents. As a result of these and other interactions, individuals require treatment for single disease areas, as well as increasingly for multiple conditions, some of which may be more chronic in nature.

The challenges in managing pharmaceuticals in emerging countries are very different from those in advanced economies. In developed countries, the challenge is to enhance the efficiency of spending on pharmaceuticals. In most of the emerging markets, the main challenge is how to expand the pharmaceutical market to include a larger share of the population. In these economies, increased spending on pharmaceuticals could lead to higher benefits in health outcomes and in some cases could also catalyze economic growth. As social health insurance plans in some of the emerging countries expand in terms of their population coverage and the range of services that are included, many are being cautious not to generate fiscal pressures at an early stage that may hamper their sustainability. The governments of many emerging market countries are attempting to expand the reach and coverage of their pharmaceutical systems in a way that avoids the high costs observed in the health systems of more developed economies.

The Unique Characteristics of Emerging Markets

For pharmaceutical manufacturers participating in pharmaceutical market growth in emerging markets, careful analysis of the structural differences in emerging market pharmaceutical systems is required. For many governments, strategies to ensure consistent access to medicines, vaccines, and health technologies in emerging markets are necessary. To ensure market growth in these countries, manufacturers often align their strategies with government strategies for improving access through innovative mechanisms like differential pricing schemes or local marketing and health education campaigns for high burden disease areas.

There are significant differences among developing countries in terms of total health expenditures, total pharmaceutical expenditures, strength of drug regulatory authorities, social health insurance, and relative division of distribution between public and private sector for pharmaceuticals. Distribution strategies will vary according to the exact set of factors present in a country. Main factors, which hold across most developing countries, are discussed below.

In mature developed markets, some form of health insurance usually reimburses pharmaceutical purchases in whole or in part. Large fractions of these populations are insured through state, employer, or private insurance. In contrast, many developing countries do not have a national health insurance system in place to ensure affordable access to services and health products for the population. Public or private health insurance is limited mainly to those with high expendable income or those employed in the formal sector. As such, the market for medicines is largely dependent on ability and willingness to pay for products. In low- or middle-income countries, purchasing of pharmaceuticals accounts for up to 40% of the total healthcare expenditure whereas in many established market economies, only 20% of costs are attributed to pharmaceutical drugs. Furthermore, in low-income countries approximately 80% of total pharmaceutical expenditures are out of pocket (**Figures 1** and **2**).

In countries like Brazil and China, newer programs of social insurance have created alternative methods to help ensure individual access to medicines through subsidies and/or targeted retail pharmacy programs for the rural poor. China has expanded its public health insurance system to cover almost

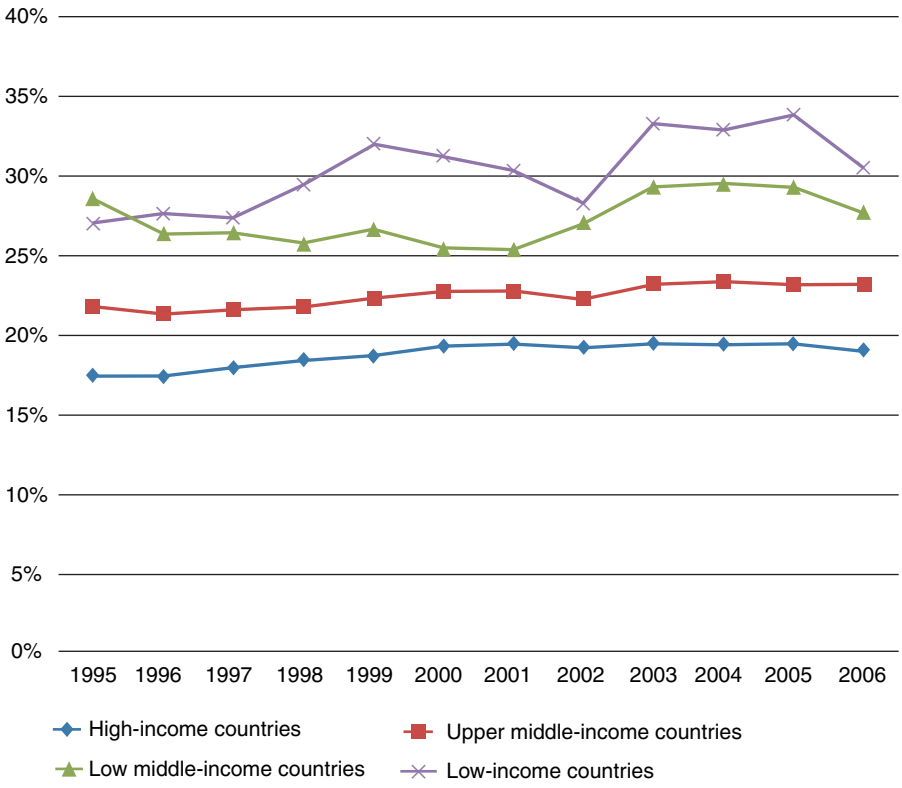


Figure 1 Average total pharmaceutical expenditure as a percent of the total health expenditure (1995–2006). Reproduced from Global Health Expenditure Database (1995–2006). National health accounts, World Health Organization. Available at: http://www.who.int/nha/en/ (accessed 03.10.13).

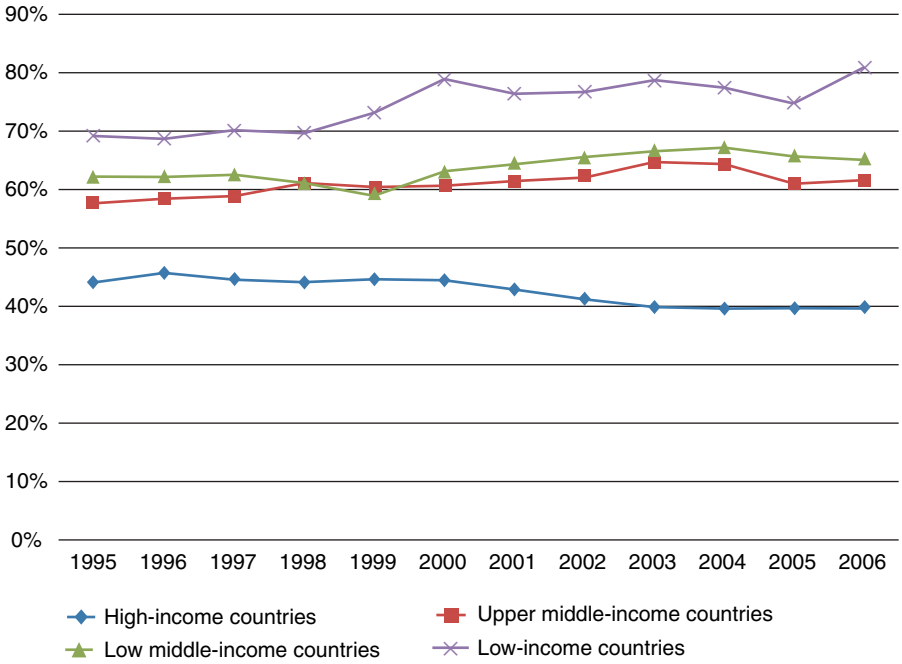


Figure 2 Average percent of total pharmaceutical expenditures that are out of pocket (1995–2006). Reproduced from Global Health Expenditure Database. National health accounts, World Health Organization. Available at: http://www.who.int/nha/en/ (accessed 18.07.13).

Table 1 Differences in overall structure of pharmaceutical market in developed and developing countries

Factor	*Developed countries*	*Developing countries*
Payer/reimbursement	Strong presence of public or private insurance companies and limited out-of-pocket expenditure	Mostly payments are made out of pocket. Social health insurance systems are expanding in many emerging markets. Private insurance plans are also growing in some emerging market countries
Regulatory structure	Strong well-defined laws and overall good ability to enforce regulations	Weak fragmented regulatory structures, ill-defined laws in some instances, and poor ability to enforce regulations
Patented generic versus branded generic	The market for prescription drugs consists of patented drugs and generics	Poor regulatory structure creates a strong market for branded generics (brand is used a signal of quality by the patient)
Prescription adherence	Prescription drugs can only be dispensed with a formal prescription	Retail pharmacies often dispense medicines and also act as the first point of healthcare contact for many patients
Balance of power in the system	Buyer (insurance companies or national health system) monopsony creates good balance of power between the manufacturer and the patients. In the US pharmacy benefit managers (PBM) and drug formularies are commonly used as a means to ensure further balance of power	Balance of power is tilted toward the manufacturer and the distribution channel. A large fraction of patients purchase using out-of-pocket funds and have little bargaining power
Price sensitivity	Pricing is crucial to gaining formulary or national reimbursement acceptance, but price sensitivity within bands is lower	Out-of-pocket payments lead to high price elasticity. Pricing is a key strategic differentiator

95% of the population. Since the early 1990s, Brazil has begun a social health insurance program called the Single Unified System (SUS) on which more than 75% of the population relies on exclusively for care. Approximately 20% of the population (wealthier socioeconomic strata and employees of certain businesses) purchases health insurance from private insurers who are regulated by the National Supplementary Health Agency. People who purchase private insurance receive a tax rebate, however, they still are required to contribute to the SUS through their income taxes. Lower-income groups tend to spend more out of pocket on medicines than higher-income groups in Brazil as higher-income groups typically purchase separate private insurance. Private sources of finance, such as out-of-pocket spending by families and companies with some direct and indirect government subsidies, fund most medicines in the Brazilian health system. Other developing countries such as Malaysia and Thailand have public health insurance plans with very high degrees of coverage. However, in most cases the breadth of services covered under the plans remains limited to catastrophic healthcare needs or very basic services such as immunization. Patients usually have to pay out of pocket for other outpatient or inpatient services and for purchasing medicines.

Also, in developed countries most pharmaceuticals are only obtained using a prescription provided by a physician and consumers purchase drugs from retail pharmacies. In many developing countries, adherence to prescription requirements is poor and patients may often obtain medicines at retail pharmacies without a formal prescription. Most developed countries have well-developed regulatory institutions to ensure the safety and efficacy of drugs. Regulatory agencies in low- and middle-income countries are weak and have very limited capacity to enforce rules. As a result, the task of ensuring quality of medicines is placed on the patients. This creates a stronger market for branded generics in developing countries than in developed countries as branded medicines are used as a signal to patients for quality products.

In low-income countries, especially in Africa, direct purchasing and distribution of medicines by the government (Ministries of Health) represents a significant portion of overall market for pharmaceuticals (Table 1).

Pharmaceutical Distribution Systems in Developed Markets

In both developed and developing countries, given the large number of medicines and packaging variants, it is difficult for retail pharmacies to purchase all of these products directly from the manufacturer and stock them in their retail stores. If all retail pharmacies and hospitals ordered all their medicines from the hundreds of different manufacturers, the number of transactions would be exponentially large and inhibit a working system. Retail pharmacies therefore depend on a well-functioning distribution system consisting of distributors, wholesalers, and prewholesalers. Distributors typically store and distribute a manufacturer's product for a fee; however, they do not own the inventory they distribute. Wholesalers are intermediaries that purchase medicines from manufacturers or prewholesalers. Wholesalers store and distribute supplies while also managing the risks associated with purchased inventory.

In most developed countries a large fraction of the pharmaceuticals used are prescribed by clinical or hospital physicians, which patients obtain at retail pharmacies. In the US, for example, almost 75% of the pharmaceutical sales are covered by retail pharmacies. There are more than 57 000 pharmacies in the US and almost 50% of the retail pharmacy market consists of chain pharmacies including food stores with pharmacies (Yadav *et al.*, 2012). In contrast, many countries in Europe do not allow chain pharmacies. Some large chain pharmacies and hospitals purchase their drugs directly from manufacturers and run their own distribution networks. Some manufacturers also ship certain specialty products directly to retail pharmacies through a specialized logistics service provider. Even with these alternatives, the greater portion of distribution occurs through wholesalers and distributors.

In Europe the manufacturer frequently sends the product from production to a prewholesaler first who then ships the product to wholesalers or large hospitals. Wholesalers then distribute the product to retail pharmacies with an average delivery frequency of twice daily. Most wholesalers stock and distribute products from a number of different manufacturers and often multiple wholesalers operate in a particular region offering competition and choice to the pharmacies. In the US manufacturers ship their product to distributors who then distribute the product to the retail pharmacies several times a week. The lower delivery frequency to retail pharmacies in the US implies that on average pharmacies in the US carry more inventory than pharmacies in Europe, although this is not always true.

In most developed regions of the world the wholesaling and distribution segment of the distribution system is concentrated amongst a few players. The three largest US wholesalers, Cardinal Health, McKesson, and AmerisourceBergen, distribute more than 90% of all pharmaceuticals sold in the US. Increasingly, these companies behave more like distributors in that they have inventory management agreements with manufacturers under which they do not necessarily own stock or carry the associated inventory risk but instead receive a fee from the manufacturers. Similar to the US, in Europe, Japan and other developed regions of the world four to five major distributors with national coverage account for 90% of the market. This is due to the underlying economies of scale in the pharmaceutical distribution business.

Financial flows within developed market distribution chains are somewhat complicated. Although money flows from health plans/insurers to manufacturers with the retail pharmacy in between may seem straightforward, the nature of price negotiations and discounting makes this system more complex. In the US, for example, insurers may negotiate prices with manufacturers. Alternatively, health plans may create or contract specialized agencies called pharmacy benefit managers (PBMs) to obtain discounted prices from manufacturers for exclusiveness on formulary or volume-based discounts. Hospitals and other providers work through Group Purchasing Organizations (GPOs) for negotiating these discounts. The discounts obtained are then shared with health plans. Similar arrangements also exist in the UK.

Order information flows from the healthcare provider to the pharmacy when a prescription is 'called in.' Pharmacies and hospitals replenish their inventories by ordering from the distributors or wholesalers. The distributors and wholesalers in turn order from the manufacturer when their stock needs replenishment. In addition, retail sales information and distributor sales information are collected by private third parties like IMS Health, which allow payers, manufacturers, and regulators to access information from each point in the supply chain. Owing to the nature of the financial flows and contracting with GPOs and PBMs, the distributor and the retail pharmacy also provide product sales information to the GPO and PBM.

Pharmaceutical Distribution Systems in Emerging Markets

Similar to the developed countries, in most emerging markets medicines in the private sector are distributed through a network of importers, wholesalers, subwholesalers, pharmacies, and drug stores (Pharmacies included a trained and certified pharmacist. Drug stores, often also referred to as chemists, are additional informal or formal retail distribution points for medicines typically without a trained/registered pharmacist. Many countries have formal regulatory mechanisms to allow for legal distribution of medicines through drug shops.) National importers and wholesalers create the link between pharmaceutical manufacturers and retail pharmacies, private clinics, hospitals, and other informal drug shops. However, for historical reasons, the pharmaceutical distribution system in most emerging markets has a different market structure compared to developed countries. The main differences include a lack of distribution networks with national reach; excessive fragmentation and too many small players; too many intermediaries between the manufacturer and the patient; poor IT and communication flow systems resulting in poor coordination across actors in the distribution channel.

For example, in 2007 the total number of retail pharmaceutical dispensing points in China was approximately five times that of the US (approximately 50 000 in the US and 140 000 in China) and the total number of wholesalers and distributors in China was close to 16 500 (Zhou, 2007). The total market share held by the top-three largest Chinese pharmaceutical distributors was only 42% as compared with that of the three leading distributors in the US, which account for 90% of the market share. Fragmentation of the wholesale market is commonly observed in most emerging markets with trends similar to that of China observed in India, Brazil, and other developing regions with large private sector markets for pharmaceuticals. Capital constraints faced by wholesalers coupled with corruption that favors certain wholesalers in purchasing by hospitals prevents large-scale consolidation in the industry. These factors do not allow scale economies to have their full effect and a fragmented wholesaling and distribution sector continues to exist.

Owing to the relative lack of distributors/wholesalers with nationwide coverage and reach, wholesalers often have to rely on regional subwholesalers or stockists, which adds another layer to the supply chain. High markups between multiple intermediaries in the distribution chain result in poor affordability and inability of the manufacturer to pursue growth strategies that rely on reaching larger proportions of the population at lower price points. Lack of information flows

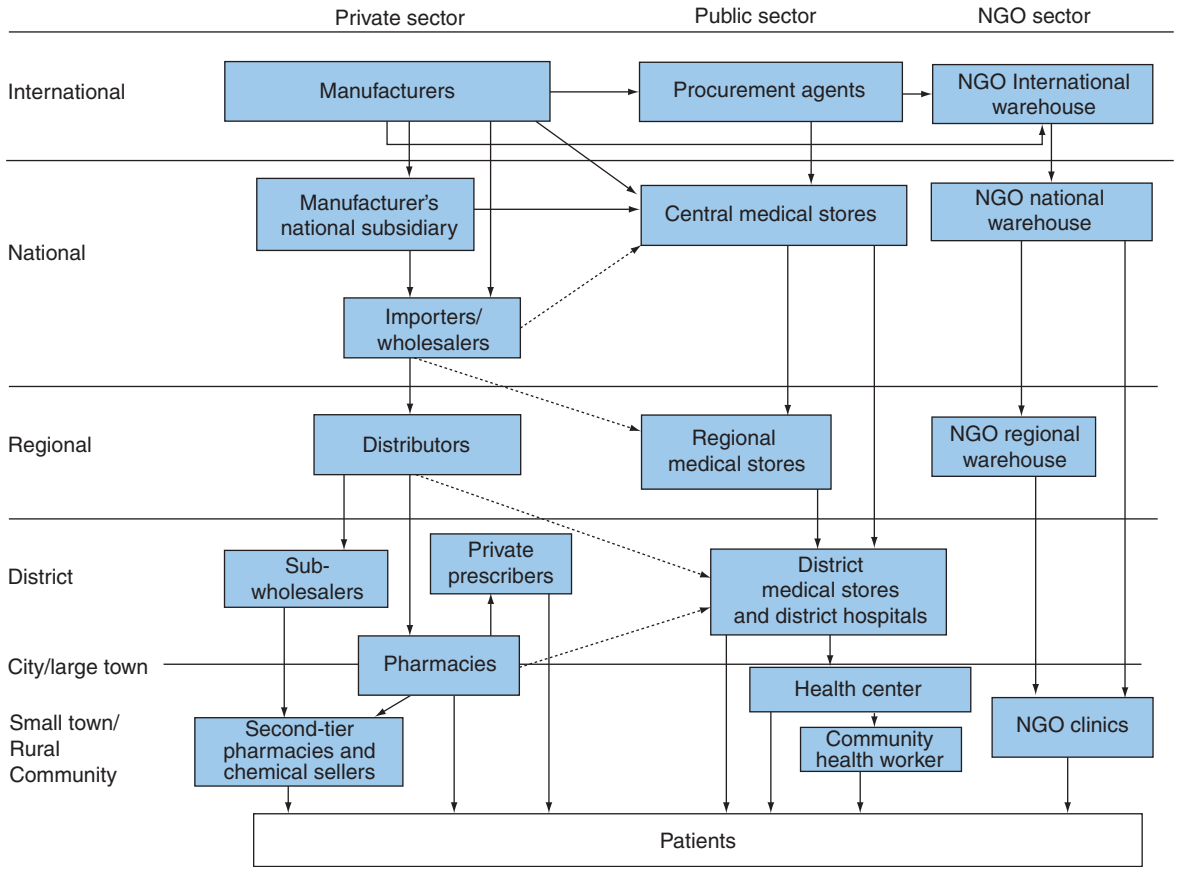


Figure 3 Distribution network for essential medicines in the public, private, and NGO channels in developing countries. Reproduced from Yadav, P., Tata, H. and Babaley, M. (2012). *The World Medicines Situation 2011: Storage and supply chain management*. Geneva: World Health Organization.

and opacity of information at different nodes in the system worsens the problem.

Apart from the private sector, in many low- and lower middle-income countries distribution of medicines also takes place in the public and nongovernmental organization (NGO)/faith-based sector (Figure 3).

In sub-Saharan Africa the predominant model is public sector distribution of medicines through a central medical store (CMS) which coordinates directly with regional or district stores. Transportation of goods is managed by a government/CMS owned and managed fleet. In addition to the CMS and regional or district stores, there are a number of primary and secondary distribution locations. These additional locations are present because of product- or program-specific supply chains setup by funding partners of the public sector. Increased availability of funding for the procurement of medicines over the past 5–10 years has highlighted weaknesses in the public sector medicine supply systems in some emerging markets.

Country case studies of Zambia and Jordan provide specific examples of distribution systems. The Zambian system represents a predominantly public sector medicine supply system. Jordan, by contrast, provides a private sector example. Each case outlines opportunities and challenges relevant to the previous distribution system review. The case study in Zambia outlines the difficulty the public sector has ensuring reliable supply and availability. In comparison the private sector in Jordan has difficulty in ensuring quality and price of medicines (Tables 2 and 3).

Distribution and delivery systems may also depend on the involvement of NGOs or faith-based organizations. Pharmaceutical delivery in this context is typically arranged according to the customer's own prearrangement, courier services, drug supply organization delivery services, or direct delivery services. Variations in the distribution structures for this specific channel are considerable across countries.

Generally, procurement, distribution, and the overarching provision of pharmaceutical goods within a particular country are disaggregated across groups with limited information sharing between public sector, private sector, and NGO/donor groups. As a result, pharmaceutical procurement and distribution strategies lack the coordination and efficiency to ensure optimal market function. High transaction costs and opacity in the market due to excess fragmentation leads to higher retail prices of pharmaceutical products. Although coordination and information sharing is a problem in developed countries too, a wider use of information technology and the presence of information consolidators such as IMS Health makes it somewhat easier.

Price is a key strategic differentiator between distribution chains in developed markets compared to those in emerging markets. Emerging market distribution chains often involve high markups between multiple intermediaries, thus making it

Table 2 Case study – Zambia

Public sector medicine supply distribution	*Distribution opportunities and challenges*
Government of Zambia stores and distributes medicines through a national medical store, Medical Stores Limited (MSL). Additional management support is contracted through Crown Agents	The MSL is engaged with typical demand expectations at each district. Such engagement enables the MSL to make discretionary decisions related to overestimated demand expectations after a period of product stockout. Although the MSL discretionary decision making may prevent a flood of stock from reaching the district level and expiry before potential use, it does not encourage transparent information flows
Districts are distributed medicines on a monthly basis by the MSL according to preset allocations of medicines. Allocations are based on reported demand at the district level, of medicines. Districts send orders directly to the MSL utilizing information on their current stock levels as well as a review the monthly stock availability report provided by the MSL	Additionally, districts often order according to the stock availability report provided monthly by the MSL. If district level representatives know an item is out of stock, their orders to the MSL will reflect what they expect to be able to receive, not necessarily the true demand. The MSL in turn, overtime, operates on accurate demand information
In addition to the demand information, all districts receive a standard, predetermined number of medicine kits every month	

Source: Reproduced with permission from Yadav, P. (2007). Analysis of public, private and mission sector supply chains for essential drugs in Zambia. *Technical Report*. Zaragoza, Spain: MIT-Zaragoza International Logistics Program.

Table 3 Case study – Jordan

Private sector medicine supply distribution	*Distribution opportunities and challenges*
The private sector supply chain in Jordan consists of several large importers and wholesalers from international manufacturers/ suppliers. There are also a number of local manufacturers that produce generic medications, primarily for export to other countries. Medicines are sold within the country to private hospitals, retail pharmacies, and drug outlets	National regulation of all importers and wholesalers ensures that pharmaceutical products have the required quality storage facilities. The local manufacturing groups have improved lead times as well as higher overall responsiveness for distribution of medicines within the country; however, their larger market is in exporting medicines to other countries
Importers and wholesalers are often granted lines of credit from suppliers and manufacturers that vary in length based on the volume of orders. The longer the period of credit requested, the larger the order required per time period. As a result, importers and wholesalers are often responsible for the larger sized order, the interest on a line of credit as well as typical, distribution and related transportation costs of their goods within the country	Prices in the private sector tend to be high. Given the relatively small market for medicines in Jordan, it is often difficult for suppliers to reach economies of scale unless they tender to public and private sector facilities. The importers and wholesalers carry additional costs resulting from poor credit provisioning in the system and the high costs of credit

Source: Reproduced with permission from Conesa, S. and Yadav, P. (2009). Analysis of the pharmaceutical supply chain in Jordan. *Technical Report*. Zaragoza, Spain: MIT-Zaragoza International Logistics Program.

difficult to ensure affordability. Reaching a larger portion of rural segments of a population with alternative pricing models can be a successful growth and pricing strategy, however, without an organized and efficient distribution system, high costs and multiple sources of uncertainty prohibit the widespread use of such approaches. Poor coordination between retail pharmacies, drug shops, private clinics, and other informal service channels precludes a systematic method for targeting inexpensively priced medicines specifically to the poor. It may also preclude increasing revenue for manufacturers at the time as improving access. Newer distribution strategies that ensure better flow tracking will enable strategies for high sales volumes (a benefit to manufacturers) with lower margins (a benefit to the general population) rather than the current model of high margin, low volume.

Strategy for Emerging Markets

Success in emerging markets requires selecting prices that lead to high affordability by the population and revenue growth for the pharmaceutical company. In some cases partnerships or equity ownership in local companies helps achieve some of these objectives.

Differential Pricing in Emerging Markets

Pharmaceutical companies have used alternative pricing models to target emerging markets. Differential pricing is based on the economic principle that the greatest profit may be derived from pricing products closest to a consumer's maximum willingness and ability to pay for that product. Typically, consumers are placed within groups according to their wealth and products are distributed at different pricing tiers to the different groups. When well designed, differential pricing for pharmaceuticals can increase affordability for patients and increase profits for the pharmaceutical company. Many pharmaceutical companies have developed and implemented differential pricing schemes with success. GlaxoSmithKline (GSK) sells a portfolio of 25 medicines targeted at both CD and NCD areas at significantly lowered prices in

low-income countries. Merck also runs a differential pricing scheme for their diabetes medication Januvia. Novartis has developed a differential pricing scheme for insulin in developing countries. A distribution system with flow tracking and information sharing is essential to ensure medicines reach those communities most in need of preferential pricing.

Manufacturing Partnerships and Acquisitions in Developing Countries

Some pharmaceutical companies use joint ventures and acquisitions of local manufacturing companies as a strategy to achieve growth in emerging markets. This allows selecting prices more appropriate to the emerging markets without the risks of developed countries asking for the same prices as different companies manufacture the products. It also allows leveraging the marketing and distribution strengths of the local company. GSK has created a strategic partnership with Aspen Pharma, a South African generics manufacturer under which Aspen manufactures and distributes many of GSK's products in the region. GSK owns 18.5% equity stake in Aspen. Abbott Laboratories, another large multinational pharmaceutical company recently acquired Indian manufacturer Piramal Healthcare Limited to accelerate its growth in emerging markets.

Distribution Strategy for Emerging Markets

Apart from differential pricing, success in emerging markets will require distribution networks to reach areas, which the current distribution model does not reach. This may require major adjustments to current business models. Improving supply chain reach and efficiency will serve as the foundation of such a strategy. Although public investments in infrastructure will automatically increase supply chain reach and efficiency, it is unclear whether infrastructure growth will be able to match the needs to enable growth and coverage in a timely fashion.

Regional or Country-Level Prewholesaling Operations

The current structure of multiple wholesalers/importers purchasing directly from the manufacturer often creates supply chain inefficiencies. Manufacturers often sell to and invest in a single wholesaler to save on transaction costs. Selling to one wholesaler limits sales volumes and geographical reach. Given the large number of wholesalers this leads to the practice of horizontal selling where the single wholesaler then sells the manufacturer's products to other wholesalers and then on to subwholesalers and retailers. Sales between wholesalers at the national level often add an additional transaction cost and mark-up to the final retail price.

Even though having additional intermediaries usually leads to higher markups, introducing a prewholesaling operation may help aggregate and organize a highly fragmented wholesaler base. Prewholesalers allow manufacturers to distribute their product to multiple wholesalers and achieve higher volumes of sales, product reach, and market penetration without necessarily adding to the costs of creating a company-owned distribution or commercial entity. A prewholesaler could also reduce markups between national wholesalers as well as reduce long lead times for orders improving the efficiency and financial stability of wholesalers who often have to find expensive working capital credit to cover lead times. In many instances the benefits (i.e., economies of scale and reduced transaction costs) outweigh the costs (i.e., increased margins) and a prewholesaler model helps improve overall supply chain efficiency.

New Retail Pharmacy Formats

Enhanced supply chain reach would require working with newer retail pharmacy formats. Instead of concentrating on a few hospitals and large pharmacies, growth will come from increasing points of sale. This would require accepting that not all retail points of sale for medicines will have the form, shape, and structure of a developed-country pharmacy. Regulatory hurdles will have to be carefully negotiated in making this change. Also cost for transit will have to be reduced significantly to reach a larger number of outlets without compromising margins. Health microfranchises represent a specific network of retail points of sale, which may be employed in the development of an optimum distribution network. These microfranchises are often developed as a means to improve access and quality of medicines and prescribing practices within emerging markets. Microfranchises are technically a part of the private sector, however, they are often accredited and maintain a certain quality standard regulated by the public sector. Accredited Drug Dispensing Outlets (ADDOs) in Tanzania, CARE shops in Ghana, and Child and Family Wellness (CFW) Shops in Kenya represent three similar microfranchise models that provide extensive sales networks, with a specific emphasis on serving remote communities.

New Models for Supply Chain Information Collection

Information flows in the supply chain where a third party information broker collects information about sales from each point in the supply chain may take years before it develops in many of the emerging markets. New models for collecting supply chain information should be examined for their feasibility. There are many new models that utilize technology to share real-time information throughout distribution networks. Logistimo, a web service run on mobile phones and Internet browsers, offers supply chain management tools specific to emerging markets. Similarly, All Indian Origin Chemists and Distributors Limited (AOICD) offers technology-based logistics services to pharmaceutical companies to improve supply chain and distribution network visibility and efficiency. SMS-for-Life is another example of technology (using SMS text and electronic mapping) that has effectively facilitated comprehensive and accurate stock counts of medicines at health facilities by district-level staff. SMS-for-life is a public–private partnership between Novartis, Vodafone, IBM, Ministry for Health of Tanzania, and the Roll Back Malaria Partnership.

Partnerships with Governments and Other Agencies

Leveraging public-sector resources to reach areas that a company cannot directly influence is another method for

improving reach. The resources required for expanding supply chain reach often are beyond the means of a single firm. Poor infrastructure in rural markets, such as roads, electricity, and telecommunications, create barriers to entry for many large manufacturers. The absence of mass media and communication platforms makes demand creation and awareness building among patients and community members very expensive, and inhibits the growth that can be achieved in rural markets. Partnerships should be formed with the government (at federal, state, and local levels) where common goals can be met through public–private partnerships.

Note that civil society and consumer groups often view new business models with mistrust. Partnering with organizations that have higher trust and confidence of the community ensures that the innovative distribution model can survive the early days of infancy without backlash from the civil society organizations and community groups because of misconceptions about the objectives of the model, etc.

Distribution Strategies of the Generics

Many small and large generic companies have achieved significant rural penetration in emerging markets. Carefully examining their distribution strategies can be a useful exercise while formulating a new distribution strategy. Many generic companies have set up rural-focused distribution channels and developed rural sales forces with locally trained staff. Mankind Pharmaceuticals and Cipla represent two such companies that have increased their product reach both domestically and within other countries through targeted distribution channels. This has given them stronger reach into both rural healthcare providers and rural pharmacies and drug shops. Some innovative pharmaceutical companies have also focused on the rural markets through patient education. The Arogya Parivar program from Novartis involves health educators, usually local women, who are recruited and trained to raise awareness about specific diseases. The initiative utilizes a special sales force to ensure that medicines are available in the most remote locations. The above strategies may not be sustainable for all companies, especially for companies with fewer and more costly products in their portfolio.

Conclusions

Emerging markets now represent a significant portion of the global pharmaceutical market and are growing at much faster rates than the more mature developed-country pharmaceutical markets. Pharmaceutical markets in emerging markets tend to be very different than developed markets with private sector out-of-pocket expenditures leading financing in Asia, some parts of Africa, and Latin America. Additionally, publicly funded medicines tend to be more prominent in other emerging market regions, especially within the African context. The nature of the distribution system used for pharmaceuticals in emerging markets is different from developed-country pharmaceutical markets in several ways. Successful growth strategies for emerging markets will depend on expanding the reach of supply chains as well as increasing its overall efficiency. This article presents strategies used for achieving those goals.

See also: Pharmaceuticals and National Health Systems. Pharmacies. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA

References

Yadav, P., Tata, H. and Babaley, M. (2012). *The World Medicines Situation 2011: Storage and supply chain management*. Geneva, Switzerland: World Health Organization.

Young, F., Critchley, J. A., Johnstone, L. K. and Unwin, N. C. (2009). A review of co-morbidity between infectious and chronic disease in Sub Saharan Africa: TB and diabetes mellitus, HIV and metabolic syndrome, and the impact of globalization. *Journal of Global Health* **5**, 1–9.

Zhou, E. (2007). China pharma basking in its spotlight. *Genetic Engineering and Biotechnology News* **27**(5), 60–64.

Further Reading

Banda, M., Ombaka, E., Logez, S. and Everard, M. (2006). *Multi-country study of medicine supply and distribution activities of faith-based organizations in Sub Saharan African countries*. Geneva, Switzerland: World Health Organization and Nairobi, Kenya: Ecumenical Pharmaceutical Network.

Bishai, D. M., Shah, N. M., Walker, D. G., Brieger, W. R. and Peters, D. H. (2008). Social franchising to improve quality and access in private health care in developing countries. *Harvard Health Policy Review* **9**(1), 184–197.

Cameron, A., Ewen, M., Ross-Degnan, D., Ball, D. and Laing, R. (2009). Medicines prices, availability, and affordability in 36 developing and middle-income countries: A secondary analysis. *The Lancet* **373**(9659), 240–249.

Conesa, S. and Yadav P. (2009). *Analysis of the pharmaceutical supply chain in Jordan*. MIT-Zaragoza International Logistics Program.

Danzon, P. and Towse, A. (2003). Differential pricing for pharmaceuticals: Reconciling access, R&D and patents. *International Journal of Health Care Finance and Economics* **3**, 183–205.

Health Strategies Consultancy, LLC. (2005). *Follow the pill: Understanding the U.S. commercial pharmaceutical supply chain*. Menlo Park, California: Kaiser Family Foundation.

Kanavos, P., Schurer, W. and Vogler, S. (2011). *The pharmaceutical distribution chain in the European Union: Structure and impact on pharmaceutical prices*. Brussels, Belgium: European Commission.

Looney, W. (2010). Strategies for emerging markets: Seven keys to the kingdom. *Pharmaceutical Executive* **30**(8), 54–60.

Macarthur, D. (2007). *European pharmaceutical distribution: Key players, challenges and future strategies*. London: Informa UK Ltd.

Velásquez, G., Madrid, Y. and Quick, J. (1998). *Selected topics in health reform and drug financing. Health economics and drugs series, No. 6. Action Programme on Essential Drugs*. Geneva, Switzerland: World Health Organization.

World Health Organization (2002). *The World Health Report 2002: Reducing risks, promoting healthy life*. Geneva, Switzerland: World Health Organization.

Yadav, P. (2007). *Analysis of public, private and mission sector supply chains for essential drugs in Zambia*. MIT-Zaragoza International Logistics Program.

Yadav, P. (2010). Differential pricing for pharmaceuticals: Review of current knowledge, new findings and ideas for action. Available at: http://apps.who.int/medicinedocs/en/m/abstract/Js18390en/ (accessed 03.10.13).

Pharmaceutical Marketing and Promotion

DM Dave, Bentley University, Waltham, MA, USA

Introduction

Between 1980 and 2010, expenditures on prescription (Rx) drugs in the US increased almost 1500%, from $53 to $831 per person (see Figure 1). Spending on Rx drugs has generally outpaced the growth in national health expenditures, doubling its share to 10%, and making it one of the fastest growing components of health care costs (see Figure 2). The growth in the share of prescription drug expenditures has coincided with the growth in pharmaceutical promotion, which increased from $11.4 billion in 1996 to $29.9 billion in 2005 (Donohue *et al.*, 2007) and $32.3 billion in 2008 (SK&A, 2011). In recent years, both Rx drug spending and promotional spending have leveled off, however, due to patent expiration on certain major drugs (such as Advair, Prevacid, and Lipitor) that are not replaced by new on-patent drugs.

Promotion of prescription drugs is generally limited to drugs on patent. It includes direct-to-consumer advertising (DTCA) on broadcast and print media as well as direct-to-physician promotion (DTPP) through visits by company representatives to providers (known as detailing), free samples provided to physicians, and advertising in professional journals. Although DTPP still comprises most of the promotional budget (approximately 83% in 2011; SK&A, 2011), the largest relative increase in promotion between 1995 and 2005 resulted from the expansion of DTCA into broadcast media. The share of total promotional spending allocated to DTCA increased from less than 1% in the early 1990s to 8.6% in 1996 and 14.5% in 2003 (see Figure 3), and has remained relatively stable since.

Pharmaceutical promotion remains controversial and is facing increased public scrutiny. At the heart of the debate is whether such marketing is welfare-enhancing. The pharmaceutical industry claims that both consumer-directed and physician-directed advertising educates patients and providers on potential treatment options, opens up lines of communication between the patient and the physician, and can even increase patient–physician contact or expand appropriate treatment for undertreated conditions, consistent with an 'informative view' of advertising. Congressional leaders and consumer groups have contended that such promotion may raise prescription drug costs. Providers may be induced into prescribing more expensive (and/or possibly inappropriate) drugs in the presence of cheaper and equally effective alternatives, consistent with brand differentiation and a 'persuasive view' of advertising.

Growth in prescription drug spending is broadly driven by increases in utilization and price, and shifts in the composition of drugs being used, all of which may be impacted by marketing. A comprehensive assessment regarding the welfare effects of pharmaceutical advertising and promotion requires information on three broad but related issues: (1) effects on primary industry-wide versus selective brand-specific demand; (2) effects on price; and (3) effects on competition. The next section briefly discusses the historical background on pharmaceutical promotion followed by a conceptual framework of advertising to help guide welfare implications, before turning to the empirical evidence with respect to each of the three issues noted above.

Background

The 1962 Kefauver–Harris Amendments to the 1938 Federal Food, Drug and Cosmetic Act shifted jurisdiction on regulating drug promotion from the Federal Trade Commission to the Food and Drug Administration (FDA) and outlined the

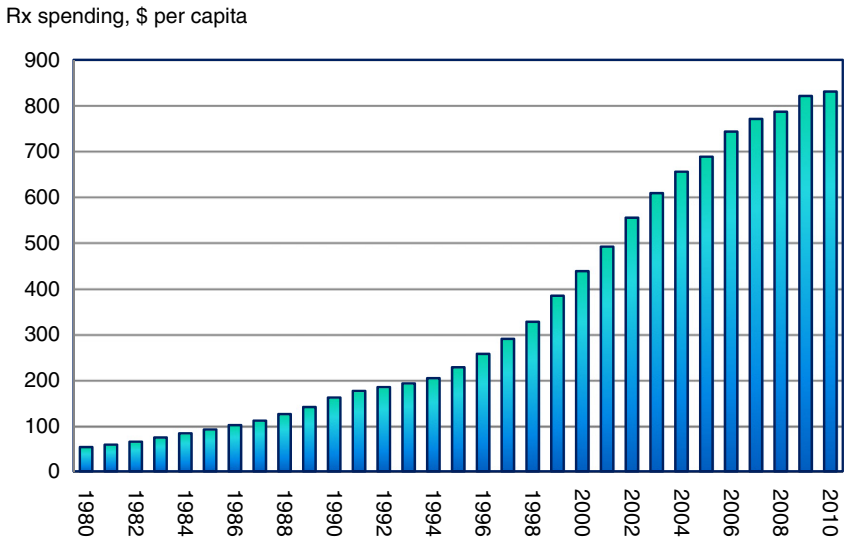


Figure 1 Prescription drug spending in the US. Data from Centers for Medicare and Medicaid Services (CMS).

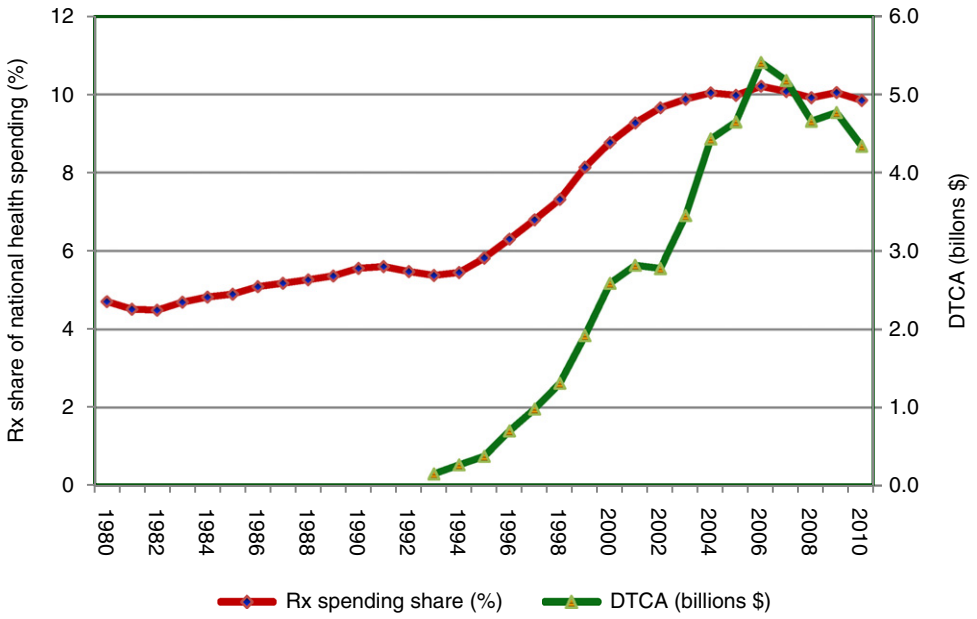


Figure 2 Rx spending share of national health expenditures and direct-to-consumer advertising. Data from CMS, Dave, D. and Saffer, H. (2012). Impact of direct-to-consumer advertising on pharmaceutical prices and demand. *Southern Economic Journal* **79**(1), 97–126; Frank, R. G., Berndt, E. R., Donohue, J. M., Epstein, A. and Rosenthal, M. (2002). Trends in direct-to-consumer advertising of prescription drugs. Kaiser Family Foundation. Available at: http://www.kff.org/rxdrugs/loader.cfmurl=/commonspot/security/getfile.cfm&PageID=14881 (accessed 26.08.09); and Bulik, B. S. (2011). Pharmaceutical marketing. *Ad age insights white paper*, Advertising Age, Kantar Media October 17.

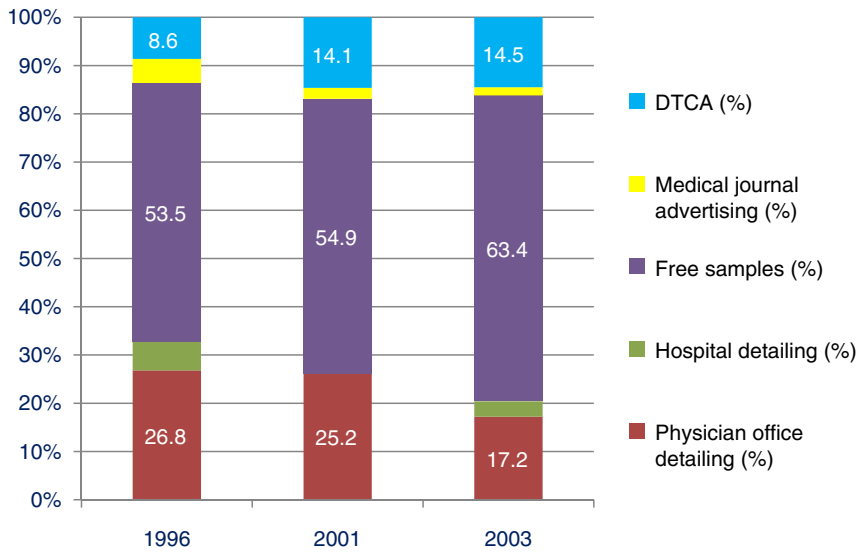


Figure 3 Components of pharmaceutical promotion. Sampling is valued based on the average wholesale price (AWP), a manufacturer's list price, and thus exceeds value based on production costs. Data from Donohue, J. M., Cevasco, M. and Rosenthal, M. B. (2007). A decade of direct-to-consumer advertising of prescription drugs. *The New England Journal of Medicine* **357**(7), 673–668; and authors' calculations from data used in Dave, D. and Saffer, H. (2012). Impact of direct-to-consumer advertising on pharmaceutical prices and demand. *Southern Economic Journal* **79**(1), 97–126.

basic requirements for acceptable prescription drug marketing (see Berndt, 2006; Dave and Saffer, 2012). Prescription drug promotional materials cannot be false or misleading, must provide 'fair balance' coverage of risks and benefits of using the drug, must provide a 'brief summary' of contraindications and effectiveness, and must also meet specific guidelines for readability and size of print. For a number of years, the FDA interpreted the 'brief summary' provision as requiring the advertiser to provide the detailed information contained in the drug's FDA-approved product labeling, thereby confining consumer-directed advertising to newspapers and magazines. There were two conditions under which firms could bypass the

'brief summary' provision: (1) if the advertising were 'help-seeking' and mentioned only disease symptoms and not any drug name, or (2) if the advertisement is a 'reminder' and mentions the drug name or its dosage without specifying what the drug is intended to treat.

Expansion of advertising into broadcast media was precipitated by the FDA's clarification of its regulation of consumer-directed advertising, particularly for broadcast advertisements. After a test period and request for public comment starting in 1995, the FDA approved the broadcast DTCA draft guidance in August 1997, eliminating the requirement that advertisements present the entire 'brief summary' taken from the product label insert. In August of 1999, the FDA further clarified the risk information requirements. Advertisements needed only to include 'major statements' of the risks and benefits of the drug, along with directions to information sources in addition to a physician, such as a toll-free phone number, a website, or a print advertisement. This shift removed a major barrier that had initially made television and radio advertising infeasible and had initially relegated consumer-directed advertising to print media only.

Between 1996 and 2000, DTCA was the fastest growing component of pharmaceutical promotion, growing at an average annual rate of 33% for gastrointestinal, cholesterol, insomnia, and antiarthritic/analgesic drugs. In comparison, detailing and sampling grew at annual rates of 12–13%, whereas professional journal advertising remained virtually unchanged (Dave and Saffer, 2012). Though the FDA's shift in guidelines specifically applied to broadcast advertising, there was also an increase in nonbroadcast advertising starting in 2000. This may be indirectly related to the FDA's new guidelines which required only 'major statements' of the drug's risks and benefits along with directions to alternate sources for more complete information. The feasibility of using television and radio advertisements may have raised the marginal product of other nonbroadcast forms. Indeed, broadcast advertisements often direct consumers to concurrent advertisements in magazines or newspapers for further information on the drug's usage and side effects.

Drugs intended to treat chronic conditions such as cardiovascular, mental health, respiratory, and erectile dysfunction conditions tend to be among the most heavily advertised. For instance, the top 5 advertised drugs in 2010 were Lipitor (indicated for high cholesterol), Cialis (erectile dysfunction), Cymbalta (mental health), Advair (asthma), and Abilify (mental health) (Bulik, 2011). The top 25 advertised drugs in 2010 (noted in Bulik, 2011) accounted for about $2.8 billion in consumer advertising expenditures or approximately two-thirds of total DTCA – suggesting a highly skewed distribution in consumer ad spending.

Recent years have witnessed a downturn in DTCA (see **Figure 2**), and total pharmaceutical sales force in the US has been cut by approximately 30% from its peak. These cuts partly reflect fewer drug launches compared to the late 1990s, and an increasing share of new drugs that are targeted at specialist physicians (for instance, cancer and orphan drugs developed specifically to treat rare conditions). Optimal promotion of such drugs may not include DTCA and, by definition, needs fewer sales representatives than the major primary-care drugs of the 1990s.

Conceptual Framework

This section draws on Bagwell (2007), which provides a comprehensive review of the economics of advertising. It is often presumed that the average consumer is responsive to advertising and promotion. However, one of the key questions in markets for healthcare inputs is whether advertising raises 'selective' or brand-specific demand versus 'primary' or industry-wide demand. The answer to this question has normative implications and relevance for public health. For instance, is advertising by the pharmaceutical industry combative and solely reflective of a market-share transfer or does it also convey information and lead to an overall expansion of the market? As a starting point, it is helpful to draw upon three principal views that have emerged with respect to why consumers may respond to advertising: (1) persuasive, (2) informative, and (3) complementary.

For firms operating under a hybrid market structure such as monopolistic competition, advertising can help them to differentiate their products and alter consumers' tastes and preferences. Under this 'persuasion' hypothesis, brand-level demand would not only shift outward in response but also become relatively less elastic, possibly leading to higher prices. Advertising-induced product differentiation and creation of brand capital may deter entry and enhance the monopolistic power of incumbent firms, especially if these established firms also enjoy scale economies in advertising and production. Thus, advertising can have anticompetitive effects under the persuasion view.

Consumers may also respond to the potential information content communicated through advertising. For instance, in markets characterized by imperfect information, advertising can effectively reduce search costs by conveying direct or indirect information regarding the existence, quality, price, and other attributes of products. With respect to pharmaceuticals, for instance, consumer advertisements may inform individuals of treatment options that they did not know existed, help them to diagnose their symptoms and seek out medical care, and remind patients to take their medications as prescribed. Similarly, detailing may provide valuable information to physicians concerning the drug's indications and contraindications allowing them to make better-informed choices. In such markets, advertising emerges as an endogenous response and solution to the information asymmetry.

There is also a distinction between search goods, wherein the consumer can determine quality before purchase though perhaps after incurring some search costs, and experience goods, wherein the consumer can assess quality only after consumption. Advertising intensity is predicted to be higher for experience goods because it can signal product quality and address the informational imbalance. The information content can also enhance the match between products and buyers in markets where consumers have heterogeneous valuations.

In contrast to the persuasive view, advertising plays a more constructive role under the informative view, and may also have pro-competitive effects. The firm's demand becomes relatively more elastic and price dispersion in the market is reduced. Advertising can thus promote competition among incumbent firms and facilitate the entry of new firms as well as new products.

The third view of advertising provides a framework under which advertising is complementary to the advertised product, and also bridges back to the informative view. For instance, if advertising enables consumers to produce information at lower cost, then consumers can more efficiently convert market goods into valued final commodities. Under this framework, a higher level of Rx drugs advertising can raise demand because the consumer now believes that he/she can obtain a greater output of the final commodity (health) from a given input of the advertised good. And, even if advertising is uninformative, it may still play a constructive role because consumers may value it directly.

The upshot of this discussion is that no single view of advertising is applicable in every setting. Furthermore, from a public health standpoint, the debate centers on whether advertising reflects a brand-switching process or a market expansion process, especially in relation to the market for health inputs. Although this is not to suggest that all brand-switching advertisements are socially wasteful (because some brand-switching may represent a better match of product attributes and consumer demand) and all market expanding advertisements are good (especially because advertisements that expand the market for unhealthy inputs may have adverse internalities as well as externalities), this dichotomy presents a useful starting point to frame some of the effects of advertising. Because advertising can affect both selective (brand-centric) as well as primary (market) demand under all three views, the question cannot be resolved based on theory alone and empirical evidence needs to bear upon the specific demand effects of advertising.

With that said, markets for over-the-counter (OTC) and prescription (Rx) medications have some predominant experience attributes. Thus, advertising intensity for the pharmaceutical industry (approximately 20% of sales) tends to be higher relative to the average industry (4–5%). These views of advertising also highlight potential effects on price, which depend on the extent to which advertising expenditures raise operating costs, affect price elasticity of demand, and allow firms to take advantage of scale economies. The concentration effects of advertising – that is, whether it facilitates entry or whether it augments the monopolistic power of established firms – depends on whether advertising is purely persuasive in nature and leads to spurious brand differentiation or whether it redresses imperfect information and makes demand more elastic.

It should also be noted that these different views of advertising may fit different forms of promotion, and are not necessarily mutually exclusive. For instance, detailing plays a role in educating providers about newer drugs and may have information value early in a product's life cycle, whereas later in the life cycle its role is predominantly persuasive, chiefly relegated to delivering samples and reminders. DTCA and detailing, by differentially targeting consumers versus providers, may also inherently play different roles in affecting primary versus selective demand. Thus, there may be a great deal of heterogeneity with respect to how consumer- and physician-directed promotion affects demand, with possible interactions with each other as well as with competition and drug characteristics. Because DTCA (and to some extent detailing) can potentially increase sales without the company having to offer a lower price or superior quality in trying to get their drug onto a preferential position with the insurer, DTCA may have the ability to undermine the insurer's formulary (Wosińska, 2002). Thus, interactions between DTCA effects and the drug's formulary position as well as between DTCA and price are also possible.

Empirical Evidence

Econometric studies have estimated the effects of DTCA and DTPP on pharmaceutical sales, patient adherence, demand for primary care, and, in a few instances, on pharmaceutical prices. Most studies have estimated an 'average' response to promotion, and very few studies have considered heterogeneity in the effects with respect to formulary placement, drug characteristics, or advertising medium. Estimating causal effects of promotion on sales and price is further complicated by potential bias due to structural endogeneity or reverse causality; promotion may affect demand, but promotional spending may also be a function of revenues.

In addition, there is potential bias from statistical endogeneity or nonrandom selection; observed and unobserved heterogeneity across Rx drugs may be driving promotion as well as sales, and prices. In the multiperiod optimization framework considered by Bhattacharya and Vogt (2003), the dynamic profit-maximizing strategy for a firm is to initially employ a relatively high level of promotion and set a relatively low price to increase current demand by raising consumers' and physicians' stock of knowledge regarding the drug. As knowledge is costly to acquire, physicians' prescribing patterns tend to be sticky and consumer use may also be sticky especially for chronic conditions. Thus, in addition to sales, price, and promotion affecting each other, they are also partly governed by the drug's life cycle and by other drug-specific unobservables, including formulary placement and implied consumer cost-sharing.

Alluding to such potential selection effects, Iizuka (2004) studies 169 brand-name drugs over 1996–99, and finds evidence that higher quality drugs (as measured by the FDA's priority rating) are more likely to engage in DTCA as are drugs with a larger potential market size (measured by the prevalence rate of certain chronic conditions from the National Health Interview Surveys). DTCA spending tends to be lower when there is a generic competitor on the market. Thus, advertised drugs are systematically different from non-advertised ones, and these differences may confound the causal relationship between promotion and demand. The more sophisticated of the studies address these concerns through instrumental variables and fixed effects.

Demand Effects

Market expansion versus product-level effects of DTCA

Rosenthal *et al.* (2003) study brands in five therapeutic classes using an aggregated US monthly time series from August 1996 to December 1999. Their results indicate that the primary impact of DTCA lies in expanding the total market size rather than affecting product market share. Specifically, at the level of

the therapeutic class, DTCA spending positively impacts sales with an estimated elasticity of 0.10. Although they do not report any significant effects of brand-specific DTCA (or detailing) on brand-specific market shares, they do caution that it may be "premature to conclude that DTCA only affects class level sales, and not individual product sales." The models estimate only a contemporaneous effect owing to the short span of the time series. This is likely to be a lower-bound estimate because advertising effects in the prescription drug market may be especially prolonged due to the multistage process, with time lags between ad exposure, scheduling a physician visit, and obtaining and filling the prescription. Wosińska (2002) shows the importance of the drug formulary in driving DTCA effects (with advertising having a greater effect on demand for drugs that have a preferential position on the insurer's formulary), and notes that the inability to differentiate across the formulary status may also explain why Rosenthal *et al.* (2003) do not find a market share effect of DTCA.

The specifications in Rosenthal *et al.* (2003) include class fixed effects, but do not control for unobserved heterogeneity across drugs within a class through drug-specific fixed effects. The study uses time to patent expiration, an indicator for 1997 (reflecting the FDA's change in policy), and interpolated monthly television advertising costs per minute as IVs that can plausibly be excluded from the sales equation. Some studies, however, have shown that the drug's life cycle is an important determinant of sales, which suggests that the product's life cycle may not be an appropriate instrument for advertising and promotion (Bhattacharya and Vogt, 2003; Dave and Saffer, 2012). Nevertheless, Rosenthal *et al.* (2003) provide one of the earliest and seminal analyses of DTCA following its resurgence in the late 1990s, and several subsequent studies confirm their market-expansion effect of DTCA.

Iizuka and Jin (2005) merge individual-level data from the National Ambulatory Medical Care Surveys over 1995–2000 with monthly DTCA. Similar to Rosenthal *et al.* (2003), their effect is identified from variation in DTCA across drugs within a given class (and over time). They also utilize an IV procedure, employing the same drug company's DTCA expenditures in other unrelated drug classes as an instrument for DTCA in a particular drug class. Consistent with a market-expansion effect, they find that each $28 increase in DTCA (which includes current advertising and a depreciated sum of past advertising) leads to an additional physician visit within a year where an Rx drug from the class is prescribed.

The market expansion effect suggests that some consumers, whose medical conditions were previously undiagnosed or undertreated, may benefit from the information provided by DTCA. Consumers, becoming aware of new treatments, may be incentivized to seek out physician care. However, the market expansion effect may also partly reflect inappropriate care, if advertising increases the use of drugs with uncertain safety profiles. Pointing to perhaps such an increase in misuse, David *et al.* (2010) find that higher levels of DTCA lead to increased reporting of adverse medical events for drugs related to certain conditions such as arthritis and depression, whereas detailing reduces the adverse event rate for high cholesterol and allergy drugs. They conclude that the effect of promotion and advertising in improving communication between patients and physicians may be welfare-enhancing if physicians can identify who is the best match for treatment. This is feasible in the case of cholesterol and allergy medications by the existence of simple diagnostic tests. In cases where there is greater uncertainty regarding diagnosis or acceptable standards for care, advertising and promotion may hinder the role of the physician as a mediator between consumer-directed promotion and proper use.

In addition to a market expansion effect, few studies do find evidence of some DTCA-induced brand-switching. Wosińska (2002), in a study of prescription claims for cholesterol drugs for Blue Shield of California over the period 1996–99, finds that current DTCA raises market share though this effect is limited to drugs that have preferential status on the insurer's formulary. Thus, it is possible that physicians suggest and prescribe advertised formulary drugs when patients inquire about advertisements for drugs that are not on the formulary. Dave and Saffer (2012) utilize monthly data on all prescription drugs in four major therapeutic classes from 1994 to 2005, exploiting the period enveloping the FDA's shift in regulations as a natural experiment and exogenous shock to DTCA. Similar to Iizuka and Jin (2005), they construct a stock of depreciated DTCA spending over the past year. Models account for unobserved heterogeneity across drugs and potential selection into DTCA by including drug-level fixed effects and various time-varying confounders including physician detailing and sampling, the drug's life cycle, competitors' advertising, generic competition, and FDA approval of new indications for the drug and labeling/marketing warnings. This study underscores the point that it is important to separately assess broadcast and nonbroadcast DTCA due to differences in their content, growth trends (because the FDA's policy change specifically affected broadcast DTCA), and their marginal impacts. They find that broadcast DTCA does significantly impact own-sales and market share with a relatively small elasticity of 0.10, though this response is higher relative to nonbroadcast DTCA. There is also some evidence that class-level DTCA may raise sales for the nonadvertised drugs. Assuming that physicians are prescribing an equally effective drug, this may be a spillover benefit of DTCA in some cases because nonadvertised drugs tend to be older and also cost less.

Prior studies, which at times found conflicting evidence on the impact of own-DTCA on own-sales, may have been confounded by aggregating broadcast and nonbroadcast forms. In periods predating the FDA's policy shift, virtually all DTCA was relegated to nonbroadcast media, whereas starting in 1998 and 1999 advertisements in broadcast media became the primary form of DTCA. Therefore, the effect of total DTCA, being a weighted average of the effect of the two separate forms, would be expected to vary depending on the time period under study and the relative composition of total DTCA between nonbroadcast and broadcast media.

Directly bypassing the potential endogeneity of advertising, Kravitz *et al.* (2005) examine how DTCA impacts the prescribing behavior of antidepressants in a randomized control trial (RCT) setting. Standardized patients, mostly professional actors, were randomly assigned to make 298 unannounced visits to family physicians and general internists. The patients made a specific brand request (referring to a DTC advertisement), a general drug request, or no request. Physicians

prescribed antidepressants for the patients portraying general depression in 54% of the visits, including 76% of visits where the patients made a general request for a drug, 53% of visits where a specific drug was mentioned, and 31% where no drug was mentioned by the patient. Patients were prescribed Paxil in 27% of the visits where they explicitly mentioned the drug, compared to 4% where there was no request for a drug and 2% where the patients made a general request for a drug. For patients portraying adjustment disorder, where antidepressants confer little or no benefits, 37% of patients requesting Paxil received a prescription for the drug, compared to 10% of patients who made a general drug request and none for patients who did not request any drugs. This study points to the role of brand-specific DTCA in raising own-demand by leading to a prescription for that brand, as well as in raising overall demand for drugs in the therapeutic class. The authors note that DTCA "may have competing effects on quality, potentially averting underuse and promoting overuse."

Observational and survey-based studies suggest that DTCA can educate consumers about health conditions and available treatments, though it may also have the potential to be misleading or uninformative. Hollon (2005) summarizes some of the survey-based evidence and notes that between 25% and 33% of adults annually have had a discussion with their physician regarding a health issue after having seen an advertisement. Although DTCA can stimulate a new diagnosis (approximately 25% of patients with DTCA visits), potentially leading to treatment of previously undertreated conditions, almost 80% of physicians report that DTCA encourages patients to seek treatments that they may not need. Thus, DTCA can mediate the patient–physician relationship through reeducation and informative discussions as well as through a likely fulfillment of the patient's request for a specific drug.

Additional evidence on the demand effects of DTCA is provided by econometric studies that examine patient adherence. Consistent with the informative view of advertising, these studies underscore an important health-promoting benefit of DTCA in reminding patients to adhere to their drug therapy as prescribed. For instance, Calfee *et al.* (2002) utilize a national monthly time series of statin prescriptions, between 1995 and 2000, and find television advertising expenditures on statins are associated with an increased proportion of existing patients who were successfully treated (existing patients with a high-cholesterol diagnosis whose total cholesterol fell below 200 mg dl^{-1}). This effect combines both a compliance effect and also a market expansion effect as successfully treated patients spread the word about the effectiveness of statin drugs and raise demand among untreated or undertreated patients.

Effects of physician-directed promotion

In addition to consumer advertising, studies have also examined the impact of promotion aimed at healthcare providers, which historically has been the primary form of promotion used by the pharmaceutical industry. As uncertainty regarding the efficacy of the drug and its safety profile tends to be high in the early stages of the drug's life cycle, DTPP can play an informative role following a drug's launch. After some point, DTPP largely takes on a persuasive role by providing samples and reminders.

Studies have generally found that the marginal impact of detailing on market share is significantly larger relative to that for consumer-directed advertising. For instance, Dave and Saffer (2012) find significantly larger sales-DTPP elasticities (0.51 for detailing and 0.34 for sampling) compared to the sales-DTCA elasticity (0.13). Wosińska (2002) also reports that the effect of detailing on market share is approximately five times higher relative to the effect of DTCA.

Beyond estimating mean effects of DTPP, some studies further assess interactions between the various marketing elements and also consider differential effects across various market, physician, and product-level characteristics. Narayanan *et al.* (2004) utilize monthly data on three branded second-generation antihistamines (and one aggregated measure of all other first-generation and other antihistamines) spanning April 1993 through March 2002. They find that detailing primarily and positively affects brand share, whereas DTCA has a significant positive effect on both brand shares and class sales. The return on investment is much larger for detailing than for DTCA, a feature that they attribute to the fact that detailing allows for a much more targeted promotional effort relative to DTCA. They also find evidence of synergy between the two forms of promotion. For instance, a sales call to a physician's office has a higher marginal impact on brand share when combined with DTCA. Studies also find that sampling and detailing are highly complementary, and that sampling can raise the effectiveness of detailing; detailing is also found to be generally more effective when targeted towards specialists followed by primary-care physicians.

International evidence

Similar to all industrialized nations except for the US and New Zealand, Canada prohibits DTCA for prescription medications. However, approximately 30% of the television viewing of Canadians in English-speaking provinces consists of US satellite and cable TV, which carries consumer-directed Rx drug advertisements (Mintzes *et al.*, 2009). Law *et al.* (2008) study the impact of such US-based advertisements on Canadian prescribing rates for three drugs (Enbrel for rheumatoid arthritis, Nasonex for allergy symptoms, and Zelnorm for irritable bowel syndrome (IBS) in women) in English-speaking provinces relative to French-speaking Quebec. The study finds only short-lived positive effects for Zelnorm, and no significant effects for the other two drugs. Zelnorm is the only drug approved for its indication in Canada, whereas the other two drugs had competitors. IBS is also underdiagnosed and undertreated, in which case DTCA can be informative. However, Zelnorm tends to be measurably effective in only approximately 1 out of 17 patients, which may explain why the effects of DTCA are short lived. Insignificant effects of DTCA on the prescribing rates for Enbrel may be because the drug requires a specialist referral and subcutaneous injection. Thus, similar to studies from the US, the impact of DTCA on drug use appears to be variable and dependent on the characteristics of the advertised drug and the medical environment.

Two major shifts in DTCA-related administrative policy occurred in Canada. In 1996, a redefinition of the boundary between 'information dissemination' and 'advertising' by Health Canada provided tacit approval for unbranded disease-oriented 'ask your doctor about available treatments'

advertisements. In 2000, manufacturers were allowed to use 'reminder advertisements' that state a brand-name but do not mention any indications, or make any therapeutic claims. Though it is not meant to imply a causal effect of this shift in policy (because many new drugs were also launched over this period), total inflation-adjusted DTCA in Canada increased from under $1.6 million (Canadian $) in 1995 to more than $22 million in 2006 (Mintzes *et al.*, 2009). Similar to the US, consumer advertisements in Canada are highly concentrated on relatively few products for treating chronic conditions. Although there has been no rigorous study on how such advertisements have impacted demand and related outcomes to inform on welfare effects, many of the heavily advertised drugs in Canada have US 'black box' warnings and have been subject to Health Canada safety advisories (Mintzes *et al.*, 2009). Thus the safety profile of some of the highly consumer-promoted drugs is questionable.

Unlike most other developed countries, New Zealand had never enacted preemptive legislation to preclude pharmaceutical DTCA, and such advertisements are implicitly permitted under conditions set by the Medicines Act and the Medicines Regulations. As in the US, consumer-directed advertising grew tremendously in New Zealand during the late 1990s. Toop *et al.* (2003) conducted a survey of 1611 general practitioners (GP) in New Zealand, with results qualitatively similar to survey-based evidence from the US. They find that 90% had experienced DTCA-generated consultations. Furthermore, 79% reported that patients frequently inquired about DTC-advertised medications, and 44% noted that they had switched to or started treatment with medicines they felt offered little added benefit over drugs they would normally use as a result of DTCA. Approximately 12% of the respondents believed that DTCA is a useful means of educating consumers about the drugs' risks and benefits, 16% felt that DTCA helped their patients get necessary medical care at an earlier stage, and 13% reported DTCA improved adherence.

Chintagunta and Desiraju (2005) report sales elasticities with respect to detailing and price for three selective serotonin reuptake inhibitor (SSRI) antidepressants (Prozac, Zoloft, and Paxil) across five markets (US, UK, Germany, Italy, and France), based on quarterly data from 1988 to 1999. The authors estimate IV models, instrumenting price with current and lagged values of the producer price index for preparations and psychotherapeutics and cost measures from the companies' balance sheets, and instrumenting detailing with the current and lagged values of wage index. Own-detailing is found to have a significant and positive impact on own sales for all three drugs in all markets. The elasticity magnitudes are generally similar for the other four countries, ranging from 0.17 to 0.59, though several orders of magnitude higher for France (2.32–2.43). The authors note that this may be due to a large marginal benefit of SSRI-related detailing in France, because pre-SSRI antidepressants had not been actively detailed. Cross-detailing elasticities are generally negative and smaller in magnitude, which is consistent with the persuasive view of advertising; detailing primarily affects selective demand and leads to brand-switching.

Berndt *et al.* (2007) present a study of the rate at which new drugs are promoted and diffused, across three therapeutic classes (antihypertensives, antidepressants, and antiepileptics) and ten countries. They find that the largest level of detailing occurs for antihypertensives, followed by antidepressants and then by antiepileptics. Spain has the highest rate of detailing for all three classes, consistent with its high rate of utilization growth of these drugs. US detail counts per capita are close to the median in their sample. Cross-national differences further indicate that it is not necessarily the case that new drugs in the US are intensely detailed over older drugs. The authors further model the diffusion process through two separate components, including the total drug therapy days per capita and the new-drug expenditure share, and estimate this framework via the Almost Ideal Demand system for the three drug classes. They generally find insignificant or very small detailing elasticities with respect to aggregate utilization for all three classes. This is consistent with the US-based studies, which generally find that detailing impacts selective brand-specific demand rather than primary market demand. However, with respect to the new-drug expenditure shares, the results show positive and significant new-detailing elasticities. The cross effect of detailing on older drugs is negative. Although the authors caution that the promotion intensity is endogenous, their results suggest that detailing can promote the diffusion of newer drugs.

Summary: Demand effects

A number of robust empirical findings emerge from these studies on the demand effects of pharmaceutical promotion. First, both the econometric as well as survey results indicate positive demand effects of DTCA. Survey results (for instance, Hollon, 2005) indicate that physicians do consider specific drug requests initiated by the patient. Although in many cases physicians appear to fulfill these requests, in other cases they take into account acceptable standards of care in prescribing an alternative drug or not prescribing at all (Kravitz *et al.*, 2005). These results are consistent with DTCA having both primary market-expansion effects and also selective brand-specific effects.

The econometric literature is able to further pinpoint the relative strength of these two effects. Studies find consistent evidence of a significant class-level market expansion effect of DTCA. Dave and Saffer (2012), for instance, find that class-level DTCA may raise the sales for lower cost, nonadvertised drugs. Thus, DTCA may bring a patient to the doctor's office, but in some cases the doctor is prescribing a lower-cost alternative. Consumer-directed promotion raises class-level sales, by encouraging patients to seek medical help, encouraging patient–physician contact, and promoting compliance with Rx drug therapy. This is reflective of the informative view of advertising, wherein DTCA plays at least some role in educating consumers and expanding treatment among those previously under-treated. However, at the same time, there is some evidence (Kravitz *et al.*, 2005; David *et al.*, 2010) that the increase in primary demand may partially reflect overtreatment or possibly inappropriate care, especially for conditions where greater uncertainty exists regarding diagnosis and acceptable standards for care. The evidence relating to the effects of own-DTCA on the specific drug's market share is mixed, though some recent studies (Dave and Saffer, 2012; Wosińska, 2002) suggest significant but relatively small

elasticity magnitudes. These studies also point to considerable heterogeneity in these own-DTCA effects, depending on the drug's formulary position, level of DTPP, other competitors in the therapeutic class, characteristics, and the composition of its DTCA. The literature frequently views such 'business-stealing' advertising as less benign, though to the extent that it results in a better match between the consumer and the product it may also confer some welfare benefits (Berndt, 2006). However, because higher-advertised drugs tend to cost more on average and to the extent that such advertising results in a higher-priced product capturing market share, it may raise healthcare costs and confer negative spillover effects. Though, here too, the effect is ambiguous if price and quality are positively correlated. Overall, stronger market expansion effects combined with weaker and mixed evidence on selective brand-specific demand effects of DTCA suggest that consumer-directed advertising is perhaps more reflective of the informative view of advertising over the persuasive view.

This literature also consistently finds that the effects of physician-directed promotion, such as detailing and sampling, on own-demand are significantly larger relative to consumer-directed promotion. This is consistent with detailing primarily driving market share, as against DTCA driving class expansion. Detailing, sampling, and medical journal advertising can shift treatment away from nondrug therapy toward the promoted drug, and can raise the number of patients treated with drug therapy, but cannot induce untreated consumers to visit the doctor. DTCA, however, can stimulate contact between untreated patients and physicians (market expansion), and can also perhaps impact prescription choice (brand demand). Thus, DTPP is relatively more reflective of the persuasive view of advertising, at least during the later stages of the drug's life cycle. During the early stages when there is greater uncertainty regarding a newer drug's attributes, DTPP may bridge an informational gap and educate physicians regarding the drug's availability, effectiveness, and safety profile.

Price effects

Advertising by pharmaceutical manufacturers does not contain price information. Because patients only pay the pharmacy their copayment, which differs across health plans, pharmacies also have no incentives to advertise prices for Rx drugs (though they do advertise price information regarding OTC drugs and sometimes their waiving of patient copays for generic drugs). In the context of manufacturer non-price advertising, promotion may nevertheless affect price through various processes. First, the increase in operating costs due to higher promotional spending may be shifted to consumers in the form of higher prices. Second, promotion may increase demand and/or reduce the absolute magnitude of the demand-price elasticity, in turn raising price. This is consistent with the persuasive view of advertising, wherein advertising-induced product differentiation and creation of brand capital may enhance firms' monopolistic power. For instance, in an oligopolistic situation advertising will raise prices if it raises product demand, makes demand less elastic, and does not substantially lower marginal costs. Detailing, DTCA, and price are also complementary strategies for the firm.

In contrast to the persuasive view, manufacturers' advertising targeted toward consumers (and physicians) may lower price if such promotion reduces search costs for consumers (and physicians) by communicating direct or indirect information regarding the existence, quality, price and other product attributes, and subsequently makes demand more elastic (Encinosa *et al.*, 2011). With respect to the pharmaceutical marketplace, the 'consumer' can also be interpreted as the pharmacy benefit managers (PBM) who negotiate discounts and rebates with drug manufacturers on behalf of the insurers. Steiner (1973) presents a dual-stage model wherein consumer advertising can affect both the manufacturer's and the retailer's margin. Manufacturers' consumer-directed advertising provides information and raises consumer demand for the brand, which may facilitate competition between retailers on the advertised brand and subsequently lower retail margins although raising manufacturer prices. It should be noted, however, that lower retail margins combined with higher manufacturer prices does not necessarily imply lower retail prices. If the increase in manufacturer prices is not large enough, and consumer price sensitivity increases sufficiently, then retail prices may fall.

The empirical evidence on the effects of advertising and promotion on price is more limited relative to the evidence on demand. This paucity of research partly derives from the difficulty in obtaining measures of net Rx drug prices due to the presence of third-party payers and unobserved rebates from drug manufacturers to third-party payers.

Kopp and Sheffet (1997) provide an early study of the effects of DTCA on retail gross margins, testing the dual-stage theory of Steiner (1973). They construct the brand gross margin ratio, which measures the percentage by which a particular drug's retail margin is higher or lower relative to the class average, based on the average wholesale price – AWP (prices paid by pharmacies) and a pharmacy survey of retail prices. They then compare this brand gross margin ratio for 13 DTC-advertised brands with the remainder of 120 top-selling drugs that were not DTC-advertised over 1986–92 (control group). Because non-DTC advertised drugs are systematically different from those that are advertised, trends in the control group may not be a valid counterfactual. DTCA during this period, which predated the FDA's policy shift, was also relatively small and confined to print media. In support of Steiner's model, the study finds that retail margins for the advertised drugs fell relative to the nonadvertised drugs.

Rizzo (1999) studies 46 antihypertensive drugs and, based on drug-specific fixed effects models, finds that increased current and past detailing efforts reduce the price elasticity. The price measure reflects the wholesale price of the drug to drug stores and hospitals. The reduction in the price elasticity may consequently result in higher prices, though Rizzo does not examine the direct link between detailing and price. He concludes that pharmaceutical promotion differentiates products, increases brand loyalty, and inhibits price competition in the pharmaceutical industry. The study is based on pooled annual data from 1988 to 1993, which predate the DTCA policy shift, and only considers promotion to physicians. Rizzo's results contrast with Narayanan *et al.* (2004) who find a negative interaction between detailing and price suggesting that detailing may raise the price elasticity, albeit for a different sample of drugs (antihistamines) and a more recent time period (1993–2002).

Law *et al.* (2009) examine quarterly level pharmacy data for Plavix (an antiplatelet drug used to prevent stroke and heart attack in at-risk patients) from 27 Medicaid programs over 1999–2005. Plavix initiated DTCA in 2001. Based on an interrupted time-series design, the study finds that, although there was no change in the preexisting trend in demand, there was a sustained increase in total Medicaid-reimbursed pharmacy cost per unit of $0.40 (11.8%) after the expansion in DTCA. They note that the extra reimbursement from Medicaid likely reflects an increase in the manufacturer's price.

Dave and Saffer (2012), discussed earlier, utilizing a larger sample of all Rx drugs in four therapeutic classes, also find that DTCA raises the AWP, though the estimated elasticity is of a relatively small magnitude (0.04). Consistent with this positive impact on price, they also find suggestive evidence that the consumer price response became relatively more inelastic during the period when DTCA was expanding. Simulations indicate that expansions in broadcast DTCA over 1994–2005 accounted for 19% of the overall growth in prescription drug spending (assuming that movements in the AWP and retail prices are proportional), with less than a third of this impact being driven by higher prices and the remainder due to higher demand.

Encinosa *et al.* (2011) examine 17 million claims for 177 Rx drugs in 19 therapeutic classes, between 2001 and 2002. They study both the AWP and the transacted retail price, which is the total reimbursement that the pharmacy receives from the insurers plus any patient copayment. The authors estimate drug-level fixed effects models and find that an increase in DTCA (from 0 to the sample mean) reduced average transacted prices by 1.8%, decreased price dispersion by 3.7%, and reduced pharmacy profit margins by 1.5% (consistent with Kopp and Sheffet, 1997). The reduced price dispersion and lower retail profit margins are interpreted as a sign of increased price shopping by the insurers' PBM. However, similar to Dave and Saffer (2012), they also find positive effects on the AWP though this effect becomes insignificant once they control for market fixed effects. The study does not assess effects of DTPP.

Limitations

One challenge faced by all of these empirical studies concerns the simultaneity between advertising and pricing decisions. For instance, as noted earlier, in the model developed by Bhattacharya and Vogt (2003), price and promotion are jointly determined over the drug's life cycle. This trajectory of higher prices and lower advertising over the drug's life cycle is also consistent with the Dorfman–Steiner (Dorfman and Steiner, 1954) condition for optimal advertising:

$$\text{Advertising/Sales} = \varepsilon_{QA}/\varepsilon_{QP}$$

The optimal advertising-to-sales ratio is a positive function of the elasticity of sales with respect to advertising (ε_{QA}) and is inversely related to the elasticity of sales with respect to price (ε_{QP}), both expressed in absolute magnitudes. Thus, the decline in advertising over the drug's life cycle is consistent with an age-related decline in the sales-advertising elasticity (Berndt, 2006). It is also consistent with an increase in the price elasticity as the drug ages and newer drugs enter. A positive association between advertising and price inelasticity may thus reflect causality in both directions – if persuasive in nature, advertising may make demand more inelastic, but *ceteris paribus* more inelastic demand also leads to a higher optimal level of advertising. Although many of these studies attempt to address this simultaneity through additional controls for the drug's life cycle, drug-level fixed effects, and exploiting the exogenous shift in FDA regulations, the results should be interpreted in the context of the limitations noted.

Another limitation relates to the measurement of drug prices. None of the price measures include rebates negotiated between PBM and other payers (for instance, state Medicaid agencies) from the drug manufacturers, because information on these rebate arrangements is confidential. Sources estimate these rebates at between 2% and 35% of drug sale prices (Dave and Saffer, 2012). This rebate does not affect the price paid by a retail pharmacy to the wholesaler, or the price paid by the PBM to the pharmacy. It is a separate transaction between the PBM and the manufacturer, and affects the net transaction price. Manufacturers of brand-name drugs that treat conditions for which alternative drugs are available have a strong incentive to grant discounts to the PBM in return for preferential positioning of their drug on the formulary. If generic equivalents are available, the manufacturer may also grant a discount to make the price of its brand-name product more competitive. Movements in the list AWP or the observed transacted retail price therefore may not be reflective of movements in the net transaction price. The growth in restrictive formularies over the period when DTCA was expanding suggests that the size of the negotiated rebates may also have expanded, leading to a decrease in the net transaction price. However, the key issue concerns the extent to which the size of the rebate is correlated with DTCA. If DTCA is targeted to raise consumer demand, provide information, and push for better positioning on the formulary, then DTCA may be associated with higher rebates leading to an overestimate of the positive effects of DTCA on the net transaction price. In this case, given that the estimated price elasticity is low, it is possible that transaction prices net of rebates may have remained unchanged or even declined. If DTCA raises market power, and reduces the rebates to PBMs, then the estimated elasticity of the net transaction price with respect to DTCA is biased downward.

Summary: Price effects

These limitations notwithstanding, the above studies do point to a few relatively consistent findings. First, DTCA may have a positive though small effect on AWP (Dave and Saffer, 2012; Encinosa *et al.*, 2011; Law *et al.*, 2009), consistent with DTCA-induced market power. Second, there is suggestive evidence that DTCA may have also reduced pharmacy retail margins (Encinosa *et al.*, 2011; Kopp and Sheffet, 1997). Both of these findings are also consistent with Steiner's (1973) dual-stage model, wherein manufacturer advertising provides information, helps differentiate brands, raises consumer demand, facilitates price-based competition (and price negotiations in the case of Rx drugs), and subsequently lowers retail margins although raising manufacturer prices. Evidence is weakly indicative that certain forms of promotion may lower the price elasticity (Dave and Saffer, 2012; Rizzo, 1999). Even then there is no strong evidence that promotion causes substantially higher retail-level prices.

Effects on entry and innovation

The above studies suggest that consumer-directed pharmaceutical promotion has information content, conveying potential treatment options to consumers and expanding the market for drug therapy, at least for certain conditions. To some extent, this also applies to DTPP, which can be educational during the early phases of the drug's life cycle. As consumers (and insurer representatives – PBMs) receive low-cost (relative to incurring search costs) information, demand can become relatively more elastic and price dispersion in the market is reduced. Under this informative view, advertising can thus promote competition among incumbent firms and facilitate the entry of new firms and new products. At the same time, some studies also point to persuasive effects of DTCA and DTPP. Such promotion-induced product differentiation may have anticompetitive effects by enhancing the monopolistic power of firms and deterring entry. Anticompetitive effects on generics, however, at least in the US market, may be muted because pharmacists can substitute generics even if the physician writes a script for the branded drug. For this reason, promotion slows as patent expiry approaches and ceases almost entirely once generics enter.

Some evidence can be gleaned from direct studies of the effects on the entry of generic and branded substitutes. Scott Morton (2000), for instance, investigates the role of prepatent expiry brand-level DTPP in impacting postexpiry generic entry. Promotion is likely to be endogenous, reflecting the same market conditions that affect entry; strong markets attract both more advertising and more entrants. Thus, the study estimates IV-based models, using as instruments the drug's life cycle, an indicator for whether the firm has other forms of the same drug still under patent, and the number of physicians expected to prescribe the drug. These IV estimates do not show any significant or substantial barrier to entry by generic firms associated with brand advertising.

Kwong and Norton (2007) study the lagged effects of promotion on pharmaceutical innovation in eight drug markets, as measured by the total number of investigational products entering clinical development in a given market, over 1995–2001. They find that detailing may have a significant positive effect on the number of new products entering into clinical development, with markets for chronic disease with high levels of detailing being more attractive to pharmaceutical firms. Other types of advertising were not found to impact product entry, however. They note that this may be due to the unique role of detailing in affecting brand-specific demand and enhancing product differentiation. As the study is unable to implement IV-based corrections or control for drug-class fixed effects due to the limited sample size, the authors acknowledge that the results may be subject to endogeneity and omitted variables bias.

Conclusion

Pharmaceutical promotion, and in particular DTCA, has emerged as a marketing force in the US healthcare system. Although the debate surrounding such promotion is unlikely to be resolved anytime soon, pharmaceutical promotion should be evaluated both in terms of its costs as well as its benefits. Welfare implications can be indirectly gleaned from the extent to which such promotion affects demand, competition, and prices.

Several studies have suggested that consumer-directed advertising provides information content regarding treatment options, induces physician contact, and expands treatment, at least for certain undertreated or chronic conditions such as depression and high cholesterol. The benefits of DTCA derive from improved health due to increases in the initiation of drug therapy and adherence with drug therapy. Detecting and treating health conditions at an earlier stage, through primary care, may also be cost-effective relative to treatment at a later stage through acute care. Many health conditions are especially undertreated for disadvantaged groups; for instance, Blacks are significantly less likely to receive Rx drug treatment for high cholesterol. Thus, if DTC advertisements provide useful information and induce patients to visit their doctors, then their potential educational benefits may help reduce health-related disparities.

There is limited direct evidence on the competitive effects of pharmaceutical promotion. Though, the few studies that have been conducted seem to indicate that, if anything, promotion may be pro-competitive. Promotion aimed at providers can facilitate entry of other products in the drug class and also positively impact the number of new products entering into clinical development. There do not appear to be strong deterrent effects on generic entry. These results are consistent with the informative-view of advertising, and studies that find advertising-induced market expansion effects generally interpret these findings as welfare improving.

One of the costs of DTCA and DTPP includes potentially higher drug prices and increased use of more expensive drugs in place of equally effective lower-priced drugs. Although there is no direct study of this latter effect, Kravitz *et al.* (2005), in a randomized setting, find that for actors portraying adjustment disorder where antidepressants confer little or no benefits, 37% of actor-patients requesting Paxil received a prescription for the drug, compared to 10% of those who made a general drug request and none for those who made no request. Higher drug and healthcare expenditures can raise insurance premiums, increasing taxpayer and individual costs, and lead to a larger prevalence of the uninsured. Cost-ineffective treatments also impose opportunity costs for public and private resources. Here too, the evidence is limited and hampered by measurement error in drug prices. However, the few studies in this area suggest that promotion may have a small positive effect on the AWP and reduce retail pharmacy margins. There is no strong evidence that DTCA or other promotion substantially raises retail-level drug prices.

Evidence from physician surveys and a randomized control study (Kravitz *et al.*, 2005) does suggest that there may be some DTCA-induced overuse and overtreatment, particularly in cases where there are no structured clinical guidelines for treatment. That physicians prescribe a certain drug in response to patients' request suggests a persuasive brand-switching response to DTCA in addition to a market-expansion component. Some econometric studies confirm that DTCA affects selective demand, which is often viewed as less benign relative to promotion that affects primary demand. However, these brand-specific effects generally tend to be small in magnitude.

In contrast, both the US-based and international studies consistently find that the brand-switching effects are far stronger for physician-aimed promotion.

Market expansion and shifting brands for nontherapeutic reasons also raise the concern of a suboptimal patient–drug match for marginal patients, carrying the risk that the drug is prescribed inappropriately and leading to a worsening of the drug's average safety profile. As shown in David *et al.* (2010), increased levels of DTCA are associated with increased reporting of adverse medical events for certain conditions. Because newer drugs generally tend to be more heavily promoted, especially with consumer-directed advertisements, a popular proposal among critics of DTCA in Congress is to impose a moratorium on such advertisements during the first 2 years of a drug's launch. In response, a group of leading pharmaceutical firms (Merck, Schering-Plough, Johnson & Johnson, and Pfizer) have agreed to a voluntary 6-month moratorium on DTCA for new drugs. This would give the FDA, providers, and patients time to learn about new safety issues for new entrants. The benefits of such a proposal also need to be balanced against the need to convey information regarding new drug therapies, which may be especially important in the early stages of a drug's launch. Optimal use of DTCA may therefore require further structured guidelines.

In summary, pharmaceutical promotion has effects which can be strongly health-promoting and welfare-enhancing, but may also have adverse effects through potential overtreatment, cost-ineffective substitutions, and potential misuse. In cases where physicians can effectively perform their role as mediators, the concern about promotion-induced inappropriate use is mitigated. However, for conditions where the diagnosis or risks may be difficult to assess, there may be a need for greater oversight and investment in postmarketing surveillance by pharmaceutical firms.

See also: Advertising as a Determinant of Health in the USA. Markets in Health Care

References

Bagwell, K. (2007). The economic analysis of advertising. In Armstrong, M. and Porter, R. (eds.) *Handbook of industrial organization,* pp. 1701–1844, vol. III. Amsterdam: North-Holland.

Berndt, E. R. (2006). The United States experience with direct-to-consumer advertising of prescription drugs: What have we learned? In Sloan, F. A. and Hsieh, C. R. (eds.) *Promoting and coping with pharmaceutical innovation: An international perspective*, pp. 221–253. New York: Cambridge University Press.

Berndt, E., Danzon, P. M. and Kruse, G. B. (2007). Dynamic competition in pharmaceuticals: Cross-national evidence from drug diffusion. *Managerial and Decision Economics* **28**, 231–250.

Bhattacharya, J. and Vogt, G. (2003). A simple model of pharmaceutical price dynamics. *The Journal of Law and Economics* **46**(2), 599–626.

Bulik, B. S. (2011). Pharmaceutical marketing. *Ad age insights white paper*, London, UK: Advertising Age, Kantar Media October 17.

Calfee, J. E., Winston, C. and Stempski, R. (2002). Direct-to-Consumer advertising and the demand for cholesterol reducing drugs. *Journal of Law and Economics* **45**, 673–690.

Chintagunta, P. K. and Desiraju, R. (2005). Strategic pricing and detailing behavior in international markets. *Marketing Science* **24**(1), 67–80.

Dave, D. and Saffer, H. (2012). Impact of direct-to-consumer advertising on pharmaceutical prices and demand. *Southern Economic Journal* **79**(1), 97–126.

David, G., Markowitz, S. and Richards-Shubik, S. (2010). The effects of pharmaceutical marketing and promotion on adverse drug events and regulation. *American Economic Journal: Economic Policy* **2**(4), 1–25.

Donohue, J. M., Cevasco, M. and Rosenthal, M. B. (2007). A decade of direct-to-consumer advertising of prescription drugs. *The New England Journal of Medicine* **357**(7), 673–681.

Dorfman, R. and Steiner, P. O. (1954). Optimal advertising and optimal quality. *American Economic Review* **44**, 826–836.

Encinosa, W., Meyerhoefer, C., Zuvekas, S. and Du, D. (2011). The inverse relationship between direct-to-consumer advertising and retail drug prices. Working Paper, January 5.

Frank, R. G., Berndt, E. R., Donohue, J. M., Epstein, A. and Rosenthal, M. (2002). Trends in direct-to-consumer advertising of prescription drugs. Kaiser Family Foundation. Available at: http://www.kff.org/rxdrugs/loader.cfmurl=/commonspot/security/getfile.cfm&PageID=14881 (accessed 26.08.09).

Hollon, M. F. (2005). Direct-to-consumer advertising. *The Journal of the American Medical Association* **293**, 2030–2033.

Iizuka, T. (2004). What explains the use of direct to consumer advertising of prescription drugs? *Journal of Industrial Economics* **52**(3), 349–379.

Iizuka, T. and Jin, G. Z. (2005). The effect of prescription drug advertising on doctor visits. *Journal of Economics and Management Strategy* **14**(3), 701–727.

Kopp, S. W. and Sheffet, M. J. (1997). The effect of direct-to-consumer advertising of prescription drugs on retail gross margins: Empirical evidence and public policy implications. *Journal of Public Policy and Marketing* **16**(2), 270–276.

Kravitz, R. L., Epstein, R. M., Feldman, M. D., et al. (2005). Influence of patients requests for direct-to-consumer advertised antidepressants: A randomized controlled trial. *Journal of the American Medical Association* **293**(16), 1995–2002, Erratum in: *Journal of the American Medical Association* **294**(19), 2436.

Kwong, W. J. and Norton, E. C. (2007). The effect of advertising on pharmaceutical promotion. *The Review of Industrial Organization* **31**, 221–236.

Law, M. R., Majumdar, S. R. and Soumerai, S. B. (2008). Effect of illicit direct to consumer advertising on use of etanercept, mometasone, and tegaserod in Canada: Controlled longitudinal study. *British Medical Journal* **337**(a1055), 557–560.

Law, M. R., Soumerai, S. B., Adams, A. S. and Majumdar, S. R. (2009). Costs and consequences of direct-to-consumer advertising for clopidogrel in medicaid. *Archives of Internal Medicine* **169**(21), 1969–1974.

Mintzes, B., Morgan, S. and Wright, J. M. (2009). Twelve years' experience with direct-to-consumer advertising of prescription drugs in Canada: A cautionary tale. *PLoS One* **4**(5), e5699, 1–7.

Narayanan, S., Desiraju, R. and Chintagunta, P. (2004). Return on investment implications for pharmaceutical promotional expenditures: The role of marketing-mix interactions. *Journal of Marketing* **68**(4), 90–105.

Rizzo, J. (1999). Advertising and competition in the ethical pharmaceutical industry: The case of hypertensive drugs. *Journal of Law and Economics* **42**(1), 89–116.

Rosenthal, M. B., Berndt, E. R., Donohue, J. M., Epstein, A. M. and Frank, R. G. (2003). Demand effects of recent changes in prescription drug promotion. In Cutler, M. and Garber, M. (eds.) *Frontiers in health policy research*, pp. 1–26, vol. 6. Cambridge, MA: MIT Press.

Scott Morton, F. M. (2000). Barriers to entry, brand advertising, and generic entry in the U.S. pharmaceutical industry. *International Journal of Industrial Organization* **18**, 1085–1104.

SK&A (2011). U.S. Pharma Company promotion spending. *Cegedim Strategic Data Report*. Available at: http://www.skainfo.com/ (accessed 15.12.11).

Steiner, R. (1973). Does advertising lower consumer prices? *Journal of Marketing* **37**(4), 19–26.

Toop, L., Richards, D., Dowell, T., et al. (2003). *Direct to Consumer Advertising of Prescription Drugs in New Zealand: For Health or for Profit?* Report to the Minister of Health supporting the case for a ban on DTCA. New Zealand departments of general practice. Christchurch, Dunedin, Wellington and Auckland Schools of Medicine. February.

Wosińska, M. (2002). Just what the patient ordered? Direct-to-consumer advertising and the demand for pharmaceutical products. Harvard Business School Marketing Research Papers No. 02-04.

Pharmaceutical Parallel Trade: Legal, Policy, and Economic Issues

P Kanavos and O Wouters, London School of Economics, London, UK

Glossary

Arbitrage A trading strategy that exploits profit opportunities arising from price differences across jurisdictions.

Commercial drug importation or pharmaceutical parallel trade The legal movement of a drug between countries by a third party without the authorization of the originator manufacturer.

Direct to pharmacy distribution (DTP) In the DTP model, pharmaceutical manufacturers bypass wholesalers to deliver drugs directly to pharmacies; wholesalers only provide logistic support and do not own the stock.

Exhaustion of intellectual property rights The exhaustion doctrine in intellectual property law mandates that once a product has been legitimately marketed, the patent holder exhausts their control over the product in that market. The type of exhaustion regime (i.e., national, regional, or international) corresponds to the patent holder's level of control of the product in foreign markets.

External price referencing (EPR) A country compares pharmaceutical prices in a basket of countries to establish a domestic reference price. The purpose of EPR may be to (1) negotiate or set prices within a country, (2) negotiate coverage and reimbursement levels, or (3) authorize product marketing.

Price discrimination The sale of goods or services to different individuals at different prices.

Ramsey pricing A linear pricing scheme in which a natural monopoly sets prices as a function of consumers' inverse demand elasticities. In the context of pharmaceuticals, this allows drug manufacturers to recover their sunk R&D costs and establishes prices based on patients' price sensitivity.

Reduced wholesaler model (RWM) In the RWM, pharmaceutical manufacturers only contract with a small number of wholesalers whose distribution behavior they can monitor more effectively; the traditional principles of wholesaling apply for RWMs.

TRIPS Agreement The World Trade Organization (WTO) Agreement on Trade-Related Aspects of Intellectual Property Rights, more commonly known as the TRIPS Agreement, is a comprehensive multilateral agreement on intellectual property. It provides the legal framework for pharmaceutical parallel trade.

Welfare A measure of an individual's or society's level of well-being.

Background

Pharmaceutical parallel trade, also called commercial drug importation, is the legal movement of a product between countries by a third party (usually a wholesaler or a distributor) without the authorization of the originator manufacturer (who is often also the patent holder). The potential for pharmaceutical parallel trade may arise given three conditions. First, price differences across jurisdictions for the same product must exist due to variation in price regulation, discordant willingness to pay, exchange rate fluctuations or other factors influencing drug prices. Second, manufacturers retain weak vertical control over the supply chain once they sell a product to wholesalers or distributors. Third, some countries adopt regimes of international or regional exhaustion of intellectual property rights (IPRs) that limit the patent holder's ability to block trade once they have marketed a product. Under these conditions, arbitrage opportunities should stimulate the export of medicines from low- to high-price countries.

Parallel trade is particularly prevalent in the European Union (EU), where it accounts for up to 20% of retail prescription drug spending in some countries. The legalization of commercial drug importation has also been proposed in the US and it remains a highly controversial issue in countries around the world. Proponents of parallel trade, including some public health authorities, argue that it increases affordability of medicines and generates health system savings. Theoretically, the mere threat of parallel trade (whether or not it actually occurs) should lead manufacturers to reduce prices. Critics, however, question its long-term effects on drug prices. They also claim that parallel distributors free ride on the marketing and service investments of authorized wholesalers and that widespread parallel trade would decrease profits for pharmaceutical manufacturers; ultimately, it would reduce pharmaceutical investment in research and development (R&D) and harm innovation. This article outlines the legal, policy, and economic issues surrounding pharmaceutical parallel trade. As it has been most common in the EU, much of this article focuses on this region.

The Legal Debate

At the international level, the legal framework governing pharmaceutical parallel trade is provided by the World Trade Organization (WTO) Agreement on Trade-Related Aspects of Intellectual Property Rights, more commonly known as the TRIPS Agreement. There have been recent amendments and additions to the TRIPS Agreement that facilitate the parallel trade of medicines (e.g., the Doha Declaration), as well as bilateral agreements that restrict it (e.g., TRIPS-plus). Regional and national legislation also affects trade and competition

 doi:10.1016/B978-0-12-375678-7.01213-X

(see Kyle (2009) for an in-depth analysis of the legal aspects of pharmaceutical parallel trade).

Intellectual Property Provisions and Parallel Trade

The TRIPS Agreement requires that WTO member states adopt minimum standards of IP protection and enforcement of IPRs. However, Article 6 of TRIPS leaves to the discretion of individual countries whether they choose to adopt a regime of national, regional, or international exhaustion of IPRs. The exhaustion doctrine in intellectual property law mandates that once a product has been legitimately marketed, the patent holder exhausts their control over the product in that market. The type of exhaustion regime corresponds to the patent holder's level of control of the product in foreign markets. Under a regime of national exhaustion, the patent holder can block the parallel import of the product from a foreign market, whereas the parallel import is legal under a regime of international exhaustion. Regional exhaustion permits parallel trade within the region, but not from nonmember countries. Frequently cited examples of national, regional, and international exhaustion regimes are the United States (US), the EU, and Kenya, respectively. The Doha Declaration on the TRIPS Agreement clarifies that members can implement the exhaustion regime that supports their domestic policy goals.

Within the EU, rules governing the single market prevent member states from imposing trade barriers on imports from other member states and mandate a regional, or community, exhaustion of IPRs. The European Court of Justice (ECJ) has taken the view that once a product is sold in one member state, it is a breach of Article 28 of the EC Treaty to prevent the product from being resold in another member state even if it is protected by a patent or other forms of IPRs. This is consistent with the principle of free trade of goods between EU members, although the regional exhaustion regime still prevents the parallel importation of a product that was first sold outside of the EU. The ECJ has upheld the legitimacy of parallel trade in numerous cases (e.g., Bristol-Myers Squibb vs. Paranova A/S, 1996, ECJ Case 427/93).

Parallel distributors must often repackage a drug to comply with the labeling and package criteria of the importing country. Manufacturers have legally challenged repackaging and claimed that it may adversely affect the consumer perception of the quality of the drug and harm the manufacturer's reputation. The ECJ, however, has generally ruled in favor of parallel traders.

Competition and Trade Policy

European Union

In response to parallel trade in the EU, several drug manufacturers have attempted various practices to discourage parallel trade, including restricting supply to exporting countries, price discriminating based on whether the drugs are meant for domestic or foreign consumption (i.e., exported) or only selling to a distributor if they agree to not engage in parallel trade. Owing to the single market in the EU, however, restrictions on parallel trade may violate clauses of the EC Treaty that pertain to collusive behavior and cartels (Article 81) and abuse of dominant position (Article 82); such practices could therefore be deemed anticompetitive.

EU courts have ruled that any implicit or explicit decision between the manufacturer and the distributor that limits the extent of parallel trade violates Article 81 (e.g., Sandoz Prodotti Farmaceutici vs. Comm'n, 1990, ECJ Case C-277/87). However, pharmaceutical manufacturers have occasionally been allowed to restrict supply to a quantity sufficient only for domestic consumption. In the case Bayer vs. Commission (2000, ECJ Case T41/96), the court ruled that Bayer was allowed to limit the supply of the drug Adalat to the Spanish and French markets to prevent outside consumption of the drug. As the domestic demand must first be satisfied before a drug can be exported (i.e., parallel trade can be banned if it induces a shortage in the exporting country), the court found that the arrangement did not violate Article 81 and ruled in favor of Bayer.

Even if draining the domestic market is subject to interpretation under Article 81, it may still constitute an abuse of a pharmaceutical manufacturer's dominant position (Article 82). In a different case (Syfait vs. GlaxoSmithKline, 2008, ECJ Cases C-53/03, C-468/06 and C-478/06), the ECJ found that GlaxoSmithKline, which was limiting supply to Greek distributors, held a dominant position but did not necessarily violate Article 82. The main reasoning behind the decision was that, unlike in other industries, intercountry differences in drug prices are often the result of discordant national regulatory environments and are only weakly influenced by manufacturers. However, the court later reversed its opinion that pharmaceutical manufacturers are not largely involved in the pricing process and found that GlaxoSmithKline violated Article 82 (Sot. Lelos kai Sia EE vs. GlaxoSmithKline AEVE, 2008, ECJ Cases C-468/06 and C-478/06).

Disputes have also arisen over whether pharmaceutical manufacturers can charge different prices based on if a drug will be consumed in the home or foreign country or selectively withdraw a product from a market to curb the volume of parallel trade. Although explicit price discrimination would violate Article 81, pharmaceutical manufacturers have again claimed that the pricing outcome was the result of regulatory pressures, not a manufacturer decision. The EU courts have offered mixed rulings on this issue depending on the circumstances of individual cases. The courts have also ruled that Article 81 prevents manufacturers from ceasing to supply a market due to the threat of parallel trade, as some theoretical models have suggested would occur; manufacturers must guarantee adequate access to needed drugs.

Overall, despite innovative attempts by pharmaceutical manufacturers to hinder parallel trade, the European courts have generally ruled in favor of parallel trade and supported the free movement of products within the EU. The current policy imperative is how to prevent the parallel distributors from capturing all of the transaction profits, in order to transfer the savings to consumers and purchasers.

United States

Parallel trade has received extensive policy attention in the US, where pharmaceutical prices have historically been higher than in other countries (e.g., Canada). Some US policymakers have championed parallel trade as a viable option to curtail pharmaceutical spending. Advocates of parallel trade argue

Table 1 Overview of key conceptual and theoretical studies on the economic impact of pharmaceutical parallel trade

Endpoint evaluated	*Results*	*References*
Innovation	Global R&D investment decreases	Li and Maskus (2006); Pecorino (2002)
	Global R&D investment increases	Grossman and Lai (2008)
	Ambiguous effect on global R&D investment	Valletti (2006)
Price competition	Downward price convergence	Ganslandt and Maskus (2004); Jelovac and Bordoy (2005)
	Upward price convergence	Kanavos and Costa-Font (2005); Costa-Font and Kanavos (2007); Kanavos and Vandoros (2010); Vandoros and Kanavos (2013)
	Ambiguous effect on prices	Pecorino (2002); Maskus and Chen (2004); Valletti (2006)
Social welfare	Social welfare increases	Maskus and Chen (2004)
	Ambiguous effect on social welfare	Jelovac and Bordoy (2005); Valletti and Szymanski (2006)

Source: The authors based on the literature.

that drug importation can reduce US drug prices without direct price regulation; critics argue that it would introduce significant safety risks due to the potentially less stringent quality assurance of foreign regulatory bodies, would not necessarily reduce prices to US consumers, and could ultimately harm pharmaceutical innovation and threaten the competitive advantage of US firms. Although legislation authorizing parallel trade was passed by the US Congress and signed into law in 2000, it was never implemented due to the aforementioned concerns. In the last decade, several other congressional bills have been introduced, but all have ultimately been defeated.

Despite the illegality of commercial drug importation, the existence of a small but steady flow has induced various responses by pharmaceutical manufacturers. Certain manufacturers have limited Canadian supply and have even circumvented distributors and supplied retail pharmacies directly to mitigate the risk of parallel trade. This has led to investigations by US antitrust authorities into possible collusive or anticompetitive behavior among manufacturers in violation of the Sherman Antitrust Act of 1890. Parallel exportation has also been opposed by the Canadian authorities, which have raised concerns about the possibility of supply shortages.

The US experience suggests that a single market may be a prerequisite for the successful introduction of parallel trade. The special regulation of pharmaceutical products under the Federal Food, Drug, and Cosmetic Act (1938) and the Prescription Drug Marketing Act (1987) also means that even if parallel trade were authorized for other goods, it would not necessarily extend to pharmaceuticals.

Conceptual Frameworks for Evaluating Parallel Trade

Several models have been proposed to explain the behavior of pharmaceutical manufacturers, payers, and regulatory agencies in the presence of parallel trade opportunities. Most studies employ a variety of theoretical approaches to explore the effects of pharmaceutical parallel trade on price competition, social welfare, and innovation; key studies are summarized in Table 1.

Price Discrimination, Innovation, and Welfare Effects

The evidence suggests that the effect of parallel trade on social welfare is an empirical issue that depends on parameters such as demand and demand-side policies, regulation, patient preferences, market structure, and innovation. The social welfare effects of pharmaceutical parallel trade are ambiguous: in the short term, lower prices may contain growth in pharmaceutical expenditure and improve patient access to medicines, whereas profit reductions to originator manufacturers may discourage pharmaceutical investment in R&D and thus harm the long-term prospects of innovative drug discovery.

Traditionally, economic analyses of parallel trade have considered it as a channel for undermining third-degree price discrimination. Without considering distribution issues and differing demand elasticities across countries for homogeneous goods, parallel trade could lead to uniform international prices. Economic theory predicts that in unregulated markets and in the absence of product differentiation, arbitrage induces price competition and stimulates downward price convergence. Nevertheless, the static welfare effects of third-degree price discrimination compared with a single price monopoly are likely to be positive if price discrimination leads to higher output and consumption, which seems plausible for pharmaceuticals.

Theoretical studies have generated opposing conclusions on the effect of parallel trade on innovation. Most studies have assumed that price regulations are fixed across countries and have shown that parallel trade may disincentivize R&D investment. In a research-intensive industry, parallel trade may threaten the ability of manufacturers to recoup their R&D investment costs. Assuming that parallel trade does not affect the total volume of consumption across all countries, parallel trade satisfies part of the demand in the importing country at a lower price. The overall profit to the manufacturer is therefore reduced. Whereas free trade is often promoted to allow all consumers to benefit from more efficient production functions, this stance may not be appropriate for the in-patent pharmaceutical market, where lower drug prices are often the outcome of more stringent regulatory regimes, not lower production costs. In this context, the use of Ramsey pricing, or a markup based on inverse demand elasticities, may maximize global social welfare and is preferable to international (or even regional) exhaustion of rights.

Other models considered a situation where the assumption of fixed pricing environments does not hold, in which case parallel trade may instead stimulate the convergence of price

regulations to those of the exporting nation and encourage R&D investment. In other words, when IPRs are exhausted internationally and different prices exist across countries, parallel trade may stimulate an international harmonization of price regulations and support innovation. The ex-post allocation of resources should motivate the importing country to provide a trade friendly environment for the firms. These theoretical findings are supported by evidence of more relaxed pricing regimes in Portugal, Italy, and France following the growth of parallel exports in those countries.

Some models have even found that parallel trade should increase the profits of the originator manufacturer in the context of a Nash bargaining game. Although such a finding is counterintuitive when one considers the opposition of pharmaceutical manufacturers to parallel trade, certain variables are absent in these models that may limit their practical application. For example, they may not consider the possibly detrimental impact of parallel trade on the ability of manufacturers to price discriminate in markets other than the ones considered in the model.

In general, the theoretical conclusions differ significantly depending on the model assumptions (Table 1). These assumptions include (1) the type of regulation in place (e.g., fixed vs. variable regulatory regimes, the threat of compulsory licensing, price caps, etc.); (2) the ability of manufacturers to discriminate based on quality (i.e., produce high- and low-quality products) or decide not to serve a particular market; and (3) the manufacturer's level of control over prices in the exporting and importing markets (to incorporate parallel trade into cost functions), among other factors.

Impact on Competition and Prices

Standard microeconomic theory predicts that parallel trade catalyzes price competition. This anticipated price outcome was supported by evidence from Sweden between 1994 and 1999, which showed that prices of drugs exposed to parallel competition dropped by 12–19%. Other studies, however, have found that the distribution chain collects most of the economic gains from pharmaceutical parallel trade and that prices remain stable at their initial level in destination countries. Due to universal health insurance coverage and low copays or copay exemptions, patients and payers (i.e., governments and health insurers) rarely benefit directly from pharmaceutical price differences. Instead, parallel distributors obtain the majority of available rents, as purchasers have limited incentives to respond to price differentials.

The volume of parallel trade does not necessarily put downward pressure on prices in destination countries. Several studies have found little evidence of sustainable price competition, even when parallel trade is actively promoted. Exchange rate fluctuations, purchasing power parities and generic penetration may be responsible for any price decreases. The pricing strategies of parallel distributors rarely deviate from those of local manufacturers in destination countries, despite discordant cost functions. Assuming both parallel distributors and originator drug manufacturers are profit maximizers and that the former face negligible marginal production costs, game theory models predict upward price convergence which minimizes the opportunity for welfare gains.

Overall, the accuracy and validity of the various theoretical models depend primarily on the respective assumptions about the behavior of originator manufacturers, parallel traders and regulatory authorities in the home and foreign markets in the presence of arbitrage opportunities. Given discordant assumptions across models (e.g., outturn of pricing negotiations, manufacturer R&D strategies or patient copayment levels), it is expected that different theoretical conclusions will be derived.

Determinants of Parallel Trade

Barriers to Entry

Despite the rapid growth of pharmaceutical parallel trade in recent years and the adoption of policies encouraging its use, this has not been without barriers. First, a parallel traded product needs to be approved and licensed by national regulatory agencies to ensure product safety. However, the ECJ and the associated jurisprudence have simplified these procedures and parallel importation can be approved quickly. Second, the nature of the pharmaceutical distribution chain suggests that obtaining market share requires a minimum scale of operations: parallel distributors must be in a position to sustainably supply a significant number of products to local retailers, otherwise they risk not becoming a preferred wholesaler and increasing retailers' costs of compliance. Third, given the fragmented structure of European wholesaling, parallel distribution requires large networks of national wholesalers from whom medicines can be purchased at favorably low prices and exported to high-price countries. Fourth, manufacturers are increasingly in a position to exercise some vertical control over the supply chain in all countries where they operate – provided this is not explicit – and this means that extra quantities for parallel exportation are difficult to obtain. Fifth, other barriers to entry may relate to the negative perception of parallel traded products among consumers in that the packaging, language, and presentation may be different to what consumers are accustomed to. Sixth, consumers may negatively perceive parallel traded products if the packaging, language, and presentation are different to what they are accustomed to. Finally, other barriers to entry include: (1) small license fees, (2) the requirement to insert a patient leaflet in the destination country's language, and (3) moderate transport costs.

Pricing Strategies across Countries

Perhaps, the greatest determinant of parallel trade is the wholesale price difference between the importing and the exporting country. This price difference provides the initial signal for parallel trade, taking transport and regulatory complicance costs into consideration. The magnitude of the price difference will determine whether parallel trade is likely to be profitable for those who exercise it. Empirical evidence finds a positive and statistically significant relationship

between the magnitude of the price gap and market entry and penetration of parallel traded products.

Drug manufacturers may attempt to limit the impact of parallel trade. Whether or not manufacturers are successful at discouraging parallel trade depends on two key parameters. First, the control that manufacturers exercise over the distribution chain (see Section Fragmentation of the distribution chain and vertical control). Second, the types of price regulation applied across countries. The widespread implementation of external price referencing (EPR) in the EU has indirectly enabled manufacturers to adopt various strategies to mitigate the impact of parallel trade. One of these is entry or launch sequencing, whereby a product is launched in countries where prices are higher to influence prices in other countries that use the launch country as a reference. Another strategy is launch delay, especially if expected prices in a country are below a certain threshold and may induce parallel trade or influence prices downwards elsewhere. A third strategy is to either not launch or withdraw from a market given the level of expected prices. This can take place at any time, especially if EPR is used repeatedly to take advantage of price changes in a country's reference basket.

Product Availability

If parallel distributors are not able to acquire sufficently large product stocks in export countries, parallel trade cannot occur. Importantly, pharmaceutical manufacturers have little, if any, interest to promote parallel trade between markets by diverting sales of their own product from high- to low-price countries. On the contrary, the ability of parallel distributors to identify sources of product is a key driver to the conduct of parallel trade and requires knowledge of price levels across jurisdictions. Sources of product can be wholesalers (for larger quantities) or retailers (for smaller quantities) in export countries. The source of parallel traded products is not always exclusively the lowest price jurisdiction, but a range of jurisdictions depending on the availability, proximity, and price difference to the destination country. The security of supply affects the capacity of parallel distributors to offer competitive prices and the long-term sustainability of parallel trade in a particular product market.

Administrative Measures and Incentives

Governments and health insurers often promote the use of parallel imported products to encourage price competition and to achieve efficiency savings. The experience of European countries is considerable in this regard and has involved both direct and indirect financial incentives and penalties to individual stakeholders; these policies are country-specific and focus primarily on the incentive structure at the retail level. The most prevalent policies are summarized in Table 2 and relate to:

- First, the award of positive financial incentives for pharmacies to actively procure and distribute parallel imported pharmaceuticals. Pharmacies can retain a proportion of the price difference between the originator and the parallel imported medicine (e.g., 33% of that difference in the Netherlands and 50% in Norway).
- Second, the presence of negative incentives, whereby pharmacies are penalized if they do not dispense a parallel traded product. In Germany, sickness funds and pharmacy associations have agreed on a parallel import quota for the latter to dispense in a given year, which is based on pharmacies' overall turnover with the sickness funds (e.g., 7% in 2003). Sickness funds receive all the financial benefits from price differences. The reimbursement for pharmacies that fail to comply with the quota is reduced accordingly. In the opposite case, pharmacies receive a credit to settle bills when the import quota is surpassed, although there is no cash benefit.
- Third, the mandatory provision of information, whereby pharmacies are legally bound to inform patients on the availability of a cheaper parallel imported drug if savings reach up to 5% on a prescribed product (e.g., Denmark). No financial benefits are envisaged for pharmacies.
- Fourth, financial incentives by means of additional lump sum payments (e.g., Sweden) and in connection with national substitution policies to enhance the use of generic and parallel traded medicines.
- Finally, discounting practices for the procurement of medicines exist in many countries (e.g., the Netherlands and the UK). Pharmacy savings on procurement prices of medicines (including parallel traded medicines) compared with list prices form part of their remuneration; regulators

Table 2 Policy measures encouraging the dispensing and use of parallel traded pharmaceuticals in Europe

Type of measure[a]	*Evidence of application*	*How it works in practice*
1. Positive financial incentive	The Netherlands and Norway	Pharmacies retain part of the price difference between parallel traded and locally sourced drugs (33% in the Netherlands, 50% in Norway)
2. Penalty	Germany	Penalties to pharmacies if they do not dispense a certain proportion of parallel traded drugs (i.e., quota system)
3. Compulsory information	Denmark	Pharmacists are required to provide information to patients about a parallel traded drug if the price difference exceeds 5% of prescription cost
4. Lump sum payments	Sweden	Payments to pharmacies for their promotion of generic and parallel traded drugs
5. Discount and clawback	The Netherlands and the UK	Discounting allowed, but a proportion of that discount to pharmacy is retained via the clawback

[a]Measures reported in this table have been implemented in the cited countries at particular points in time and may be of transient nature or subject to changes.
Source: The authors based on the literature.

can retain part of that discount by implementing a clawback when reimbursing pharmacies.

Other Factors Influencing the Extent of Parallel Trade

Several other factors are likely to influence the amplitude of parallel trade. These are discussed below.

Fragmentation of the distribution chain and vertical control

The fragmentation of the distribution chain in exporting countries is positively associated with parallel trade; the more fragmented it is, the more likely wholesalers – and possibly retailers – are to parallel export part of their stock. For wholesalers, the financial incentive to engage in parallel exporting can be greater than distributing products to the domestic retail chain. To counteract this incentive, regulators in some European countries require wholesalers to register and report the destination of their products (Spain) or to keep a stock at 25% more than historical demand (Greece). Recently, pharmaceutical manufacturers have sought to increase their control of the distribution chain, either by bypassing wholesalers and delivering directly to pharmacy (direct-to-pharmacy (DTP) model) or contracting with a small number of wholesalers (reduced wholesaler model (RWM)), typically two or three, whose distribution behavior they can monitor more effectively. In the DTP model, wholesalers only provide logistical support and do not own the stock, whereas the traditional principles of wholesaling apply for RWMs. In practice, however, the 'public service obligation,' requiring all registered wholesalers to be sufficently stocked to further supply the product to the retail distribution chain, is a significant barrier to the wider implementation of either model.

Market size and geographical proximity

One of the likely indicators of the potential for parallel trade in individual countries is the country's 'economic size.' The larger a particular market, the more attractive it is for both manufacturers and distributors to undertake production and parallel trade, respectively. The gravity model of international trade predicts that the flow of goods between two locations is positively related to their size (or income levels) and negatively related to the distance between them, after controlling for factors that may affect trade (e.g., price differences and differences in the salient features of regulatory frameworks). Empirical evidence suggests that distance appears to reduce the total volume of parallel trade into a country; however, a gravity specification exhibits an opposite effect for volume (as a proxy for market penetration). This may be because once parallel distributors have established contact with a potential source, distance does not become a significant barrier and geographically distant sources may have incentives to become better connected.

Patent expiry and genericization

The effect of generic entry on the extent of parallel trade has not been studied in a systematic way. Genericization leads to competition between originator and generics, which may result in the price of the originator declining. If this is combined with administrative policies (e.g., compulsory generic substitution), the market share of originators (locally sourced or parallel imported) is likely to be small. The only reason why genericization might enhance parallel trade is if there is differential patent expiry among jurisdictions, in which case, patent-expired originators in one jurisdiction may be acquired at a significant discount and resold in jurisdictions where the same product remains patent protected.

Exchange rate fluctuations

Exchange rate fluctuations can indirectly influence the extent of parallel trade, insofar as currency appreciations or depreciations affect relative prices of pharmaceutical products across countries. *Ceteris paribus*, currency depreciation could make exportation more attractive, or could even reverse the flow of trade, if price differences between the export and the import country are small. In the UK, moderate average price reductions in the context of the Pharmaceutical Price Regulation Scheme (PPRS), along with the depreciation of sterling *vis-à-vis* the Euro, reversed the flow of trade in 2007. Since then, the UK has been a net exporter of pharmaceutical products, compared with the opposite situation until that point; parallel imports accounted for approximately 20% of the UK prescription in-patent market in 2002.

Parallel Trade and Its Economic Impact

Stakeholder Positions

The economic and noneconomic impact of parallel trade on health systems is heavily debated by the key stakeholders in both exporting and importing countries, including parallel distributors, drug manufacturers, health insurers, distribution chain actors and patients. To understand the position of different stakeholders on the subject, it is important to evaluate the incentive structure that motivates them. The available evidence from the EU, where the subject has been studied at some length, is important in this context.

Parallel distributors argue that pharmaceutical parallel trade promotes competition by forcing down the prices of their domestically sourced counterparts and enhances access to medicines; this should contribute to a lower public pharmaceutical expenditure and result in direct and indirect health care savings. Parallel distributors are motivated by the private pecuniary benefits of arbitrage, as the potential for price equalization in regulated pharmaceutical markets is low and price differences are sustained over time. The key issue that affects the volume of parallel trade is the sustainability of supply in potential export markets.

Pharmaceutical manufacturers oppose parallel trade. They argue that it undermines the profitability and ability of manufacturers to recoup R&D investment costs; it therefore harms future investment in R&D and the potential for discovering new treatments. Pharmaceutical parallel trade may provide an unfair playing field, as manufacturers take all the risk in developing drugs and parallel distributors profit from the R&D of manufacturers. As a result, areas where parallel trade takes place or is encouraged may become increasingly unattractive for conducting business, which could cause job losses, cutbacks, and industry relocation in the long term.

Statutory health insurance organizations in exporting (source) countries are not in a position to realize any pecuniary benefits, as products that are meant for the treatment of patients under their jurisdiction are parallel exported to other markets. Instead, parallel exportation can result in product shortages in exporting countries and limit patient access to medicines.

Statutory health insurance organizations in importing (destination) countries, however, may benefit in three ways from a conceptual standpoint. First, savings from parallel trade could accrue partly or entirely to them. In the Netherlands and Norway, the government maximizes its financial benefits by surrendering part of the price difference to pharmacists (see Table 2). Second, health insurance organizations can implement 'clawback' arrangements to ensure that part of the pecuniary benefits accruing to retailers from more competitive purchasing are accounted for through lower reimbursement. Third, price competition may lower pharmaceutical expenditure in destination countries; however, it is difficult to determine the extent to which this is occurring, which is likely to be product-specific. These benefits are static and need to be balanced against the countries becoming less attractive R&D investment locations due to parallel trade.

The retail distribution chain can also benefit in countries where pharmacy margins are not determined by regulation, or where pharmacies are explicitly financially incentivized to dispense a parallel imported product. In this case, pharmacies can negotiate discounts with distributors and parallel distributors, making it profitable to stock and dispense parallel imported medicines that may carry the same reimbursement price as a locally sourced equivalent.

The benefits to patients in destination countries theoretically result from the lower prices of parallel imported drugs, which may reduce patients' overall medication costs and improve their access to medicines. This also assumes that patients pay a significant proportion of their medication out-of-pocket. Such benefits, however, may be transient unless parallel trade can be a sustainable source of cheaper products. The above arguments have little applicability in health systems offering comprehensive prescription drug insurance coverage, however, where patients contribute modestly to the cost of their medicines. The type of cost sharing also plays a role: fixed prescription charges do not offer any financial benefits to patients, whereas a coinsurance might.

Evidence of Economic Impact and Welfare Implications

Parallel trade has generated considerable interest about its economic impact, welfare implications, and effects on the stakeholder community. Few empirical studies have examined its impact on stakeholders and price competition and all focus on the EU experience.

Static gains from parallel trade

Direct benefits from parallel trade can be significant, are static in nature, and arise because of the price differences between exporting and importing countries resulting in a redistribution of revenue from pharmaceutical manufacturers to a variety of stakeholders.

Financial gains to health insurers are modest both in absolute terms and as a share of total prescription drug spending; evidence from six European countries suggests that on average they are less than 2% of the inpatient–outpatient prescription drug market (ranging from 0.3% in Norway to 3.6% in the Netherlands). It has been found that the cumulative financial gain to parallel distributors is much greater than the gain to health insurers. The retail distribution chain also profits directly from the conduct of parallel trade.

Market entry, penetration, and competition

Much of the debate on pharmaceutical parallel trade has focused on its indirect benefits and whether it creates sustainable price competition in the long run. Empirical observations suggest that in importing countries where parallel traded products occupy a significant market share, the difference between the highest and lowest parallel distributors' price rarely exceeds 7%. The distributors with the largest market share are those with prices toward the lower end of the spectrum or those with the lowest price in the range. This indicates that payers and retailers exhibit some price sensitivity to parallel import prices. Still, prices of locally sourced equivalent products do not seem to be influenced by the extent of parallel trade. In several cases, they have actually increased over time, despite seeing their domestic market share declining in the presence of parallel imports. The price spread between parallel traded and locally sourced drugs can vary significantly from shipment to shipment, and from destination to destination, depending on (a) the acquisition price in a source country and (b) the incentives associated with the distribution and sale of parallel imported products in a destination country. In the UK, for example, this difference can be zero as parallel distributors may choose to have a list price at parity with the locally sourced equivalent, enabling them to offer further discounts to retailers.

Few studies, however, have analyzed whether price competition is sustained over time. Empirical evidence from several countries suggests a commensurate average price change of parallel imported and locally sourced products over time; this indicates that there is a co-movement in prices in the presence of parallel trade. The prices of locally sourced and parallel traded products appear to converge upwards rather than downwards. These results are sensitive to the product mix they refer to, the country settings where they are derived from and the time period studied.

The absence of a strong and sustainable price effect is likely due to a number of reasons. First, the fragmentation in the supply chain generates an environment where price differentials between exporting and importing countries are typically divided among several rent-seeking agents. Second, the lack of security in procuring the necessary quantities of product from potential export countries in a sustainable manner undermines downward price convergence. Third, parallel distributors may not always import from the lowest price countries for a variety of reasons (e.g., quantity available at a particular time or geographic proximity to the destination country). Finally, the regulatory and operating environment in importing countries may not be conducive to parallel distributors offering health insurers any price advantage to locally sourced products.

Safety and quality of parallel traded products

The need to repackage some parallel traded medicines increases the risk of counterfeiting. From a regulatory perspective, parallel distributors are obliged to notify the drug regulator in the destination country, as well as the originator manufacturer, of any changes made to the relevant product to obtain an import license for a particular shipment of the product; this makes them liable in case there is tampering with the medicines and counterfeits enter the distribution chain. For example, thousands of packs of three parallel traded counterfeit medicines entering the UK supply chain were recently recalled. The counterfeit medicines were manufactured in China and entered the EU in Luxembourg, where they were resold, without being checked, to UK and Belgian wholesalers. As the drugs entered Luxembourg as 'transit' goods, they were exempt from the checks that would otherwise apply.

Impact on exporting countries

Even if the impact of parallel trade on patients is neutral in the destination country, it is important to evaluate its impact on drug availability in the source country. Arbitrage incentivizes parallel distributors to maximize the quantities they acquire in the exporting country for resale elsewhere. Wholesalers in the source country may also have an incentive to supply parallel distributors with maximum quantities (and provide them with a discount), rather than supply their own market, because they reduce their overall distribution costs when they sell to a single buyer rather than to several buyers (e.g., pharmacies). Consequently, while the distribution system in the export country favors parallel trade, or has few available limits to prevent its extent, this may have an adverse effect on the availability of medicines in that country and thus harm the health of the patients.

Evidence suggests that parallel trade can induce shortages in drugs that are exported intensively. This has been documented in Greece, Spain and, more recently, the UK. The governments of all three countries now require wholesalers to ensure that local demand is first satisfied. Greece implemented short-lived parallel export bans in 2011, 2012, and 2013. In the UK, many products were officially listed in short supply and government authorities threatened to punish manufacturers, wholesalers, pharmacists, and doctors who breached duties to supply medicines to the local market. One could claim that drug manufacturers should increase production in exporting countries to meet demand, but, given the incentives to domestic wholesalers discussed above, it is not guaranteed that increased production will satisfy this demand.

Impact on R&D and innovation

It is further argued that parallel trade has a deleterious effect on R&D and innovation due to the eroded profitability of innovator manufacturers. If a strong research-based industry that invests in R&D and discovers new molecules is a health (and industrial) policy objective in both exporting and importing countries, then profit erosion through parallel trade would provide an obstacle to achieving this goal. If nations compete on the basis of comparative advantages, parallel trade may discourage research-based manufacturers from locating in areas where it is widely practiced or is actively promoted. The encouragement of parallel trade by destination countries constitutes a beggar-thy-neighbor practice whereby higher-price countries borrow regulatory practices from lower-price countries.

Parallel trade and overall welfare implications

Given a range of stakeholder considerations and the available empirical evidence on welfare implications, the overall welfare effects of parallel trade are likely negative. The potential short-term benefit of parallel trade is cost containment in importing countries; if savings are sustained through price competition between locally sourced and parallel imported medicines, parallel trade may stimulate welfare gains. However, such gains are likely to be transient and negated by the adverse long-term effects of parallel trade. First, despite the perceived benefits from increased affordability, the empirical evidence suggests that parallel trade does not lead to sustainable price reductions in destination countries. Instead, price reductions occur in an opportunistic way, reflecting the availability of surplus product stock in source countries. Second, parallel exportation often leads to shortages in source countries, resulting in access problems for patients. Third, a reduction in manufacturers' profitability could lower R&D investment and levels of future innovation. Fourth, global welfare is also likely to suffer, as manufacturers are, in principle, disinterested in launching new products and in supplying countries with low-priced drugs.

Conclusions

Pharmaceutical parallel trade is a form of arbitrage caused by intercountry price differentials, which are in most cases generated by national regulatory practices. Depending on how the principle of exhaustion of IPRs is applied, the same product can be made available in different jurisdictions without the authorization of the rights holder. The practice of parallel trade has found wide applicability in the EU single market, where it accounts for a significant share of the retail prescription drug market. From a global perspective, the Doha Declaration leaves it up to the WTO members to handle it in a way that protects their national interests. Overall, the key determinants of parallel trade penetration are the price difference between the importing and the exporting country, the overall pharmaceutical market size of a country, and the fragmentation of the distribution chain.

The available evidence on the impact of parallel trade highlights three key conclusions. First, parallel trade does not seem to promote price competition. Instead, the prices of parallel traded and locally sourced products usually converge upwards; prices are more likely to be guided by domestic drug policy measures (e.g., regulatory and generic policies) than by parallel trade. Therefore, although the pharmaceutical market ostensibly offers a suitable environment for arbitrage, the expected price outcomes may be illusory. If national governments wish to further control their drug spending through supply-side measures (focusing on prices), parallel trade is probably a weak policy option. Second, the pecuniary and welfare gains from parallel trade in destination countries have proved to be at best modest; by contrast, exporting countries have suffered significant shortages, which negatively affect

access to treatments and social welfare. Third, there is an evolving jurisprudence that acknowledges the potential value of competition where parallel trade is permitted, but also recognizes the importance of protecting the commercial interests of originator manufacturers.

In the future, several important operational developments may mitigate the extent of pharmaceutical parallel trade. First, the widespread application of EPR enables manufacturers to streamline their launch sequences and opt to not launch if the expected prices would initiate parallel trade. Second, product shortages in export markets are likely to increase resistance to parallel trade. Finally, changing distribution practices (e.g., DTP distribution and RWM) allow manufacturers to exercise greater vertical control over their stock and its destination.

See also: Biopharmaceutical and Medical Equipment Industries, Economics of. International Trade in Health Services and Health Impacts. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Pharmaceuticals and National Health Systems

References

Costa-Font, J. and Kanavos, P. (2007). Medicines parallel trade in the European Union: A gravity specification. *LSE Health Working Paper Series*, WP6. London: LSE Health, London School of Economics.

Ganslandt, M. and Maskus, K. (2004). Parallel imports and the pricing of pharmaceutical products: Evidence from the European Union. *Journal of Health Economics* **23**(5), 1035–1057.

Grossman, G. and Lai, E. (2008). Parallel imports and price controls. *RAND Journal of Economics* **39**(2), 378–402.

Jelovac, I. and Bordoy, C. (2005). Pricing and welfare implications of parallel imports in the pharmaceutical industry. *International Journal of Health Care Finance and Economics* **5**, 5–21.

Kanavos, P. and Costa-Font, J. (2005). Pharmaceutical parallel trade in Europe: Stakeholder and competition effects. *Economic Policy* **20**(44), 751–798.

Kanavos, P. and Vandoros, S. (2010). Competition in prescription drug markets: Is parallel trade the answer? *Managerial and Decision Economics* **31**(5), 325–338.

Kyle, M. (2009). Parallel trade in pharmaceuticals: Firm responses and competition policy. In Hawk, B. (ed.) *International antitrust law & policy: Fordham competition law*. New York: Juris Publishing.

Li, C. and Maskus, K. (2006). The impact of parallel imports on investment in cost-reducing research and development. *Journal of International Economics* **68**(2), 443–455.

Maskus, K. and Chen, Y. (2004). Vertical price control and parallel imports: Theory and evidence. *Review of International Economics* **12**(4), 551–557.

Pecorino, P. (2002). Should the US allow prescription drug reimports from Canada? *Journal of Health Economics* **21**, 699–708.

Valletti, T. (2006). Differential pricing, parallel trade, and the incentive to invest. *Journal of International Economics* **70**(1), 314–324.

Valletti, T. and Szymanski, S. (2006). Parallel trade, international exhaustion and intellectual property rights: A welfare analysis. *Journal of Industrial Economics* **54**(4), 499–526.

Vandoros, S. and Kanavos, P. (2013). Parallel trade and pharmaceutical prices: A game-theoretic approach and empirical evidence from the European Union. *The World Economy*, doi: 10.1111/twec.12063. Published online: 21 March 2013.

Further Reading

Danzon, P. (1998). The economics of parallel trade. *Pharmacoeconomics* **13**(3), 293–304.

European Commission (2013). Single market for goods. Available at: http://ec.europa.eu/enterprise/policies/single-market-goods (accessed 20.03.13).

Food and Drug Administration (2010). Information on importation of drugs. *Prepared by the Division of Import Operations And Policy, FDA.* Available at: http://www.fda.gov/ForIndustry/ImportProgram/ucm173751.htm (accessed 20.03.13).

Kanavos P., Schurer, W. and Vogler, S. (2011). The pharmaceutical distribution chain in the European Union: Structure and impact on pharmaceutical prices. Report by EMINet; *Commission of the European Communities, DG Enterprise.* Brussels. Available at: http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/structimpact_pharmaprices_032011_en.pdf (accessed 06.09.13).

Malueg, D. and Schwartz, M. (1994). Parallel imports, demand dispersion and international price discrimination. *Journal of International Economics* **37**(3–4), 167–195.

Maskus, K. (2000). Parallel Imports. *World Economy* **23**(9), 1269–1284.

Richardson, M. (2002). An elementary proposition concerning parallel imports. *Journal of International Economics* **56**(1), 233–245.

Szymanski, S. and Valletti, T. (2005). Parallel trade, price discrimination, investment and price caps. *Economic Policy* **20**(44), 705–749.

Varian, H. (1985). Price discrimination and social welfare. *American Economic Review* **75**(4), 870–875.

World Health Organization (2000). The TRIPS Agreement and pharmaceuticals. Available at: http://apps.who.int/medicinedocs/en/d/Jh1459e (accessed 20.03.13).

World Trade Organization (2013). TRIPS material on the WTO website. Available at: http://www.wto.org/english/tratop_e/trips_e/trips_e.htm (accessed 20.03.13).

Pharmaceutical Pricing and Reimbursement Regulation in Europe

T Stargardt, University of Hamburg, Hamburg, Germany
S Vandoros, London School of Economics, London, UK

Introduction

Although the presence of a European single market has led to price convergence for many products, pharmaceutical markets still remain very fragmented, with the 27 European Union (EU) countries following different pricing and reimbursement policies. The level of pharmaceutical spending and the willingness to pay for innovation differs between and within EU countries as discussed by Drummond and Towse (2012), whereas the regulatory approaches dealing with the increase in pharmaceutical spending are fairly similar across countries and can be standardized, for example, see Vogler *et al.* (2008). Thus, this article will focus on the most common regulatory instruments used in the EU countries.

The Concept of Pricing and Reimbursement in the EU

In the EU, pricing and reimbursement of pharmaceuticals are closely related and interdependent, with the two terms often becoming mixed-up. As a result, 'pricing' and 'reimbursement' are frequently used as synonyms, but in reality they are not. Although the term 'reimbursement,' i.e., the amount or the percentage of the price of a pharmaceutical paid by a (public) payer for a predefined population or indication can easily be defined, it is much more difficult to explain the term 'price' in the EU context.

There are multiple numeric figures throughout the supply chain of pharmaceuticals that qualify for the term 'price.' The manufacturer's price refers to the price of a pharmaceutical set by the manufacturer, which may include value added tax (VAT). The wholesale price, which is defined as the price at which the wholesaler sells the medicine to the pharmacy, includes a mark-up for the wholesaler (in the EU countries usually defined by law) on the manufacturer's price and VAT. The pharmacy price – also known as retail price – defined as the price at which the pharmacy sells the pharmaceutical includes a mark-up for the pharmacy (in the EU countries usually – again – defined by law) on the wholesaler's price and VAT (Figure 1). Prices are called 'list prices' if they are published and do not include any rebates.

The term 'price' may not be defined only for whom it is set, but also to whom it is charged. When the price of a pharmaceutical refers to what is paid by the consumer, it is usually called 'copayment' or 'out-of-pocket payment' instead of 'price' and when the price of a pharmaceutical refers to what is paid by the (public) payer, it is usually called 'reimbursement' or 'reimbursement price,' i.e., the list price net of any copayments and rebates.

Pharmacy and wholesaler mark-ups differ from country to country, as does the VAT on pharmaceuticals. In their comparison on pharmaceutical regulation in the EU, Vogler *et al.* (2008) have found that wholesaler and pharmacy mark-ups are usually a linear or regressive (i.e., decreasing) function of the price. For example, in Germany, both mark-ups are regressive: the wholesaler mark-up on manufacturer prices is between 6% and 12% of the manufacturer's price, whereas the mark-up of the pharmacy consists of a fixed component, i.e., €8.10 per package, plus a linear mark-up of 3% on the wholesaler's price.

VAT differs even more drastically among the EU countries. According to the data of the Federal Union of German Associations of Pharmacists (2012), a small number of countries, i.e., Denmark, Bulgaria, and Germany charge their full VAT to pharmaceuticals, i.e., 25%, 20%, and 19%, respectively, whereas the majority of countries apply a reduced VAT between 2.1% (France) and 12% (Latvia) or do not apply VAT for outpatient prescription drugs at all (Ireland, Malta, Sweden, and the UK). These differences in mark-ups and VAT make price comparisons of list prices at the pharmacy level among the EU countries very difficult to interpret. Nevertheless, in most EU countries, these rules are applied only for prescription drugs to ensure that a pharmaceutical has the same retail price in all pharmacies in a country, regardless of whether the pharmacy is in an urban or rural area or whether the pharmacy faces competition or not.

The Process of a Price and Reimbursement Decision in the EU

Similar to multitiered reimbursement lists (formularies) in the US, many European countries operate so-called positive lists,

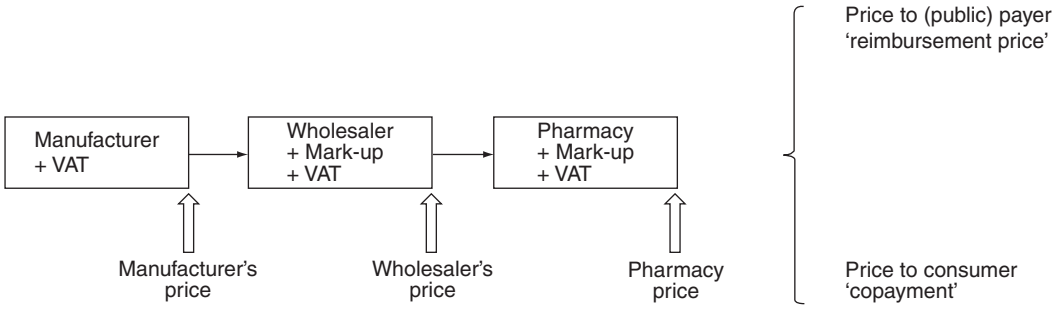


Figure 1 Various levels of 'price' in the pharmaceutical market.

i.e., lists containing all pharmaceuticals that are reimbursed up to a certain amount or for a certain percentage of the pharmaceutical price. The main difference as compared with the US is, however, the scope of a reimbursement/coverage decision. Although a negative decision by one health plan in the US affects only coverage of a drug for patients in that one (regional) plan among many, a negative decision in a European country usually applies for the whole public system, i.e., it excludes that pharmaceutical from reimbursement for the entire population; except for those few with private health insurance. However, there are exceptions in decentralized healthcare systems in Europe, such as in Spain and Italy, where a decision is only valid for a region or where a centralized decision may be overruled by a region.

Some countries, for example, the UK and until recently Germany do not regulate the manufacturer prices directly and only exert influence on the reimbursement price by means of health technology assessment, negotiated rebates, or reference pricing. Other countries, however, combine the reimbursement decision with statutory pricing or price negotiations, for example, France, Italy, and Spain. For those countries that regulate prices of pharmaceuticals, it is common to only regulate prices if manufacturers apply for reimbursement. Also, if a country operates public and private healthcare systems side by side like in Germany, regulation of prices may refer to both systems, whereas reimbursement regulation usually refers to only public health insurance.

The reimbursement process usually starts with the manufacturer applying for reimbursement before the launch of their product (e.g., France, Italy, and Spain). Exceptions apply, as products may already be sold while the application is being submitted or evaluated (e.g., Germany and Austria). In some countries, the decision-making procedure is not linked to market launch directly but may be initiated at any point of time by the regulator (UK: England). The process itself may include two components (**Figure 2**):

1. A pricing decision by the regulator. Either by (a) an agency/authority that sets list prices, i.e., statutory pricing or (b) by negotiation of list prices between the manufacturer and an agency/authority.
2. The reimbursement/coverage decision, i.e., the decision on who will get access to the pharmaceutical (patient population and conditions) and on the amount of reimbursement or the percentage of the list price being reimbursed. Again, the decision may include negotiable (e.g.,

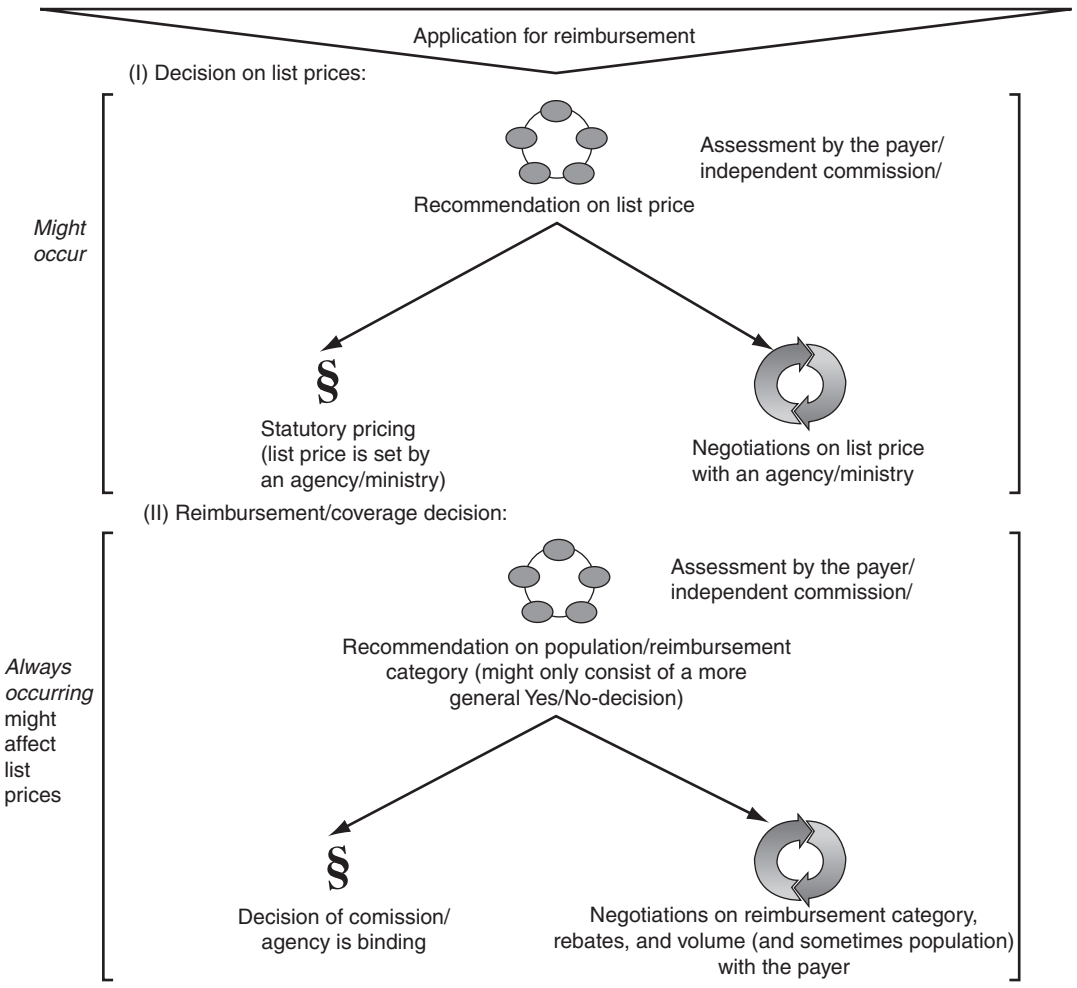


Figure 2 The process of a reimbursement decision.

negotiations on a rebate between the manufacturer and an agency/authority on list prices) or nonnegotiable elements.

Both decisions are usually based on some kind of benefit assessment of the pharmaceutical in question. Sorenson (2010) has found in her comparison that nearly all the EU countries apply health technology assessment to some degree to classify pharmaceuticals according to the therapeutic improvement they deliver. Although the UK applies cost-effectiveness analysis directly that results in a Yes/No decision on the reimbursement status of a pharmaceutical for a particular indication, countries such as France and Germany determine therapeutic improvement (added medical benefit compared with existing therapies) and later on negotiate rebates or prices in order to guarantee value for money. According to a systematic literature review by Erntoft (2011) and comparisons of reimbursement decisions by Blankart *et al.* (2011), other criteria such as budget impact and type of disease (acute vs. chronic and common vs. orphan) may also play an important role when making reimbursement decisions, especially when deciding on the percentage of list price that is to be reimbursed. Sometimes, the therapeutic value of a pharmaceutical itself (not in comparison to existing therapies) or the underlying product properties, i.e., whether a drug prolongs life, whether it improves, or whether it only maintains the patient's condition are used as well.

Demand- and Supply-Side Regulation Structure

Apart from the initial reimbursement decision and classification into a reimbursement category, additional reimbursement regulation exists in all EU countries to influence price, prescription volume, quality of prescribing, or total spending. These approaches are often categorized as demand-side interventions, i.e., interventions that target pharmaceutical consumption at the point of utilization (patients, physicians, and hospitals), and supply-side interventions, i.e., interventions that target pharmaceutical consumption at the point of production or service delivery (manufacturers, wholesalers, and pharmacies) (**Table 1**).

Table 1 Overview of the regulation on the demand and supply-side

Supply-side (industry, wholesalers, and pharmacies)	*Demand-side (patients, physicians, and hospitals)*
Price controls	*Price controls*
Based on clinical performance	Copayments
Based on economic performance	(Enforcement of) generic substitution
Costs of existing treatments	(Enforcement of) parallel trade
Costs (cost-plus regulation)	
Price in neighboring countries (external reference pricing)	*Volume controls/quality of prescribing*
Prices of comparable products (internal reference pricing)	Patient education/information
Temporary price freezes	Prescriber education/ information
(forced) rebates on price/ tendering	Prescription guidelines
Regulation of pharmacy/ wholesaler margins	Prescription monitoring/ auditing
	Prescribing targets (with monetary incentives)
	Classification of products as Rx or OTC negative lists
Spending controls	
Payback/clawback mechanisms	*Spending controls*
Price-volume agreements	Drug budgets
Rebates/discounts	Payback/clawback mechanisms
Risk-sharing agreements	
Industrial regulation	
Profit controls/rate of return	
Tax benefits	

Source: Reproduced from von der Schulenburg, F., Vandoros, S. and Kanavos, P. (2011). The effects of drug market regulation on pharmaceutical prices in Europe: Overview and evidence from the market of ACE inhibitors. *Health Economics Review* **1**, 18.

Internal Reference Pricing

Reference pricing is one of the most common supply-side measures used in many European countries. The measure refers to reimbursing the same amount for a group of comparable pharmaceuticals (so-called reference pricing cluster), i.e., the difference between the list price and the reference price is borne by consumers as a copayment. Manufacturers whose products are priced above the reference price will usually lower their list prices to the reference price (**Figure 3**). In this case, the consumer is not affected by reference pricing at all as all drugs in a reference pricing cluster are available without an additional copayment.

Although the basic idea behind reference pricing – paying the same amount of money for the same therapeutic value being delivered – sounds reasonable, whether all pharmaceuticals put into one reference pricing cluster are therapeutically equivalent remains a highly debated question.

Generally, it is important to differentiate between (1) therapeutic reference pricing and (2) generic reference pricing. Although in generic reference pricing a common reimbursement limit is established for the off-patent originator and its generics (same active ingredient), therapeutic reference pricing includes all pharmaceuticals deemed to be comparable by the regulator, usually based on mechanism of action, pharmacological properties such as duration of action and form of administration, and/or indication. Depending on the design of the reference pricing scheme, it is possible that a group includes several originator drugs in the same therapeutic category and their generic versions if available. Although generic reference pricing is usually limited to the off-patent market, therapeutic reference pricing may also include pharmaceuticals still on patent.

The method of determining reference prices varies between countries. Countries use the lowest or average of the few lowest prices in a group of pharmaceuticals or define the reference price to ensure that a certain number of products is available at or below the reference price. Depending on the method used, freedom of choice for physicians among fully reimbursed pharmaceuticals is being more or less limited. For example, in Germany, the reference price is set in such a way

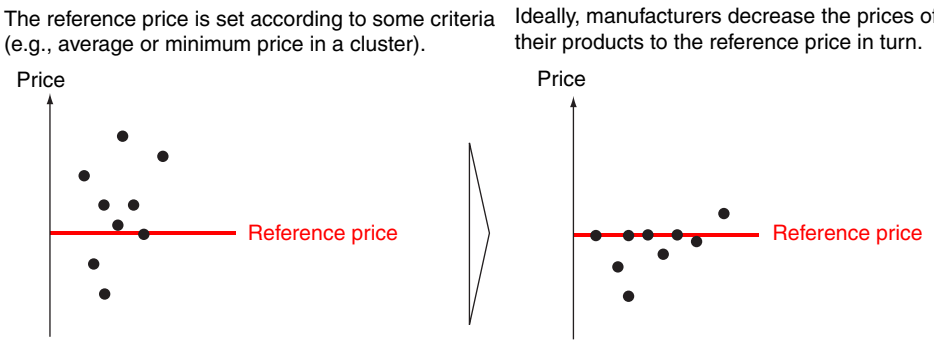


Figure 3 The basic concept of reference pricing within a group of comparable pharmaceuticals.

that roughly one-third of all packages in one group are available at or below the reference price. Although some countries use complex mathematical formulas to determine a reference price per package, others, for example, Hungary, determine the reference price as the lowest price per defined daily dose.

Galizzi *et al.* (2011) have found in their systematic review on reference pricing that the result of generic reference pricing has been consumers switching to less costly products of the same active ingredient, which usually also leads to increases in the market share of generics in general as well as a decrease in drug expenditure. In the case of therapeutic reference pricing, switching may not always occur to less costly drugs within a group of pharmaceuticals but also to close substitutes that are not subject to reference pricing. Schneeweiss *et al.* (2003, 2004) found a decrease in expenditure of CAD$ 6.7 million for angiotensin converting enzyme inhibitors (i.e., 6% of total expenditure for all cardiovascular drugs before the policy change) and CAD$ 1.6 for calcium channel blockers (CCBs) (i.e., 12% below projected expenditure for CCBs from the prepolicy trend) in the first 12 months for the Canadian province Ontario, respectively, whereas Stargardt (2010) found net savings after considering costs due to switching between €94 million and €108 million in the first 12 months for statins in Germany (i.e., between 7.7% and 8.8% of prepolicy expenditure for statins).

Although the majority of studies, for example Schneeweiss *et al.* (2003), do not find evidence of patient's health being affected by policy-induced drug switching, there is also some evidence of increased health-care utilization that points in the opposite direction. Stargardt (2010) found more hospital visits, especially due to cardiovascular disease, for the subgroup of patients that switched their medication more than once following the introduction of reference pricing for statins. However, given the different designs of reference pricing schemes, the variability of manufacturers' reactions to regulation and differences in the study design with regard to measuring the impact of reference pricing, a final drawn conclusion is perhaps unwarranted by the few Canadian and the one German study that use patient-level data to measure healthcare utilization after the application of reference pricing to a drug class.

Regarding the effect of reference pricing on drug prices, Galizzi *et al.* (2011) have found evidence in support of a reduction in prices in the first instance of forming a reference pricing cluster to be dominant in the literature. The long-term results of reference pricing on drug prices, however, are not very clear. Stargardt (2011) provides evidence from Germany that suggests that reference pricing also generates a long-term downward trend in prices, whereas other studies, such as Kanavos *et al.* (2009), argue that the reference price also works as a price floor, below which no producer has an incentive to decrease their price. Again, these differing results are unsurprising, given differences in design of reference pricing systems. In general, however, there is evidence of increased generic competition under reference pricing.

External Reference Pricing

In addition to internal reference pricing, i.e., referencing to prices of other pharmaceuticals within a country, the majority of EU countries also applies external reference pricing, i.e., referencing to prices of the same pharmaceutical in other countries. According to a review article on external reference pricing by Leopold *et al.* (2012), 25 countries, nearly all except for Denmark, Sweden, and the UK, make use of this instrument to some extent. The degree to which information on foreign countries' prices is used differs: It can be used either as the main criterion or only as supportive information in pricing and reimbursement decisions. Also, external reference pricing may be applied either to newly launched pharmaceuticals until the completion of Health Technology Assessment only (e.g., Austria) or to all pharmaceuticals on the market.

The design of external reference pricing varies greatly among the EU countries: some countries define the lowest price (e.g., Hungary and Poland), the third lowest price (e.g., Greece and Bulgaria), or an average of the six lowest prices (e.g., Slovakia) of a basket of foreign prices – as their reimbursement limit – whereas others use an average of all foreign countries' prices included in the basket (e.g., Austria and Ireland), or the price of the country from which the drug has been imported (Luxembourg). Furthermore, according to Leopold *et al.* (2012), the number of countries included in the index varies considerably between 2 (Latvia) and 26 (Slovakia). Portugal, Ireland, and the Netherlands reference only neighboring countries, whereas Austria, Greece, and Slovakia do so to almost all of the EU-25 nations. Generally, however, basket composition seems to be in accordance with the rank of each country according to gross domestic product

(GDP), i.e., countries seem to reference countries with similar levels of GDP.

The extensive use of external reference pricing may, on one hand, create cross-border spill-over effects of price reductions that are welcomed by policy makers and may reduce health-care costs in general. On the other hand, Stargardt and Schreyögg (2006) and Richter (2008) argue that cross-border spillover effects also provide strong incentives for strategic product launches as well as for launching delays and lobbying activities and can affect the effectiveness of regulation or the final decision of a manufacturer to launch pharmaceuticals at all. In particular, Danzon and Epstein (2009) show that external referencing can lead to both reduced access to new drugs and higher prices in lower income countries in the EU.

Sometimes, both types of reference pricing are applied simultaneously or in different instances, for example, external reference pricing is used before or while a reimbursement decision is being taken, whereas internal reference pricing is applied if the molecule loses patent protection. Eventually, the increased use of external reference pricing may contribute to convergence of prices within Europe. Whether this increases or decreases social welfare remains a subject for future research, given the differences between European countries in both consumer preferences and willingness to pay.

Copayments

Compared with the US, copayments in Europe are of minor importance. Historically, copayments were initially introduced as a prescription fee that restricts pharmaceutical consumption, i.e., volume, and does not exert influence on price. Later on, the instrument evolved, as it became partly dependent on prices. Today, many EU countries operate a system of multiple reimbursement categories (positive lists), each of which is subject to a different level of reimbursement, i.e., 90%, 80%, 70%, etc. It is part of the initial coverage decision taken at launch in each market, and thereby also determines copayments. Unlike the US system, there are typically no differences between 'preferred' and 'nonpreferred' brands within a therapeutic category. For example, although one or two of the eight statins may be in the preferred brand tier of a US drug plan (with the others being in the nonpreferred tier or the tier for generics), the same reimbursement category would be valid for all statins reimbursed in nearly all healthcare system in Europe. Thus, the same percentage of reimbursement prices will be reimbursed whether it is branded or generic.

In countries that apply internal reference pricing there might be two types of copayments: (a) prescription fees or price-dependent copayments as defined by the reimbursement category and (b) additional copayments, if the list price of a pharmaceutical is above the reference price, i.e., the difference between the list price and the reference price, regardless of the reimbursement category.

As many studies have found that copayments impact adherence, European healthcare systems usually limit the total annual copayments for (a)-type-copayments to a fixed percentage of the individual's annual income, or exclude vulnerable groups such as severely diseased, retired, or children from copayment on general principle. However, when patients do anticipate that they will be spending more than their annual deductible, the marginal costs of pharmaceutical consumption are reduced to zero and patients will be price-insensitive. This makes those who consume the most nearly price inelastic as they are either generally exempted or can expect to hit the maximum annual copayment threshold anyway. Thus, besides their contribution to financing pharmaceutical care – with the exception of some of the eastern European countries – the impact of (a)-type-copayments on consumption behavior is in those subgroups that consume pharmaceutical the most, rather small. Also, individuals may adapt to the level of copayment after sometime. Lessons from behavioral science and psychology, for example, adaption-level theory have shown the tendency of people to adapt to stimuli over time. Thus, there may be decreased effectiveness of copayments in reducing demand over time.

In contrast to (a)-type-copayments, (b)-type-copayments are seen more as the result of private luxury, i.e., the consumption of a health-care product is considered nonessential as there are always other comparable alternatives with no (b)-type-copayments available. Consequently, such surcharges are not included within annual limits. Thus, this clearly impacts choices of all consumers. For example, when reference pricing was applied to statins in Germany in 2005, this resulted in (b)-type-copayments for atorvastatin whose manufacturer kept the list price above the reference price between €18 and €109 per package (in addition to the 'regular' (a)-type-copayments of €5 per package). As a result, the market share of atorvastatin among all statins fell from 33.2% to 6% within 1 year according to Stargardt (2010).

Spending Caps, Drug Budgets, Prescribing Targets, and Prescription Guidelines

Supply-side regulation also includes measures to contain the volume of prescribing and – as regulators would call it – measures to influence quality of prescribing. Prescription guidelines aim to influence the choice between different active ingredients, clearly differentiating first- and second-line treatment options. This is related to the concept of a preferred product within a group of comparable pharmaceuticals, albeit not being connected to financial incentives. The use of the instruments ranges from nonbinding guidance for therapy choice, to enforceable rules as laid down in national law with electronic prescription monitoring being employed.

As a result of a systematic review of studies by Sturm *et al.* (2011), drug budgets that affect physicians have been found to decrease drug expenditure per prescription, to decrease expenditure per patient, and to decrease prescribed volume. Also, in order to prevent under prescribing in regions where there is a larger concentration of elderly or severely diseased, budgets need to take into account various demographic and epidemiological factors. Andersson *et al.* (2009) have found decentralized drug budgets to result in a higher degree of cost awareness by physicians as compared with a more centralized approach. Germany, for example, employs drug budgets at the physician level. The so-called practice-specific targets of cost control and appropriate prescriptions have been implemented since 2002.

Physicians exceeding 125% of the prescription target are required to compensate the sickness funds, unless they prove the necessity of their prescriptions from a medical viewpoint.

In the UK, physicians' individual prescribing behavior is being compared with national and regional levels. Specifically, data on prescription volume and costs of individual GP practices is collected and statistics are disseminated quarterly to encourage awareness of prescription volume and costs. Monitoring activity results in physicians being notified if they exceed regional averages. Similar systems exist, according to Vogler *et al.* (2008), for example, in France, Germany, the Netherlands, Sweden, and Italy. Sometimes, financial incentives, i.e., bonuses or penalties, are connected to prescribing targets in order to influence regulatory compliance of physicians. If electronic, auditing allows physicians to be constantly aware of their prescribing patterns. However, it also permits authorities to follow-up or intervene in cases of overprescription or underprescription. In any case, prescription monitoring systems rely on comprehensive electronic medical records that allow for comparisons at the regional level.

As continuous physician education and information by health authorities is also vital, prescription guidelines are an important tool to assist physicians with the appropriate/efficient choice of treatment in all countries. By excluding treatments or differentiating between first-line and second-line use of medication, differences in cost effectiveness are accounted for. If compliance with prescription guidelines becomes mandatory, then the use of a drug shall be limited to cases in which its utmost value is delivered according to the decision makers.

Parallel Trade and Promotion of Generic Competition and Substitution

Price regulation at the national level has generated differences in pharmaceutical prices across EU member states. In the presence of a Single European Market, within which circulation of goods is free, arbitrage opportunities occur. As the free movement of goods cannot be restricted, agents can capture economic rents by buying products in low-price countries and selling in other EU countries where prices are higher. Although the parallel traded products are the same as locally sourced ones and produced by the same manufacturers, parallel trade makes the pharmaceutical industry lose the difference between the local price in the importing country and the price in the exporting country, which is captured by parallel traders minus any transportation and (possibly) repackaging costs. Parallel trade may thus reduce the ability of manufacturers to price-differentiate across EU markets.

Parallel trade may lead to savings for health insurances in destination countries in the short run. The long-term effects are, however, debatable: Although some of the previous studies, for example, Ganslandt and Maskus (2004), have shown parallel trade to encourage competition besides significant reduction of manufacturers prices, others have shown that parallel traded products are usually sold at the same level as that of locally sourced ones or just below. Kanavos and Vandoros (2010) have found evidence that manufacturers use nonpricing strategies to respond to parallel trade and that the induction of price convergence by parallel trade seems to be upward rather than downward. Also, there is a discussion on whether parallel trade may lead to welfare losses because of a reduced incentive toward innovation.

For the low-price EU countries, parallel trade may lead to supply problems in the local market due to manufacturers' strong incentive to restrict and/or control volume of sale to local wholesalers. For example, after price cuts following the financial crises in Greece, parallel exports have led to local shortages of some drugs, resulting in a ban on exports of particular products. Such bans may swiftly resolve such problems, but authorities have to ensure that the policy measure is enacted early enough, before shortages are actually observed.

Some EU countries encourage parallel trade by sharing savings from lower prices with the distributional chain or by obliging pharmacies to dispense parallel imports, if available and cheaper by law. For example, in the Netherlands, health insurance shares any price differences between locally sourced products and parallel imported ones with the pharmacies. Thus, some of the savings resulting from increased competition are being transferred to the distribution chain. In Germany, pharmacies and the Federal Association of Sickness Funds even negotiate a minimum parallel import quota in order to exploit the full gains to the public payer.

Although parallel trade refers to exploiting price differences of an active ingredient by a manufacture across markets, generic competition exploits price differences between multiple producers of the same active ingredient within a country. At least for compounds with a large market size, alternative producers of the same active ingredient (generic manufacturers) enter the market on patent expiry. Sometimes, a so-called 'authorized generic' may enter 3–6 months before patent expiry through licensed production by the former innovator. This way, a generic manufacturer may exploit a first-mover advantage whereas the former innovator will extract some of the rents generated by the generic competitor.

In a study on pricing in the US, the UK, France, and Germany, Magazzini *et al.* (2004) have found a tendency for prices of branded products to decrease over time after generic entry in the highly regulated European markets whereas a recent study by Vandoros and Kanavos (2013) has shown opposite effects. Lu and Comanor (1998) have found the intensity of price competition to depend on the therapeutic area, type of disease (e.g., acute vs. chronic), and the launch price. In the US, however, according to Grabowski and Vernon (1996) and Frank and Salkever (1997), the prices of branded drugs have increased after generic entry, whereas the prices of generic drugs have decreased over time.

Generic penetration varies widely in EU countries as do policies to stimulate generic competition. Uptake seems to depend on (1) the price of the branded original and the difference between the price of the branded original and the price of generics that also determines the number of generic competitors in a market, (2) regulation that encourages or discourages the use of generics, for example, whether prescribing by international nonproprietary name(s) rather than brand name being mandatory or by the degree of generic substitution being mandatory at the pharmacy level, (3) the type of drug, i.e., when brand loyalty appears to be larger for some

disease area than for others, and (4) the enforcement of generic policies.

Rebates, Price-Freezes, Price-Volume Agreements, Tendering, and Risk-Sharing Agreements

Rebates or discounts on pharmaceutical prices are also commonly used instruments in EU countries. They can be imposed (1) on the national level by law (also known as price cuts or forced rebates) or (2) by following negotiations between manufacturers and payers at various levels, i.e., at the federal level, the regional level, or for each sickness fund. For example, in response to the financial crisis, Greece has commissioned two rounds of price cuts with a weighted average of 21.5% and 10.2%, according to Vandoros and Stargardt (2013). Germany too applies forced rebates for generics or pharmaceuticals not subject to reference pricing. The so-called price-freezes, i.e., the obligation of a manufacturer not to increase prices for a predefined time period, or the so-called price-volume agreements have been used in combination with price cuts and forced rebates. For example, in France, price-volume agreements on therapeutic drug classes are used to impose spending caps on pharmaceutical expenditure. Compliance with budget constraints is guaranteed by a clawback mechanism in case of overspending.

Competitive tendering has been extensively used for the procurement of drugs in inpatient markets of EU countries. In European outpatient drug markets, according to Kanavos *et al.* (2009), this policy tool is being implemented in Germany and the Netherlands only. In generic tendering, producers submit their bids (prices); one or more of the cheapest product(s) gain reimbursement status or at least some kind of preferred supplier status in return. The policy creates conditions of extreme competition between manufacturers, as only the cheapest product(s) is reimbursed by health insurance in that particular (reimbursed) segment of the market. Therefore, tendering can lead to very aggressive price competition, leading to prices close to the per-unit cost of production.

In the Netherlands, tendering was initially only used for three off-patent products (omeprazole, pravastatin, and simvastatin) in 2005. According to Kanavos *et al.* (2009), price reductions were in the range of 88% (omeprazole), 84% (simvastatin), 85% (amplodipine), 88% (citalopram), and 93% (alendroninezuur) of wholesaler prices. Consequently, the number of products subject to tendering was expanded. In the long run, however, there is a threat that winner-take-all tendering may eventually reduce generic competition: Manufacturers who systematically fail to win any bids may go out of business. If this happens, prices may experience an increase later on, although they are unlikely to reach their initial level. In addition, as pharmacies are remunerated on the basis of a markup of the manufacturer price, pharmacists' income may experience a significant decrease following the implementation of tendering.

One of the more recent forms of agreement between payers and pharmaceutical manufacturers is risk-sharing agreements. In these types of pay-for-performance arrangements, prices, and/or reimbursement are linked to patient outcomes, i.e., manufacturers have to refund a negotiated percentage of prices if predefined treatment goals are not met. This may be related to disease progression, progression-related death, or unacceptable toxicity of a drug.

On the one hand, Cook *et al.* (2008) argue that risk-sharing is a way to increase the confidence of payers in higher prices of innovative drugs. It thus provides an additional option in negotiations for manufacturers and payers that can increase overall efficiency by overcoming risk aversion of payers. On the other hand, this may involve high implementation costs. Barros (2011) argues that the monitoring of a drug's effectiveness will sometimes require special documentation by the treating physician or the dispensing pharmacist, and a suitable data infrastructure. Also, pharmaceutical companies are most likely to adjust prices upwards before negotiation in order to cover cost of failure in some of the patients. They thus see risk-sharing as a means to expand treatment to groups, which do not benefit as much from an innovation as requested by the regulator. Overall, there is still little proof of their long-term performance available.

Concluding Remarks

Owing to the need to limit pharmaceutical expenditure growth for public health insurance in European countries, the regulatory environment of pharmaceutical markets is constantly changing. In many European countries, a 'major reform' of pharmaceutical policy takes place almost on an annual basis. Given the costs of adapting to regulatory interventions by patients, physicians, and the pharmaceutical industry, the evaluation of goal attainment of policy measures in terms of cost containment, budget impact, effects on the provision of health services, and effects on the health of affected individuals, all are of great importance. Often, short-run effects of a policy such as cost-saving may be accompanied by unwanted long-run implications, for example, the worsening medication compliance or lower incentives for R&D. If not evaluated on a regular basis, the design of future measures cannot be based on past experience, which may – unnecessarily – waste scarce resources and harm patients.

See also: Adoption of New Technologies, Using Economic Evaluation. Biopharmaceutical and Medical Equipment Industries, Economics of. Medical Decision Making and Demand. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Pharmaceutical Parallel Trade: Legal, Policy, and Economic Issues. Pharmacies. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Rationing of Demand. Value of Information Methods to Prioritize Research

References

Andersson, K., Carlsten, A. and Hedenrud, T. (2009). Prescribing behaviour after the introduction of decentralized drug budgets: Is there an association with employer and type of care facility? *Scandinavian Journal of Primary Health Care* **27**, 117–122.

Barros, P. P. (2011). The simple economics of risk-sharing agreements between the NHS and the pharmaceutical industry. *Health Economics* **20**, 461–470.

Blankart, C. R., Stargardt, T. and Schreyögg, J. (2011). Availability of and access to orphan drugs: An international comparison of pharmaceutical treatments for pulmonary arterial hypertension, Fabry disease, hereditary angioedema and chronic myeloid leukaemia. *PharmacoEconomics* **29**, 63–82.

Cook, J. P., Vernon, J. and Manning, R. (2008). Pharmaceutical risk-sharing agreements. *PharmacoEconomics* **26**, 551–556.

Danzon, P. M. and Epstein, A. J. (2009). Launch and pricing strategies of pharmaceuticals in interdependent markets. *NBER Working paper.* Cambridge, MA: National Bureau of Economic Research.

Drummond, M. and Towse, A. (2012). Is it time to reconsider the role of patient co-payments for pharmaceuticals in Europe? *European Journal of Health Economics* **13**, 1–5.

Erntoft, S. (2011). Pharmaceutical priority setting and the use of health economic evaluations: A systematic literature review. *Value in Health* **14**, 587–599.

Federal Union of German Associations of Pharmacists (ABDA) (2012). Zahlen, Daten, Fakten. Available at: http://www.abda.de/fileadmin/assets/ZDF/ZDF_2010/mwst_auf_arzneimittel_32_33_2010.jpg (accessed 12.08.12).

Frank, R. G. and Salkever, D. S. (1997). Generic entry and the pricing of pharmaceuticals. *Journal of Economics and Management Strategy* **6**, 75–90.

Galizzi, M. M., Ghislandi, S. and Miraldo, M. (2011). Effects of reference pricing in pharmaceutical markets. *PharmacoEconomics* **29**, 17–33.

Ganslandt, M. and Maskus, K. E. (2004). Parallel imports and the pricing of pharmaceutical products: Evidence from the European Union. *Journal of Health Economics* **23**, 1035–1057.

Grabowski, H. G. and Vernon, J. (1996). Longer patents for increased generic competition in the US. The Waxman-Hatch Act after one decade. *PharmacoEconomics* **10**(supplement 2), 110–123.

Kanavos, P., Seeley, E. and Vandoros, S. (2009). Tender systems for outpatient pharmaceuticals in the European Union: Evidence from the Netherlands, Germany and Belgium. *Report to European Commission, DG Enterprise and European Medicines Information Network (EMINet).* Available at: http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/study_pricing_2007/tendering_systems_en.pdf (accessed 12.08.12).

Kanavos, P. and Vandoros, S. (2010). Competition in prescription drug markets: Is parallel trade the answer? *Managerial and Decision Economics* **31**, 325–338.

Leopold, C., Vogler, S., Mantel-Teeuwisse, A. K., et al. (2012). Differences in external price referencing in Europe: A descriptive overview. *Health Policy* **104**, 50–60.

Lu, Z. J. and Comanor, W. S. (1998). Strategic pricing of new pharmaceuticals. *Review of Economics and Statistics* **80**, 108–118.

Magazzini, L., Pammolli, F. and Riccaboni, M. (2004). Dynamic competition in pharmaceuticals. Patent expiry, generic penetration, and industry structure. *European Journal of Health Economics* **5**, 175–182.

Richter, A. (2008). Assessing the impact of global price interdependencies. *PharmacoEconomics* **26**, 649–659.

Schneeweiss, S., Dormuth, C., Paul, G. and Soumerai, S. B. (2004). Maclure Malcolm. Net health plan savings from reference pricing for angiotensin-converting enzyme inhibitors in elderly British Columbia residents. *Medical Care* **42**, 653–660.

Schneeweiss, S., Soumerai, S. B., Malcolm, M., et al. (2003). Clinical and economic consequences of reference pricing for dihydropyridine calcium channel blockers. *Clinical Pharmacology and Therapeutics* **74**, 388–400.

Sorenson, C. (2010). Use of comparative effectiveness research in drug coverage and pricing decisions: A six-country comparison. *Issue Brief* **91**, 1–14.

Stargardt, T. (2010). The impact of reference pricing on switching behaviour and healthcare utilisation: The case of statins in Germany. *The European Journal of Health Economics* **11**, 267–277.

Stargardt, T. (2011). Modelling pharmaceutical price changes in Germany: A function of competition and regulation. *Applied Economics* **43**, 4515–4526.

Stargardt, T. and Schreyögg, J. (2006). Impact of cross-reference pricing on pharmaceutical prices: Manufacturers' pricing strategies and price regulation. *Applied Health Economics and Health Policy* **5**, 235–247.

Sturm, H., Austvoll-Dahlgren, A., Aaserud, M., et al. (2011). Pharmaceutical policies: Effects of financial incentives for prescribers. *Cochrane Database of Systematic Reviews* 10. Available at: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD006731/abstract

Vandoros, S. and Kanavos, P. (2013). The generics paradox revisited: Empirical evidence from regulated markets. *Applied Economics* **45**(22), 3230–3239.

Vandoros, S. and Stargardt, T. (2013). Reforms in the greek pharmaceutical market during the financial crisis. *Health Policy* **109**(1), 1–6.

Vogler, S., Habl, C., Leopold, C., et al. (2008). Pharmaceutical pricing and reimbursement information report (PPRI) – Comparative report. *PPRI Report.* Vienna: European Commission, DG Health and Consumer Protection, and Austrian Federal Ministry of Health, Family and Youth.

Pharmaceuticals and National Health Systems

P Yadav and L Smith, University of Michigan, Ann Arbor, MI, USA

Background

Pharmaceuticals used to treat a variety of health conditions continue to be a critical aspect of quality healthcare. However, consistent access to medicines in many low- and lower-middle income countries persists as a major challenge. Although there has been a global increase in proportion of gross domestic product (GDP) spent on health, there are significant inequalities in the spending on pharmaceuticals across countries. When looking at the share of pharmaceutical expenditures to total health expenditure (THE) by country income group, one can see large differences between the mean spending in high-income countries, compared to low-income countries. This may be reflective of multiple factors including disease burden, health system infrastructure, differences in cost of service delivery and health policies (Table 1).

In 2006, 1.5% of the global GDP represented pharmaceutical spending. Pharmaceutical expenditures are negatively associated to GDP by income group with total pharmaceutical expenditure (TPE) as a share of THEs ranging from a mean of 1.41% in high-income countries to a mean of 1.62% in low-income countries. Lower-income countries spend a greater percentage of their total health costs on pharmaceuticals relative to their GDP (Table 2).

Further, among lower-middle and low-income countries, the private sector is where a large share of pharmaceutical expenditures are made. A considerable portion of this spending is based on a community's ability and willingness to pay for health products. In many low- and middle-income countries numerous individuals purchase pharmaceuticals using out-of-pocket (OOP) monies. Up to 50% of THEs are made using OOP monies of which, up to 90% go toward the purchase of medicines. The opposite trend appears to be true for high-income countries where health insurance and other financing or pricing policy may be in place (Table 3).

According to IMS Health, the largest segment of global spending growth for pharmaceuticals is expected in pharmaceutical markets in emerging economies such as China, Brazil, India, Russia, Mexico, Turkey, Poland, Venezuela, Argentina, Indonesia, South Africa, Thailand, Romania, Egypt, Ukraine, Pakistan, and Vietnam or pharmaceutical markets in emerging economies. Population growth, new healthcare reforms, and economic growth are expected to increase spending in these markets. In 2006, global spending on pharmaceuticals within pharmaceutical markets in emerging economies was 14% compared to developed markets including the European Union, Japan, the US, Canada, and South Korea. In 2011, this share increased to 20% and in 2016, it is predicted to increase to 30%. Given the forecasted increases in global spending, ensuring consistent access to medicines will remain a central focus for healthcare stakeholders. This article will outline the traits of pharmaceutical sectors within low- and lower-middle income country health systems. Further understanding of

Table 1 TPE share of THE, 2006 (%)

Income group	*Mean (%)*[a]	*Median (%)*	*Minimum*	*Maximum*
High	19.7	18.2	8.7	32.4
Upper middle	23.1	22.0	10.4	36.8
Lower middle	27.6	26.6	9.8	67.6
Low	30.4	29.5	7.7	62.9

[a]Weighted mean by population.

Source: Adapted from Lu, Y., Hernandez, P., Abegunde, D. and Edejer, T. (2011). *The world medicines situation 2011: Medicine expenditures*, p. 6. Geneva, Switzerland: World Health Organization; WHO NHA database.

Table 3 Per capita TPE by income group, 2006 (%)

Income group	*TPE*	
	Public (%)	*Private (%)*
High	61.3	38.7
Upper middle	38.8	61.2
Lower middle	33.5	66.5
Low	23.1	76.9

Source: Adapted from Lu, Y., Hernandez, P., Abegunde, D. and Edejer, T. (2011). *The world medicines situation 2011: Medicine expenditures*, p. 7. Geneva, Switzerland: World Health Organization.

Table 2 TPE and THE as a percentage of GDP by income group, 2006 (%)

Income Group	*TPE*					*THE*	
	N	*Mean (%)*	*Median (%)*	*Minimum (%)*	*Maximum (%)*	N	*Mean (%)*
High	46	1.41	1.40	0.30	2.70	49	11.3
Upper middle	37	1.45	1.30	0.40	2.70	54	6.4
Lower middle	44	1.63	1.45	0.40	3.80	47	4.4
Low	34	1.62	1.50	0.40	3.60	41	5.3

Abbreviation: *N, number of countries.*

Source: Adapted from Lu, Y., Hernandez, P., Abegunde, D. and Edejer, T. (2011). *The world medicines situation 2011: Medicine expenditures*, p. 8. Geneva, Switzerland: World Health Organization; WHO NHA database.

these system traits will enable strategic reform and improvements in this sector.

Pharmaceuticals and Health Systems

Health systems vary in form, but irrespective of the form of the health system, pharmaceuticals play a critical role within it to prevent and treat health conditions. It is important to understand the organization of each country's pharmaceutical sector in order to ensure consistent access to pharmaceuticals. Elements of financing, procurement, distribution, and service provisions must be effectively aligned to reach patients with medicines. Each of these components contributes to the overall effectiveness and efficiency of the health system.

Efficiency in a health system is a metric by which performance may be measured. Health system efficiency includes both technical efficiency – the method for producing a good or service at minimum cost – and allocative efficiency – the right collection of outputs provided in a health system to achieve overall health improvement goals. Evidence shows that in health systems with similar health expenditures per capita, technical and allocative efficiency help explain differences in population health outcomes (Figure 1).

Technical efficiency of the pharmaceutical subsystem within national health systems implies achieving best pharmaceutical related health outcomes at the lowest cost. The degree of technical efficiency varies with the structure by which financing is collected and potentially pooled to be used to procure medicines. For example, pooling arrangements may optimize technical efficiency of national health systems through the purchase of larger volumes of medicines at the most competitive prices. Allocative efficiency for pharmaceuticals in national health systems consists of product selection and resource allocation decisions that ensure medicines of greatest need and health benefit are available.

Technical efficiency and allocative efficiency couple together sustainability and equity goals of the health system. Further discussion of the elements impacting health system efficiency will be reviewed in individual detail in the sections that follow.

How Are Pharmaceuticals Financed?

In the overall health system the financing function typically: (1) collects revenue from multiple sources, (2) pools funds and spreads risks across groups, and (3) allocates funds to purchase goods and services. Within the general structure of health financing, pharmaceuticals may be financed through a variety of mechanisms.

Out-of-Pocket Spending

OOP spending typically makes up a large portion of financing for medicines in developing countries. Individuals make OOP payments most often for outpatient and chronic care. Spending of this kind typically impacts lowest income households, usually part of the informal economy, to the greatest degree. OOP can account for up to 50% of total healthcare expenditure in low- and middle-income countries, whereas in higher income countries this amount is estimated to be much less (approximately 15%). Of OOP expenditures, up to 90% are spent on medicines.

OOP spending may be individually financed from personal savings or through borrowing funds and accepting debt. In a study of 15 African countries done by Leive and Xu in 2008,

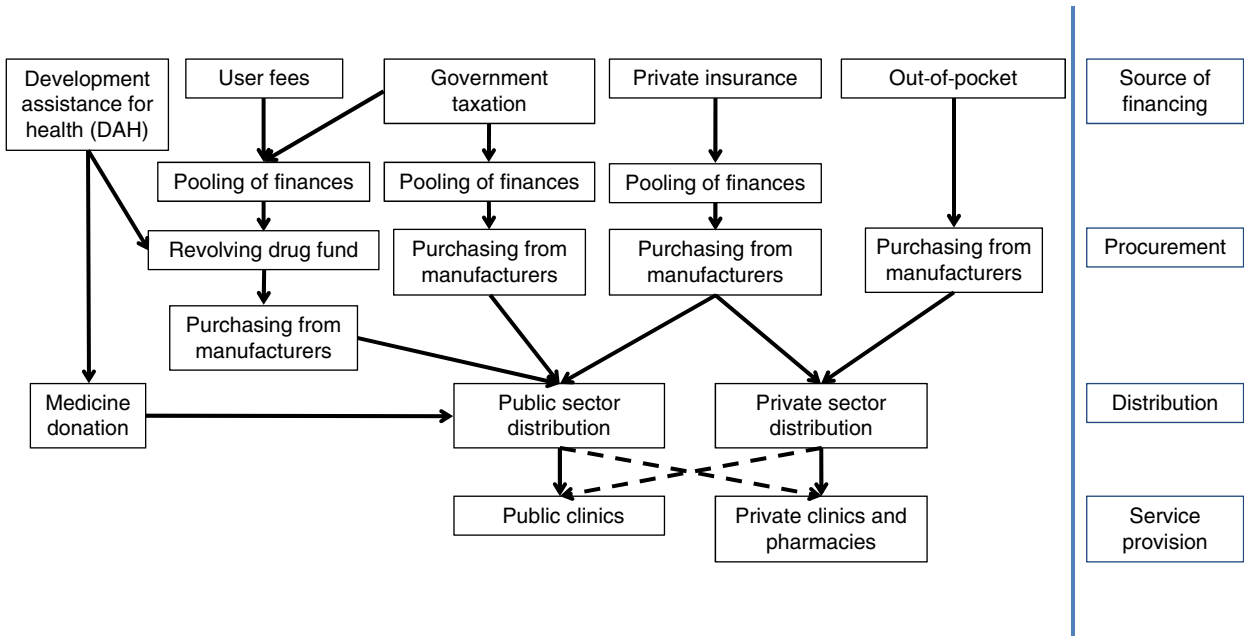


Figure 1 Pharmaceuticals in the overall health system.

approximately 30% of all households surveyed financed their OOP health expenditures by borrowing or selling assets. This form of healthcare financing is largely dependent on the populations' willingness and ability to pay for medicines.

Private Prepaid Funds

Prepayment of healthcare services and medicines is another financing mechanism available in certain contexts. Prepayment schemes are usually managed by health plans within countries. For example in South Africa, a separate healthcare company Yarona Care has developed prepayment schemes at discounted rates. Patients may purchase vouchers in advance for their expected health needs and redeem them throughout the year. Vouchers are priced to save community members money and guarantee them a network of quality service providers. This system presumes individuals have sufficient funds and are willing to pay for services in advance. Within populations with a high burden of chronic diseases, the expectation to pay for services and medicines throughout the year may be greater. Within populations with fewer long-term conditions to manage (or poor awareness and few available services for such conditions) and/or a larger emphasis on episodic care, individuals may be less inclined to purchase care in advance as their expectations for healthcare costs may be less certain.

The prepayment method of healthcare financing provides a structure to help improve healthcare resource allocation and planning at an individual level. It may also integrate technology to minimize administrative management and associated costs of voucher utilization as with the South African care model.

Revolving Drug Funds

Revolving drug funds (RDFs) involve a one-time investment of capital utilized to develop a self-sustaining medicine supply system. In low-income countries capital funds are often provided by external organizations as part of development assistance for health (DAH). Once established, RDFs are reimbursed through the sale of drugs. Funds are pooled before placing new orders of medicines directly with suppliers. RDFs create a consistent pool of monies with buffer financing to ensure timely procurement of essential medicines and consistent access for patients. RDFs act as a separate fund of money protected from fluctuations in available government resources and shifting political priorities. A fund of this nature does necessitate its own administrative management as well as cooperative coordination between invested stakeholders and funders. RDFs are a financing strategy utilized by multiple national healthcare systems in low- and lower-middle income countries to promote access to medicines.

Ensuring success of an RDF requires thoughtful consideration of the country context. Previous research has outlined a set of guidelines essential for a successfully implemented RDF. Where elements are lacking within these guidelines, further preparation or planning may be required before establishment of a revolving drug fund and/or alternative methods for improving access should be examined.

Private Insurance

Private healthcare insurance is typically paid for or provides care directly to employees by their employment firm. Employers offer insurance as a benefit to their employees, however, employees typically still pay copayments on medicines and premiums for services. The risk pooling that takes place is usually dependent on the size of the employer and/or the size of the insurance company offering a health plan to the employer. Employee-based insurance plans are less common in lower- and lower-middle income countries as many individuals work in the informal economy.

Private health insurance funds, separate from employer-supported insurance are voluntary and typically less common in most developing market contexts. Individuals contract with an insurance entity that pools the risk of all members. In general, individual private health insurance funds have low membership, low contributions, low coverage and weak regulatory environments. Private health insurance may be organized as a nonprofit entity or a for-profit entity. Typically nonprofit entities charge premiums just as for-profit entities. Nonprofit insurance plans are often arranged by religious groups, civic groups, hospitals and physician associations. Namibia and South Africa represent two countries where private-for-profit health insurance is relatively common. Private-for-profit health insurance providers represent the predominant prepaid plans in these contexts. For-profit insurance plans are funded through equity from private stakeholders as well as premium payments by enrollees. Uptake may be low as a result of the inability to afford annual fees and/or the perception that services and practitioners available for service are poor quality and so advanced payment is seemingly less appealing. Individuals may rather purchase healthcare only when necessary in these instances.

Community Health Insurance

Community-based health insurance (CBHI) is a voluntary insurance mechanism in which an organization coordinates a community of payers in order to pool risk to cover all or part of healthcare costs. At times, the organization may be a buyers' cooperative managed by representatives of the community. Studies examining the impact and effectiveness of CBHI in low-income countries have found that there is little evidence to support community-based insurance as a viable healthcare financing mechanism. Although successful in certain specific contexts, in general CBHI is not able to sufficiently mobilize financial resources. There is some evidence that suggests CBHI relieves some of the burden of OOP for patients and increases utilization of healthcare. Even with CBHI, ensuring benefits reach the lowest income community members continues to be a challenge in healthcare financing. This is a specific population group where access to care is only marginally affected by CBHI as many individuals cannot afford premiums or contributions to the insurance scheme.

In general, CBHI does not address barriers to accessing health care (i.e., affordability, perceptions of care quality, and geographic distance to healthcare facilities). Additionally, reimbursement processes tend to be burdensome for plan participants. Certain specific examples demonstrate that dropouts of CBHI membership are common. The result of this is a

smaller risk pool, which may in turn have a negative impact on the attractiveness of the health plan and future enrollment. Further research is needed to systematically summarize the characteristics of contexts where CBHI has shown a larger impact.

Social Health Insurance

Social insurance plans are a universal coverage health financing program in which membership is required for a population and members are provided a nationally determined benefit package of care. The World Health Organization has recently promoted social health insurance (SHI) plans as a means to reduce the burden of OOP on lower-income community members. Social insurance programs are typically financed through mandatory contributions from workers, self-employed, enterprises, and government. Most programs determine levels of contribution based on income along with contribution ceilings. Unemployed individuals within certain contexts are typically covered by others' contributions or by government assistance. Health risks are typically pooled across larger populations than with community-based insurance programs and private insurance (employer-based, nonprofit, and for-profit) programs and over a longer period of time in lower- and lower-middle income countries. A prerequisite to SHI is having a sufficient portion of a given population participate in the formal economy. In lower- and lower-middle income populations that have larger informal economies, payments into a national SHI may place a higher burden on the formal economic members.

The costs of running a SHI program are greater, with complex administrative, allocative and accountability mechanisms. Variations of SHIs exist across Europe, Latin America, and parts of Asia where the programs have been instituted. In these contexts, entire populations are included for coverage or coverage may be more selective with medicine coverage provided to a portion of individuals. Some SHI arrangements allow for individuals to opt out voluntarily whereas others cannot afford to include all individuals and so selectivity becomes financially necessary. Social insurance programs typically determine a formulary of covered services and essential medicines which the program will finance. These lists are often based on World Health Organization (WHO) recommended essential medicines, however, updates to lists may be slow with access to medicines potentially lagging behind healthcare need.

Taxation

This form of financing is collected in large pools of funds that are not controlled by consumer payments. For members of the formal economic sector, indirect taxation through purchases and direct taxes on income may be an effective way to collect revenue to fund public healthcare. Indirect taxes may include taxes of goods or services (e.g., alcohol or tobacco purchase taxes) or taxes on lotteries and betting. Direct taxes are typically taxes on personal income, business profits and transactions, imports and exports, and property. A portion of taxes is typically allocated to healthcare through the ministry of health.

Balance of Power in Procurement: Manufacturer, Payer, and Patient

In national procurement arrangements the balance of power between manufacturers and purchasers is more often seated with the manufacturer especially for patented medicines. For medicines that are only available from a single source (i.e., patented products or in some cases single registered manufacturer in the country), both pricing (i.e., affordability) and production (i.e., availability) are influenced to a great degree by the decisions of the monopolistic manufacturer.

In developing country markets there is often fragmentation of orders among multiple purchasing groups. This results in lack of coordination and significant wastage of resources (Figure 2).

In some instances the opposite may also be true (i.e., a single large buyer has influence over supplier(s)). In contexts where one group is responsible for the majority of purchasing for a specific product, the single buyer may be better able to negotiate prices with the manufacturer. With multisource products (i.e., products with more than one manufacturer supplying to the market) the purchasing body achieves the best possible prices (Figure 3).

In Australia and many countries in Europe, government is the principal payer for pharmaceuticals and the market resembles a monopsony. In Australia, the Pharmaceutical

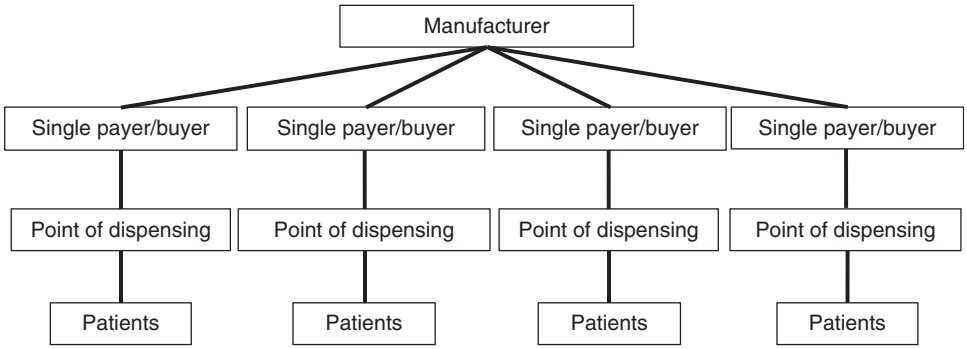


Figure 2 Manufacturer monopoly in pharmaceutical markets. *Point of dispensing refers to facilities such as hospitals, health clinics, and retail pharmacies.*

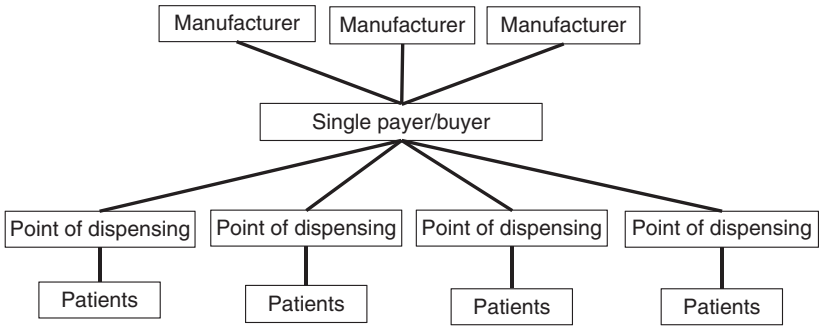


Figure 3 Buyer/payer monopsony in pharmaceutical markets. *Point of dispensing refers to facilities such as hospitals, health clinics, and retail pharmacies.*

Benefits Scheme (PBS) selects appropriate pharmaceutical reimbursement levels. In the US, Pharmacy Benefit Managers (PBMs) help align the balance of power between manufacturers and purchasers. PBMs provide services to health plans including price discount negotiation with retail pharmacists, rebate negotiation with manufacturers, and managing mail-order prescription services and claims processing systems. PBMs also help health plans develop appropriate medicine formularies, review prior authorization for specific products and substitution scenarios for generic versions of brand name drugs. In their relationships with both retail pharmacies and pharmaceutical manufacturers, PBMs manage payment for retail and mail order drugs. Acting on behalf of health plans, PBMs have been able to obtain discounted prices and improved process efficiencies. In other country contexts, more than one, but still a small number of payers may purchase from suppliers. This would most closely resemble a monopsony with the balance of power still with the payers; however, a continuum of different procurement scenarios and related balances of power exist between the two extremes, monopoly and monopsony.

In many low- and middle-income countries pooled procurement creates a balance of power between buyers and suppliers. These procurement arrangements influence affordability and the end patient price of drugs by pooling orders at a global level from multiple groups or countries and negotiating orders through bids from multiple suppliers. Pooling arrangements may take place within country or across multiple countries. Various pooling schemes are outlined in the section that follows.

Pharmaceutical Procurement and Distribution in the Public Sector

Medicine purchasing within the public sector may take place in multiple ways. Typically, procurement of medicines is based on epidemiological needs, funds availability, and sales and stock information collected from service facilities. Many countries rely on a registered list of essential medicines to determine which pharmaceuticals ought to be procured or reimbursed. Procurement varies based on the health system context and in many countries where health infrastructure continues to evolve so too does the procurement and distribution architecture. In health systems that are more centralized, often in geographically smaller countries, procurement is typically done by the ministry of health or a centralized agency acting on behalf of the ministry of health. In health systems that are decentralized, often larger countries with more extensive healthcare infrastructure, procurement of pharmaceuticals may be appropriated to the state or equivalent regional level.

National Pharmaceutical Procurement

Brazil provides an example of decentralized procurement of medicines. In Brazil, the Regulatory Council of the Pharmaceutical Market (CMED) approves medicines prices and adjusts products available on the market annually. The CMED helps to regulate purchasing prices paid by the government for medicines included in the Ministry of Health's essential medicines list. Procurement within the Brazilian Unified Health System (SUS) is entirely decentralized and performed by the Federal Union, 5564 municipalities, 26 states and the Federal District. Current evidence suggests that this fragmented administrative structure although allowing certain flexibilities may lead to allocative inefficiencies of financing. Specifically, in municipalities with smaller populations, procurement of medicines is often more expensive because of lower negotiating power on smaller quantities purchased.

Several strategies have been recommended to mitigate these procurement inefficiencies including development of pharmaceuticals in public production laboratories, creation of consortiums of municipalities to engage in small-scale pooled procurement, predetermined pricing regulation that is consistent across states and centralization of purchases for pharmaceuticals at the national level for products manufactured by single provider and/or those that have the most expensive pricing and/or products that require importation. Of these strategies, municipal consortiums for medicine procurement have been implemented and examined in southern Brazil. The Intermunicipal Health Consortium (CIS-AMMVI) improved access to medicines by reducing the purchase price and the number of stockouts. These benefits may impact smaller municipalities most as they are able to reach economies of scale and better negotiate prices in a larger tender process.

Similar to the strategy to set pricing standards across states or equivalent regions, Mexico has implemented the Coordinating Commission for Negotiating the Price of Medicines (CCPNM) and other health inputs. This national-level entity coordinates across public health institutions to collect background information including economic documentation to assist in annual negotiations for public procurement prices for patented medicines. Pricing negotiations are reported to save the country substantial financial resources; however, the political will and sustainability of this group may be less certain in the future. A coordination commission of this nature requires ongoing political support, appropriate performance indicators, and predetermined methods for assessing impact and transparency with key stakeholders.

Several centralized procurement models were setup in India at the state level to ensure consistent access to medicines. The Delhi Model Drug Policy pooled procurement for all hospitals within the state with a storage and distribution center. This policy not only organized procurement but it also pushed for implementation of standard treatment guidelines and a standard essential list of medicines, which fed into the development of a formulary for the state. In addition, a nongovernmental organization (NGO), Delhi Society for the Promotion of Rational Use of Drugs (DSPRUD) was contracted to implement technical activities related to the state policy. The Delhi model was largely seen as a success increasing access to medicines in many government hospitals.

Similarly, in Tamil Nadu, the state instituted a centralized drug purchase organization, the Tamil Nadu Medical Services Corporation Limited (TNMSC). TNMSC was developed to create a systematic method for streamlining the purchase, storage, and distribution of essential medicines in the public sector. TNMSC setup information systems including the provision of computers to warehouses for tracking stock in and out of storage and passbooks at clinics and hospitals to record inventory received. The TNMSC procurement gave structure to the previously fragmented purchasing structure. It laid out guidelines for the selection of suppliers, payment procedures, and standard essential medicines. This pooling mechanism has helped to purchase medicines at lower prices and to better ensure consistent availability of products. This model is now being used as a national benchmark for centralized procurement within each Indian state.

Centralized pooling of procurement for pharmaceutical products increases the level of influence a community of buyers may have on suppliers. Pooling higher volume orders enables suppliers to reach more efficient production at economies of scale. Buyers are better able to realize these efficiency savings because they are aligned and can negotiate in a unified fashion for lower prices. Pooling may also provide benefits to suppliers as they will be provided with a forecast of demand from a larger community of purchasers rather than relying on each individual tender. As a result, they may be better prepared with installed manufacturing and supply capacity to serve the needs of their buyers. However, in nationally centralized systems the administrative and management costs required to ensure information is collected and pooled in a timely fashion may be challenging and complex. Decentralized purchasing creates more flexibility among purchasers to order whenever necessary. The result is a procurement system that is more responsive to fluctuations in demand and may be better able to prevent stock outs. The tradeoff between the flexibility of decentralized purchasing and lower prices obtained through centralized purchasing should be considered according to each context.

Global Pharmaceutical Procurement Groups

Global pooled procurement of medicines enables countries to negotiate contracts with suppliers at the global level. As with pooling within countries, pooling orders across countries provides smaller countries or countries requiring fewer medicines for specific diseases with increased negotiating power. Joint procurement arrangements typically involve an organizing, intermediate buyer. Often, donor agencies will help facilitate this intermediate step either by setting up a new intermediary buyer as the Partnership for Supply Chain Management in the case of President's Emergency Plan for AIDS Relief or working with existing procurement groups as with vaccines for GAVI-eligible countries procured through United Nations Children's Fund (UNICEF) (Table 4).

Multiple actors in the pooling mechanism may introduce the same inefficiencies and delays in payment and shipments as with traditional individual tenders. In certain pooling arrangements, revolving funds are used to address delayed payments from buyers to ensure payment is assured to suppliers and medicines are received as needed. Revolving funds require additional management and may be relied on too heavily to fill in financing gaps. Global pooling arrangements may not be best suited to all contexts. Countries with limited domestic resources to purchase medicines (highly resource constrained) as well as contexts with large purchasing volumes and thereby individual purchasing power (i.e., Brazil, India, China) may not be well-suited to a multicountry pooling mechanism.

Setting Prices/Price Ceiling or Reimbursement Levels for Pharmaceuticals

Because pharmaceuticals constitute a large portion of the overall health expenditure, payers and governments use different levers for managing the prices of pharmaceuticals. The exact nature of the method used depends on the way the overall health system is organized.

The most direct form of controlling prices is a statutory price control used at the ex manufacturer, wholesaler, retailer, or some combination of these levels in the distribution chain. Most countries in Europe, Australia, Canada, and many Francophone African countries use such an approach. In many instances the health authorities set a price for a medicine based on the prices for that product in other countries in its region, income class, or countries with other similarities. For example, the prices in Greece are selected to be the average of the three lowest prices in the European Union (EU). In some countries medicine prices are set based on a comparison with medicines that have similar active substances. In India the price of a select group of medicines (scheduled drugs) is regulated whereas others are not regulated. Indonesia, South

Table 4 Examples of pooled procurement

Group name	*Characteristics*
UNICEF vaccine purchasing pool for GAVI-eligible countries	● GAVI-eligible low- and lower-middle income countries register with UNICEF to participate in pooled procurement primarily for vaccines ● Buyer specifies order needs with UNICEF and UNICEF responds with pricing estimate ● Buyer formally places order and pays full amount in advance ● UNICEF ensures appropriate flow of information to manufacturer(s) and monitors shipments to countries and distribution within countries ● Manufacturers deliver directly to countries
Organization of Eastern Caribbean States (OECS) (formerly Eastern Caribbean Drug Service (ECDS))	● Community of small island countries submits annual medicine, medical supplies and X-ray consumable needs to OECS ● OECS evaluates needs and selects tenders from prequalified supplier bids ● OECS awards annual contract and places orders directly with supplier ● Supplier ships directly to countries and OECS monitors delivery and quality ● Countries reimburse ECCB drug accounts upon receipt of medicines
Pan American Health Organization (PAHO) EPI Revolving Fund	● Central contracting model with Latin American and Caribbean countries specifying annual vaccine needs and submitting to PAHO ● PAHO evaluates needs and selects 1–3 supplier bids to fulfill needs ● PAHO submits prices to the buyer and the buyer confirms orders quarterly ● PAHO places confirmed orders with supplier quarterly and suppliers deliver directly to countries ● PAHO pays suppliers using a revolving fund within 30–45 days of delivery ● PAHO invoices countries and countries reimburse revolving fund within 60 days of delivery
Gulf Cooperation Council (GCC) Group Purchasing Program	● Group contracting model with countries in Persian Gulf region (7 countries) submitting pharmaceutical, vaccine, laboratory supplies and chemical needs for the next year to council group ● GCC combines information and shares tender information within prequalified suppliers ● GCC selects preliminary supplier bids and reports information to countries ● Countries are given 4 weeks to confirm or adjust orders ● GCC confirms final orders with suppliers and from then on suppliers work directly with countries utilizing purchase orders ● Supplier deliver directly to countries 1–3 times a year and countries pay directly to the supplier; suppliers pay GCC 0.5% fee
The Global Fund Voluntary Pooled Procurement Program (VPP)	● Procurement arrangement for Global Fund (GF) grantees with procurement agreement; voluntary for other GF countries ● Countries specify order details and delivery dates to GFVPP to receive quote ● GFVPP reviews and submits to prequalified procurement service agent (PSA) ● PSA invites bids from suppliers and submits price quotes to countries ● Countries review quotes and once approved, PSA prepares invoice and GFVPP disburses funds ● PSA confirms orders with suppliers and coordinates delivery to country; suppliers ship directly to countries ● Countries confirm orders in country and PSA reconciles accounts

Source: Reproduced with permission from Privett, N. and Yadav, P. (2012). Analysis of the procurement and pricing architecture for vaccines. *Working Paper*, NYU-Wagner School, New York, NY.

Africa, and many other lower middle-income countries also have such schemes.

Most countries in the EU regulate their wholesalers and retailer margins in addition to control on ex manufacturer prices. The wholesaler and retailer markup regulation takes multiple forms such as fixed percentage markup, fixed absolute markup, and regressive markup (i.e., the markup decreases with increases in the product price resulting in incentives for the channel to also stock and promote lower cost products). For example, in Spain for products with a selling price lower than €22.90 the wholesale margin is 10.3% of the price; for products priced higher than €22.90 and lower than €150, a margin of 6% on the portion of the price higher than €22.90 is charged; and for products above €150 a 2% margin is allowed on the part of the price over €150. South Africa and India use a single exit price regulation where the retail price is set once the price negotiations with the manufacturer are carried out.

A less direct form of price control is through the use of formularies, which specify which drugs will be used for which condition and the price to be paid for it. In return for putting a drug on the formulary the payer (insurance company or hospital) asks the pharmaceutical company to offer discounted prices. In some cases these pricing negotiations are

carried out on behalf of the payer by specialized organizations called PBMs.

Some countries such as the United Kingdom (UK) and Australia also use pharmacoeconomic assessments and health technology assessments (HTAs) to set the prices of new medicines. This involves a cost–benefit analysis of the new medicines relative to existing treatments. Based on HTA, recommendations are made on the price and reimbursement level for the medicine evaluated. The National Institute for Health and Clinical Excellence (NICE) of the UK is a pioneer in the use of such techniques.

More recently risk-sharing between the manufacturer and the payer is gaining ground in some countries such as the UK. Under such arrangements the pharmaceutical company gets a smaller price/reimbursement level at the start and as health outcomes are realized from the use of the product the remaining reimbursement is made. These are like payment by result schemes where part of the payment is based on outcomes achieved in practice. Numerous examples exist, but a noteworthy one is where the pharmaceutical company agreed to pay NHS if the product atorvastatin failed to reduce LDL-C levels to agreed targets and a risk-sharing program for Bortezomib used to treat multiple myeloma.

Domestic Production Versus Import of Pharmaceuticals

Economic and public health views on the issue of pharmaceutical access

Domestic production as a strategy to increase access through the reduction of production and shipping lead times, decline in importation costs and the development of the local economy is an important component of the ongoing debate around improving access to medicines. Within this discussion there is often tension between economic interests and public health interests. Domestic production of pharmaceuticals is thought of as a means to create new jobs and increase the skill-level of local communities. However, this industrial economic argument for local production may sometimes be at odds with public health interest to improve access, both availability and affordability, of medicines. If quality medicines cannot be produced efficiently and cheaply in the local context, local production may not be the best investment of resources.

It is reasonable to expect that countries would want to become self-sufficient with their production, especially as they have seen domestic pharmaceutical industries developed in other developing country markets (i.e., India, China, Brazil, South Africa). However, historical examples should be considered with thorough understanding of the current realities of global trade, regulation, international economics of the pharmaceutical industry as well as the aforementioned assumed tension between economic and public health interests. Global trade, more specifically patent policy as a part of the Trade Related aspects of Intellectual Property Rights (TRIPS) agreement, has been cited by many as a complex, resource-intensive area to navigate for developing country governance structures. Context-specific information will help to establish the case for national manufacturing self-sufficiency when appropriate.

Domestic Production Business Models

Pharmaceutical companies in low- and middle-income countries are broadly organized into four main business models. The first is a pharmaceutical subsidiary of a large multinational company. The locally situated business will manufacture branded products for local and regional markets. The second business model consists of generic manufacturers producing a large portfolio of generic drugs for the global market. These drugs typically meet global quality standards and are competitively priced. The third business model consists of domestic generic manufacturers with a national focus on operations. Most manufacturers that fall into this category produce drugs for their country of residence or neighboring countries. Some, but not all manufacturers meet good manufacturing practices (GMP) standards for their products. Small-scale local manufacturers make up the fourth business model. These manufacturers typically serve local or regional markets and often do not meet GMP standards. Sometimes, small-scale manufacturers may be owned or managed by a local NGO or large hospital group. The portfolio of medicines produced is often focused on fewer drugs.

In addition to the aforementioned business models, domestic production of drugs may be setup as a combination of models. The level of production in locally situated manufacturing facilities may also vary based on the governing business model. Typically, most domestic production of medicines in low- and middle-income countries focuses on formulation and packaging of products. Chemically synthesized products (i.e., the active pharmaceutical ingredient (API)) are typically purchased and imported for local formulation into a complete product. Chemical synthesis tends to be a more complicated process; however, many generic manufacturers serving the global market are able to produce APIs for sale to others and/or as a part of their drug production. At any level of production pursued by domestic manufacturers, large capital investments are required up front to finance initial production facility development and technology transfer. Joint ventures like Cipla Ltd. Joint venture with Ugandan manufacturer Quality Chemicals and GlaxoSmithKline's joint venture with South African manufacturer Aspen may provide some examples of ways domestic production may begin in new developing country markets.

Domestic Production Decision-Framework

The decision to pursue domestic production versus importation of medicines is dependent on quality costs, regulatory costs, size of the local market, competitiveness within the local market, availability of skilled manpower, and economic status of the country. In certain contexts, domestic production does not make economic sense and investments would be better made elsewhere (i.e., investment in healthcare infrastructure or stimulation of the existing local market). Successful domestic production requires a functional ecosystem to support business sustainability. An active national regulatory authority is needed

to manage quality reviews and enforcement of standards. Further, international trade regulations may require additional investment of resources, both time and money, to compile with global policy. To cover initial capital costs related to regulatory requirements and installation of capacity, significant market share and sales volumes are needed to reach economies of scale. Without these elements, domestically produced drug prices may remain high and will struggle to be competitive both locally and globally. To address the issue of access to medicines through domestic production, countries should conduct a thorough market analysis considering the current business ecosystem. Batson, Evans and Milstein developed a framework that may be used for pharmaceuticals and vaccines to determine the production model that works best given a country's market size, GDP per capita income and current technical capacity.

Similarly, strategic policy options, similar to those Seiter outlines for different market contexts should be considered to create the appropriate mix of domestic investments (Table 5).

How Are Medicines Distributed?

Medicine distribution may take place through supply systems that are run by governments or the public sector, the private sector and NGOs or faith-based organizations (FBOs). Within each of these supply chains, characteristics such as the type of commodity, geography, flow of finances, cost of commodities, public versus private treatment seeking will be different based on the country context. Generally, procurement, distribution and provision of pharmaceutical goods are managed across groups with involvement from public, private, and NGO/FBO sectors.

Private Sector Distribution

Private sector supply chains for medicines typically include a network of importers, wholesalers, sub wholesalers, pharmacies, and drug stores. In most emerging markets, pharmaceutical manufacturers sell products to national importers and wholesalers. Beyond the national level, there are often a large

Table 5 Policy options for different market contexts

Market context	*Proposed strategies*
Countries with sizeable home market and an existing competitive industry, such as India, Brazil, South Africa	● Strengthen governance and the regulatory framework to ensure that the local industry can produce for the global market ● Gradually abandon subsidies or preferential market access to expose local industry to global competition at a pace that allows it to adapt and become stronger ● Open the market to foreign companies willing to invest in the local industry, encouraging the introduction of new technology, and development of R&D capacity to further improve competitiveness
Countries with a medium sized or small home market, or an existing industry with questionable competitiveness	● Strengthen governance and the regulatory framework to ensure safety of drugs in circulation ● If the local industry is used to a protected environment, allow two to three years for adaptation, but implement a clear path toward full enforcement of GMP standards and a market that is open for globally operating competitors ● Encourage collaboration and mergers between companies; invite foreign companies to take over local manufacturers. The goal should be to eliminate substandard manufacturing but keep as many jobs as possible ● Assist the local industry in exploring export markets
Countries with a medium sized home market, no relevant local industry, but good infrastructure	● Focus resources on developing an efficient procurement system, distribution chain, and payment systems with incentives for rational use of pharmaceuticals ● Develop strong regulatory capacity to secure safety of drugs in circulation ● Explore willingness of larger global companies to invest in local pharmaceutical manufacturing, but consider the incremental costs of upgrading the governance and regulatory framework ● If subsidies are considered for industrial policy reasons, they should neither become a burden on the health budget nor lead to higher drug prices
Countries with a small or medium sized home market, or with limited infrastructure	● Focus resources on developing an efficient procurement system, distribution chain and payment systems with incentives for rational use of pharmaceuticals ● Develop strong regulatory capacity to secure safety of drugs in circulation ● Shut down substandard manufacturing operations; assess the possibility of changing the business model from manufacturing to pharmaceutical wholesale and distribution

Source: Reproduced with permission from Seiter, A. (2005). *Pharmaceuticals: Local manufacturing*. HNP Brief # 3. Washington, DC: World Bank.

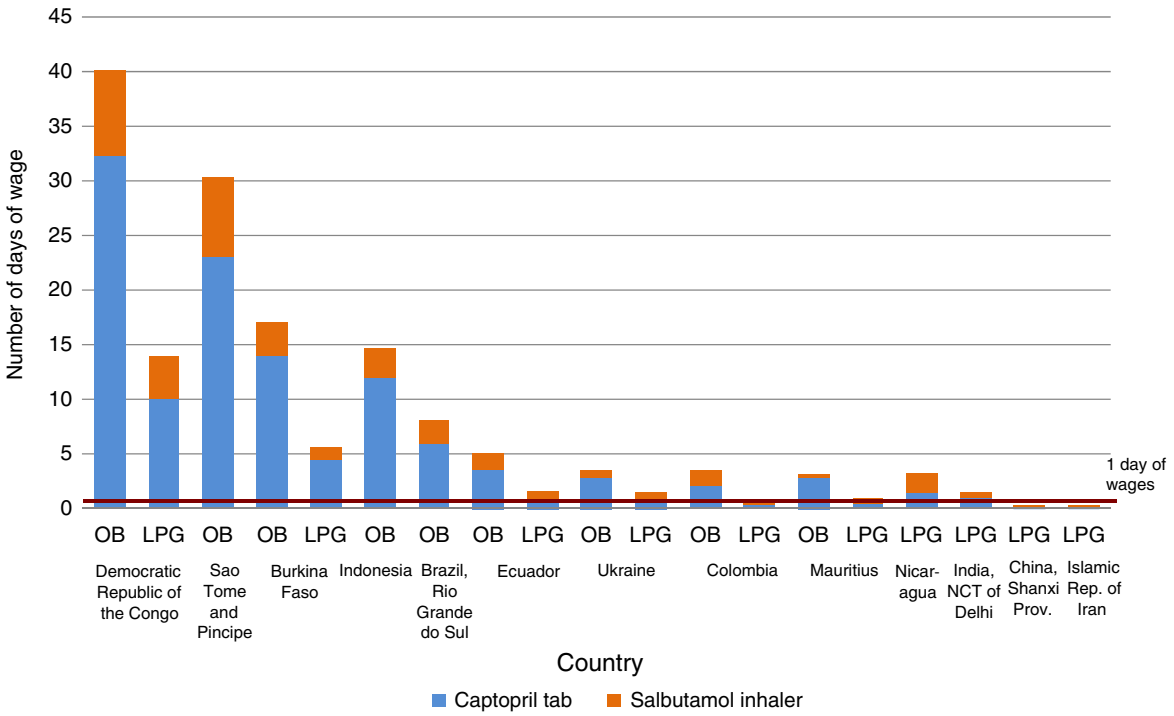


Figure 4 Number of days of wage income needed by lowest-paid government worker to pay for 30 days of drug treatment for an adult with hypertension and a child with asthma, between 2007 and 2011. *Abbreviation*: OB, originator brand; LPG, lowest-priced generic equivalent. Reproduced with permission from United Nations (2012). The global partnership for development: Making rhetoric a reality. *MDG Gap Task Force Report: Millennium Development Goal 8*, p. 65. New York, USA: World Health Organization/Health Action International survey data. Available at: http://www.haiweb.org/medicineprices (accessed 26.07.13).

number of intermediaries between manufacturers and patients (wholesaler, sub wholesaler/stockist, retailer). Many private sector wholesaler networks struggle to reach more remote communities and so they rely on others to increase their distribution networks. Markups at each tier of the supply chain results in higher overall markups on medications. As a result, a complicated private sector distribution chain may negatively impact affordability to the patient. This may in turn impact a manufacturer's ability to increase sales volumes and once economies of scale have been met, lower price points (Figure 4).

Fragmentation and opacity of information as a result of fragmentation across groups may cause poor coordination in distribution of medicines. Poor coordination impacts availability of health products as stock information may be poorly communicated; however, this may be less of a concern in the private sector, than in the public sector, as business profits may provide a greater incentive for better reporting in the private sector. In addition to availability, poor coordination may also impact the ability to track suboptimal and poor quality products entering the private sector medicines market (Table 6).

Table 6 Median availability of selected generic medicines in public and private health facilities in low-income and lower-middle income countries during the period 2007–11 (percentage)

	Mean	*Maximum*	*Minimum*
Public sector	50.1	87.1	21.2
Private sector	67.0	90.7	22.2

Source: Adapted from World Health Organization/Health Action International (WHO/HAI). Medicine price and availability surveys (2007–2011). Available at: http://www.haiweb.org/medicineprices (accessed 26.07.13).

Public Sector Distribution

In most public sector distribution systems, particularly in Sub-Saharan Africa, a centrally located warehousing and distribution point, often called a central medical store (CMS) manages the top tier of distribution. CMS then distributes medicines to regional or district stores depending on the geographic characteristics (i.e., the size of the country and relative distribution network) and product- or program-specific supply systems. Often donor funding for specific products and/or health programs creates multiple vertical supply systems in the public sector.

In many public sector supply systems, the risk of stockout at the health clinics is high because of skeletal distribution and reporting systems. As is such, availability tends to be lower in the public sector when compared to the private sector (see Table 5). In some cases, where pharmaceuticals are not free, increasing the prices of certain medicines covers distribution costs. This approach may lead to disparate access to certain medicines for which there are no vertical

Table 7 Mean number of day's wages of the lowest-paid unskilled government worker needed to purchase a course of treatment, by WHO region

	Africa	*Americas*	*Eastern Mediterranean*	*Europe*	*Southeast Asia*	*Western Pacific*
Adult respiratory infection; amoxicillin 250 mg capsule/tablet, three per day for 7 days						
Private sector OB	2.9	1.9	1.6	1.4	1.2	0.5
Private sector LPG	0.5	1.0	0.6	2.9	0.6	0.4
Public sector LPG	0.5	0.2	0.3	7.9	0.4	0.4
Diabetes; glibenclamide 5 mg capsule/tablet, two per day for 30 days						
Private sector OB	8.4	4.5	2.1	0.5	1.3	1.6
Private sector LPG	1.8	1.5	0.9	1.8	0.4	0.7
Public sector LPG	1.1	0.1	0.5	2.5	0.6	0.7
Asthma; salbutamol 0.1 mg/dose inhaler, 200 doses						
Private sector OB	4.4	2.0	1.6	3.6	1.2	1.4
Private sector LPG	2.5	1.0	0.8	5.0	0.6	0.7
Public sector LPG	1.6	0.6	0.7	15.0	–	1.1
Ulcer; ranitidine 150 mg capsule/tablet, two per day for 30 days						
Private sector OB	35.4	9.0	8.5	21.1	2.7	5.5
Private sector LPG	5.0	2.8	3.8	4.6	0.5	1.7
Public sector LPG	6.3	0.6	1.3	6.3	2.2	1.2

Abbreviations: OB, originator brand; LPG, lowest-priced generic.

Source: Reproduced with permission from Cameron, A., Ewen, M., Ross-Degnan, D., Ball, D. and Laing, R. (2008). Medicine prices, availability, and affordability in 36 developing and middle-income countries: A secondary analysis. *The Lancet* **373**(9659), 240–249.

funding systems. In general, medicines provided in the public sector are more affordable to patients when compared with prices to mean number of day's wages in the private sector (Table 7).

Nongovernmental Organization/Faith-Based Organization Distribution

NGOs and FBOs may also play an important role in distribution of pharmaceuticals in emerging markets. Distribution managed by NGOs and FBOs is often context specific, however, typically arranged according to a customer's own prearrangement, courier services, drug supply organization delivery services, or direct delivery services. Medicines are often purchased according to customer inventory needs (pull system) or given through prepacked kits of essential medicines (push system). The level of involvement of NGOs/FBOs sector varies considerably across countries.

Summary

Pharmaceuticals play an integral role in the prevention and treatment of a variety of health conditions. Consistent access to pharmaceuticals remains a challenge in many national health systems. This is despite increasing levels of healthcare investment through domestic expenditures and large increases in DAH for low-income countries. This article reviewed the attributes of pharmaceutical sectors within low- and lower-middle-income country health systems. Such analysis is important to ensure the long-term sustainability of national health systems and also to ensure that DAH investments have the intended impact of improving access to medicines.

An optimally designed health system will operate at a high level of technical and allocative efficiency. In this form, pharmaceuticals may be purchased and distributed at the lowest cost possible and the most appropriate set of pharmaceuticals will be provided to serve the needs of each specific population. To achieve these goals, elements of financing, procurement, distribution, and provision of pharmaceuticals must be effectively aligned.

For pharmaceuticals functions of collecting funds, pooling funds and spreading risks across groups, and allocating resources to purchase products often occur through a hybrid of multiple financing strategies. The most common forms of financing include OOP payments, private prepaid funds, RDFs, private healthcare insurance, CBHI, SHI, and government taxation.

With financing secured, pharmaceuticals must be purchased at prices to ensure affordability and long-term sustainability of the manufacturers. When thinking about different procurement structures it is important to consider the influence of different stakeholders in decision making. In decentralized models, purchasing power is distributed to a larger number of individuals within the health system. Decentralized procurement may provide individuals with more autonomy and flexibility, in turn lowering the number of stock outs. Conversely, decentralized procurement may also disempower smaller groups when negotiating lower prices with national or global suppliers. In nationally centralized systems the administrative and management costs required to ensure information is collected and pooled in a timely fashion may be challenging and complex. However, if done well, centralized systems with single payers, do improve the negotiating power of the payer as compared to the supplier. The tradeoff between flexibility with decentralized purchasing and lower

prices with centralized purchasing should be assessed according to each context. Beyond national pooling, international pooled procurement arrangements are often used to facilitate pharmaceutical purchasing in low- and lower-middle income countries.

During purchasing and once pharmaceuticals have arrived in country, payers and governments use different levers to manage the prices of medicines. Price controls utilized include creating a comparison pricing standard, regulating wholesale and retail margins on pharmaceuticals, directive formularies as a part of health plans and health technology assessments. Outcomes-based pricing and risk-sharing arrangements are also gaining popularity in countries such as the UK especially for expensive chronic care medicines.

Another option often considered by countries to reduce medicine costs is to develop a domestic market for pharmaceutical production. Domestic production is thought to increase access to medicines through the reduction of production and shipping lead times, decline in importation costs and the development of local economy. Although it is reasonable to expect countries to seek self-sufficiency, not every context can support a domestic pharmaceutical industry. The decision to pursue domestic production versus importation of medicines is largely dependent on quality costs, regulatory costs, the size of the local market, competitiveness within the local market, and the economic status of the country.

Once a source is identified and medicines have been purchased, distribution is the final step required to ensure access. Pharmaceutical distribution often takes place through a combination of public sector, private sector, and NGOs/FBOs. Fragmentation within each of these sectors and across sectors often equates to poor information flows and opacity in the distribution chain. Improvements and investment in national healthcare distribution systems may facilitate more consistent availability and affordability of pharmaceuticals.

See also: Pharmaceutical Company Strategies and Distribution Systems in Emerging Markets. Pricing and User Fees

Further Reading

Batson, A., Evans, P. and Milstein, J. B. (1994). The crisis in vaccine supply: A framework for action. *Vaccine* **12**, 963–965.

Cameron, A., Ewen, M., Ross-Degnan, D., et al. (2009). Medicines prices, availability, and affordability in 36 developing and middle-income countries: A secondary analysis. *Lancet* **373**(9659), 240–249.

Carrin, G. (ed.) (2011). *Health financing in the developing world: Supporting countries' search for viable systems*. Brussels: University Press Antwerp.

Ekman, B. (2004). Community-based health insurance in low-income countries: A systematic review of the evidence. *Health Policy and Planning* **19**, 249–270.

Gómez-Dantés, O., Wirtz, V., Reich, M., Terrazas, P. and Ortiz, M. (2012). A new entity for the negotiation of public procurement prices for patented medicines in Mexico. *Bulletin of the World Health Organization* **90**, 788–792.

Kanavos, P., Das, P., Durairaj, V., Laing, R., Abegunde, D. O. (2010). Options for financing and optimizing medicines in resource-poor countries. *World Health Report: Background Paper*, p. 34. Geneva, Switzerland: World Health Organization.

Kaplan, W. and Laing, R. (2005). *Local production of pharmaceuticals: Industrial policy and access to medicines. Health, nutrition and population discussion paper*. Washington, DC: The World Bank.

Lu, Y., Hernandez, P., Abegunde, D. and Edejer, T. (2011). *The world medicines situation 2011: Medicine expenditures*. Geneva: WHO.

Pauly, M. V., Zweifel, P., Scheffler, R. M., Preker, A. S. and Bassett, M. (2006). Private health insurance in developing countries. *Health Affairs* **25**(2), 369–379.

Ranson, M. K., Sinha, T., Chatterjee, M., et al. (2005). Making health insurance work for the poor: Learning from the Self-Employed Women's Association's (SEWA) community-based health insurance scheme in India. *Social Science and Medicine* **62**(3), 702–720.

Roberts, M., Hsiao, W., Berman, P. and Reich, M. (2008). *Getting health reform right: A guide to improving performance and equity*. New York: Oxford University Press.

World Health Organization (2005). Social health insurance: Sustainable health financing, universal coverage and social health insurance. *Fifty-Eighth World Health Assembly: Provisional Agenda Item 13.16*. Geneva, Switzerland: World Health Organization.

Yadav, P., Smith, R. and Hanson, K. (2012). Pharmaceuticals in the health sector. In Smith, R. and Hanson, K. (eds.) *Health systems in low- and middle-income countries: An economic and policy perspective*, pp. 147–168. New York: Oxford University Press.

Yadav, P., Tata, H. and Babaley, M. (2012). *The World Medicines Situation 2011: Storage and supply chain management*. Geneva, Switzerland: World Health Organization.

Pharmacies

J-R Borrell, University of Barcelona, Barcelona, Spain
C Cassó, University of Barcelona, Barcelona, Spain, and Northeastern University, Boston, MA, USA

Glossary

Double marginalization Phenomenon in which pricing of any good or service turns out to be excessive because not only manufacturers that have upstream market power overcharge in equilibrium equating its marginal costs and marginal revenue from its residual demand but retailers also overcharge as they can profit from their dominant position downstream equating again marginal costs to the marginal revenue from its residual demand. It is called double marginalization when upstream and downstream firms are separated. When both are vertically integrated in one unique firm, the phenomenon disappears and price turns out to be lower as the vertically integrated firm equates marginal cost to marginal revenue just once.

Over-the-counter (OTC) medicines Medicines that can be offered in pharmacies, or in some jurisdictions in other retail outlets, and that can be readily available to the public from the store shelves. Usually, these are medicines that have been in use for some time and which may only cause minor adverse effects when not used properly. There is usually enough competition among providers of such medicines among innovator brands, other brands, and generic alternatives. In many jurisdictions, direct to consumer advertising is only permitted for OTC drugs, but in others such as in the US, direct to consumer advertising is also permitted for pharmacy-only and prescription-only medicines.

Pharmacy-only medicines Medicines that can only be obtained from a pharmacy outlet, but which do not require a doctor's prescription. Such medicines may be dispensed only with the advice and assistance of a pharmacist. Usually, these are medicines that have been available for some time and that have no serious adverse effects when properly taken.

Prescription-only medicines Medicines for which a doctor's prescription is a legal requirement.

Introduction

Dispensing medical drugs is a profession that combines the particularities of a professional service and retail industry. The focus here is on retail pharmacy, leaving aside the special character of hospital pharmacies.

First, pharmacists are responsible for a range of professional services including offering advice and assistance to those receiving their medication. Pharmacists are responsible for making sure that people receive their medication safely and professionally. Second, a network of pharmacies and other retail outlets is a major vehicle for public policies designed to make pharmaceuticals available and affordable to the public throughout a jurisdiction. However, different jurisdictions do this differently. There is no single 'template' policy model that fits all jurisdictions.

In the middle- and low- income countries, there is tension between policies aimed at safe and proper dispensing and those aimed at availability and affordability.

This article focuses on the policy challenges that arise in setting up the rules and regulations governing this professional trade, the tradeoffs that need to be evaluated when designing markets in health care and the role of health economics in promoting public policy.

The article is organized into four sections. After the introductory section, the key policy instruments that are available to policy-makers and societies are highlighted and described (Section Challenges and Policy Options). The focus is on the differences between the quality, entry, and price regulations that characterize policies based on a competitive market and on those that rely much more on a regulated command and control framework.

In Section Benefits and Risks from Regulating Pharmacies an assessment of the pros and cons of quality, entry, and price regulations is dealt with.

Challenges and Policy Options

Societies face a twofold challenge when organizing professional retail pharmacy markets.

First, the structure of the market should address the following to ensure that

1. the consumers get assistance and advice that is free from the personal interests of prescribers;
2. the consumers receive assistance and advice that is free from the personal interests of dispensers, particularly when a prescription is not filled; and
3. that the pharmacists and their assistants are professionals qualified in medicine and pharmacy.

Second, pharmacies are retail outlets that serve the cause of better access: through which professionals make affordable pharmaceuticals available to the local public. Key drivers of availability and affordability are the following:

1. Reimbursement arrangements: These should ensure that there are sufficient retail outlets or other distribution channels for pharmaceuticals across the country to dispense the medicines that are reimbursed by health care organizations.

2. Competition: Competitive pressure upstream helps retail pharmacists to get the best deal from wholesalers and manufacturers, with pharmacies ideally offering good deals to their customers and passing on to them most of the gain from the upstream competitive interaction.
3. Behavioral regulations: These ensure that providers, whether independent pharmacies or companies running pharmacy chains, do not abuse the public if they reach a dominant position in distribution and dispensing in their catchment geographical areas.

Given the objectives of safety/proper dispensing and availability/affordability, legislatures and governments regulate key elements of the industry. There are three well established traditions in high-income countries regarding the two main objectives. In most countries of continental Europe, the trade is generally reserved to licensed community pharmacies owned by independent pharmacists or by national, regional, or local governments. There are therefore rules limiting ownership of pharmacies to those who are licensed or employees of or contractors to the State. There are also rules that limit the number of pharmacies that each licensed pharmacist can own (usually a one-pharmacy-per-pharmacist rule). Moreover, in many countries, community pharmacies are subject to entry restrictions based on various tests of need.

In most European countries, chains are not permitted; vertical integration of retail pharmacies with medicine wholesalers or manufacturers is not permitted. Community pharmacies are obliged to sell the full line of authorized medicines in each country and deliver some health care services for other mandated health care providers.

By contrast, in the US and Canada there are no entry restrictions: there is free entry. Pharmacy chains owned by limited or public companies are common, although such companies contract licensed pharmacists to manage the service under stringent professional codes. There are also many independent pharmacies owned by professionals that compete with the company-owned chains. Both independent pharmacies and chains are obliged to sell the full line of authorized medicines and drugs in each country. However, they contract in a voluntary basis with mandated health care programs (Medicare and Medicaid in the US), and with health maintenance organizations, the extra services related to the reimbursement of drugs within the health plans.

The UK, Ireland, and the Netherlands adopt an intermediate regulatory stance. There is no formal regulation limiting the entry of pharmacies. Entry is effectively restricted by contracts with the mandatory health care organizations. Potential entrants do not effectively enter without securing a contract and such organizations award them only after considering carefully the incremental benefits and costs of new pharmacy openings.

In these three countries, pharmacies ('chemists' in the UK) are independent retail outlets that contract to fill prescriptions covered by the mandated health care organizations. They also dispense pharmacological products prescribed by physicians not paid by the mandated health care organizations, and they also sell other medical products not requiring a prescription, as well as a wide range of common hygiene and health-related products. Some have broadened their range to embrace photography and other chemical-based product lines, and even products and services having little or nothing to do with health or health care. These 'chemists' compete with other retail outlets that sell over-the-counter (OTC) drugs without doctor prescription, other medicinal products and any other product and service.

In the middle- and low-income countries, all types of regulation can be found depending on the historical circumstances of each. However, middle- and low-income countries often have unregulated retailers that effectively sell many types of pharmaceutical drugs usually with neither a pharmacist-manager on the premises nor qualified assistant. Such outlets typically neither satisfy the obligation to sell a full line of drugs, nor ensuring that they dispense prescription-only drugs only when a qualified prescriber prescribes them.

There are six policy instruments used by most countries in the world that distinguish these three different pharmacy models: (1) restrictions to practice: professional licensing; (2) ownership restrictions; (3) separation of prescribing and dispensing; (4) pharmacist management, supervision, and assistance; (5) zoning; and (6) price regulation.

Restrictions to Practice: Professional Licensing

Professional licensing in retail pharmacy, that is, restricting the practice of pharmacy to qualified professionals is the law in almost all countries across the globe.

Countries differ in licensing retail pharmacy according to how 'the practice of pharmacy' is defined: (1) whether, or not, the ownership of a pharmacy is part of the practice of pharmacy, and as such, it is reserved only to pharmacists; (2) whether the management of the pharmacy is part of the practice of pharmacy, and as such, it is also reserved only to pharmacists; and (3), whether dispensing of medicines is part of the practice of pharmacy and as such is also reserved only to pharmacists.

In high-income countries, pharmaceutical retailing is reserved to pharmacists. It includes professional owners of independent pharmacies and professionals, who should be hired to manage, organize, even supervise or assist part or all retailing of medicines at each one of the outlets of any company operating a chain of pharmacies. This restriction is commonly observed and generally enforced.

By contrast, in middle- and low-income countries there are gray or second-tier outlets that sell medicines. Such outlets have no pharmacist owning the outlet, or to manage, organize, supervise, or assist the dispensing of medicines.

Availability and affordability problems in middle- and low-income countries are so severe that the authorities do not enforce professional licensing regulations on the grounds that such enforcement may further reduce the reach of their weak distribution networks. Having a licensed pharmacist performing these duties is perceived as an excessive cost of servicing in the outlet, partly fixed and partly marginal, that would eventually increase the price of the dispensing service. This is particularly true when medicines are sold without any pharmacist involvement in the preparation of packages with convenient labels and patient information leaflets containing clear and comprehensible directions for use. Mexico is among the few countries in the world in which the law clearly allows

pharmaceutical products to be sold in retail outlets without any pharmacist managing, supervising, or assisting the dispensing. Pharmacist involvement is only mandated in the case of dispensing psychotropic drugs.

Ownership Restrictions

Many countries reserve the ownership of pharmacy retail outlets to professionals or the state. Some countries allow the competition of nonpharmacist-owned outlets (usually company chains) with independent pharmacies (owned by pharmacists). Others forbid professional pharmacists to own the pharmacies in which they work.

The question of pharmacy ownership is controversial. Those in favor of restricting ownership to pharmacists claim that the key to make the pharmacy trade professional is the mandatory membership of the pharmacist-owner to a professional organization. They claim that only by having the pharmacist-owner subject to the rules and supervision of the professional body will pharmacists advise and assist patients in buying their medication according to the standards of safety and proper dispensing.

Without such ownership restriction and professional supervision, the claim goes, the personal interests of pharmacist would supersede the interests of the patient. At the same time, they claim that the standards of safety and proper dispensing should be decided by these professional bodies to which pharmacist-owners are affiliated and not by other government or industry bodies, which may in turn again give priority to their interests before the patient interests.

The main argument against the restriction of ownership of pharmacies to licensed pharmacists is that pharmacists-owners can also give priority to their interests before those of their patients. Professional bodies, which are mainly controlled by pharmacy owners, do not always set up the appropriate standards of conduct, nor do they always enforce sanctions against those affiliated owners who misbehave. It may even be easier to supervise and enforce standards of conduct over company chains that are liable for the conduct of their managers and employees and over pharmacists employed to manage outlets in pharmacy chains.

There are different perceptions of legal enforcement: for example, about the effectiveness in different jurisdictions of the different mechanisms for setting standards of behavior, organizing external supervision, and enforcing sanctions when there is misbehavior. Ultimately, enforcement depends on how clear the set of rules is and how tough the judiciary is in enforcement and penalty.

Ownership restrictions are widespread in Europe. As many as 18 out of 27 EU Member States reserve ownership of any pharmacy to a licensed pharmacist. It is only the Netherlands (since 2000) and Ireland among the old EU Member States that have free pharmacy ownership, together with most of the new EU Member States: Poland, the Czech Republic, Estonia, Lithuania, Malta, Slovakia (since 2005), and Slovenia (recently deregulated). Ten of the 27 EU Member States have state-owned community pharmacies: Bulgaria, Cyprus, Czech Republic, Hungary, Italy, Luxemburg, Malta, Poland, and Slovenia. Sweden is the only case in which the entire pharmacy trade was reserved to a state-owned enterprise from 1971 to 2009. In 2009, the Swedish government started to sell part of the state-owned pharmacy chain in clusters and new private pharmacies have been established, some in joint ventures with multinational chains. The new chains compete with the remaining part of the state-owned pharmacy chain.

Among middle- and low-income countries, Tanzania, Kyrgyzstan, Uganda, and Guatemala have state-owned pharmacies and other special arrangements such as contracted independent pharmacies and franchised Non-Governmental Organization pharmacies. The Dominican Republic project 'People's Pharmacies' is a successful program sponsored by the government to make available and affordable essential generic medicines for low-income families in all state-owned and managed health care premises, including state-owned pharmacies and hospital pharmacies. Sixteen out of the 27 EU Member States allow pharmacists to own only one independent pharmacy. In the remaining 11, pharmacy chains are allowed and widespread: Austria, Ireland, Netherlands, and Sweden among the old EU Member States, and the Czech Republic, Estonia, Latvia, Lithuania, Poland, Romania, and Slovakia among the more recent ones.

Chains are widespread and compete with independent pharmacies in the US, Canada, Mexico, and the Philippines. South Africa (since 2004) and Kyrgyzstan have corporate managed care preferred pharmacy networks.

Ownership restrictions to just one independent pharmacy go together with the prohibition on vertical integration with wholesalers or manufacturers. By contrast, when ownership is free and chains allowed, it is very common for some chains to integrate vertically with wholesalers or manufacturers. In such cases, economies of scale and scope may be attained and passed on to consumers. Those economies may not be passed on to consumers when chains reach dominant positions in the distribution of the medicines of some affiliated manufacturers, or in the distribution of all medicines in some catchment areas.

In the EU, the courts have upheld ownership restrictions when they have been questioned as being contrary to the freedom of establishment in the internal market. Pharmacy is excluded from the rules of the internal market and the service directive only when member States reserve the pharmacy trade to a regulated health profession or a state-owned entity as a means of protecting health in the context of a mandated or public organized health care system.

Prescribing and Dispensing Separation

Separating prescription of medicines to doctors (or prescribing nurses) and dispensing to pharmacists is a way of reducing conflicts of interest. Most countries forbid doctors from dispensing drugs and classify medicines in one of the following categories, depending on the level of advice and assistance that pharmacist should provide:

1. Prescription-only medicines: Medicines only with the prescription of one authorized to prescribe.
2. Pharmacy-only medicines: Medicines that can be obtained only from a pharmacy outlet, but which do not require a prescription from an authorized practitioner.

3. Over-the-counter (OTC) medicines: Medicines that can be made readily available to the public in other retail outlets – as well as in pharmacies.

In the legislation, prescribing by a doctor or a nurse-prescriber and dispensing by a pharmacist is the general rule in Europe, the US, and Canada. However, Austria, Belgium, Czech Republic, Cyprus, Finland, France, Greece, Hungary, Malta, Slovenia, and the UK allow exceptionally doctors in rural areas to perform both the prescribing and dispensing of pharmaceuticals in order to make sure the availability of service in remote areas.

By contrast, many countries, particularly in Asia and Latin America, have dispensing doctors or integrated health centers and pharmacies. The problem with this integration is that dispensing doctors are influenced in their prescribing decisions by their personal financial incentives at dispensing. The Philippines have encouraged the separation of prescribing and dispensing since 1995, and fees have been designed to financially promote such separation since 2002. By contrast, South Korea has clearly mandated the separation of prescribing and dispensing. However, enforcement of prescription-only and pharmacy-only rules varies strongly across countries. In many higher-income countries, and in most middle-and low-income countries, dispensing prescription-only medicines without any doctor prescription is widespread.

In these cases, the advice of pharmacists is the key to making sure that the safety and proper dispensing of medicines to patients is paramount. In such settings, financial incentives also apply: there is a lot of evidence that pharmacists' financial interests have an impact on the kind of advice and sales services they provide.

Pharmacist Management, Supervison, and Assistance

Most countries require that a pharmacist should manage the service and be responsible for ensuring that the pharmacy (whether independent or within a chain) complies with all professional rules, regulations, and standards. They also require that at least one pharmacist is always present and in charge of supervising or assisting in the dispensing of pharmaceuticals.

This is easier said than done in independent pharmacies, particularly in 'mom-and-pop' pharmacies in middle- and low-income countries. It is not clear cut whether enforcement of the rules of the profession is easier in company-owned pharmacy chains as competitors and pharmacists unions usually track compliance, or when all pharmacies are independently owned and supervised by a professional body governed by pharmacy owners or independent experts.

As mentioned before in the section Restrictions to Practice: Professional Licensing, Mexico is among the few countries in which the law allows pharmaceuticals to be sold in retail outlets without any pharmacist managing, supervising, or assisting in their dispensing. Pharmacist involvement is required only when dispensing psychotropic drugs. Moreover, as previously observed, many low- and middle-income countries do not enforce the law requiring pharmacist management, supervision, and assistance when retailing medicines.

There is a cost–benefit tradeoff to be evaluated by policy makers: having a pharmacists manage, supervise, and assist in the dispensing of medicines is expensive, particularly in the countries that lack qualified professionals, and the pricing of medicines will reflect these extra costs.

Zoning: Restricting Entry and Location of Pharmacies by Tests of Need

In most of continental Europe, independent pharmacists are subject not only to tight licensing regulations that restrict the trade to licensed professionals but also to government regulations that limit the number of pharmacies that can be open to the public in any given catchment geographic area. In 17 of the 27 EU Member States, entry restrictions under the formal form of zoning, quotas, or distance regulations apply. Among these countries, Slovakia has deregulated entry since 2005 and Slovenia is deregulating it. Hungary experienced some deregulation of entry and reregulation between 2007 and 2010. Portugal introduced some less restrictive entry conditions in 2006.

When zoning is in place, pharmacies are authorized to enter after some needs test, usually when the population to be served reaches some specified threshold. Three EU Member States have explicit distance regulations: Greece, Hungary, and Spain. Another three Member States (the Netherlands, Ireland, and the UK) indirectly control the number and location of pharmacies by awarding contracts from national health services to a restricted number of community pharmacies. Among the Member States, only, Bulgaria, Cyprus, the Czech Republic, Estonia, Germany, and Poland do not restrict entry according to any population need test.

Outside the EU, Norway at one time restricted entry. Currently, entry is free but pharmacy market shares are restricted. Entry is also free in Iceland, the US, Canada, and the Philippines. In Latin America, there are cases like the Dominican Republic where minimum distance entry regulations are in operation, and in countries like Chile and Mexico where the number and location of pharmacies is freely determined. South Africa has a restrictive system as new pharmacies have to obtain a certificate of need. It also operates a system of competitive price bidding for franchises. Mali has placed some limits to opening a pharmacy at the capital, Bamako, with the (unsuccessful) intention of moving new entrants to rural areas. India has passed more liberal legislation between 2000 and 2004. Mexico and the Philippines have been very successful in promoting the entry of pharmacy chains only for dispensing generics brands, boosting availability, and affordability of medicines throughout their territories.

Price Regulation

Pricing regulation in the pharmacy industry takes the form of mandated dispensing fees, maximum margins (percentage over the final price), or markups (percentage on the manufacturer's or wholesale price). Margins or markups can be fixed or regressive with respect to prices, and there might be rules mandating that no discounts and promotions are offered to

the public, or that such discounts should be subject to a maximum.

Entry restrictions in Europe are typically coupled with price or retail margin regulations. Seventeen out of 25 EU Member States (all but Romania and Bulgaria, for which the authors do not have information) set the pharmacy markups by regulation. Discounts are not allowed. The other eight set maximum markups or fees for services allowing competition in discounts and promotions.

Discounts to final consumers are allowed only in Cyprus, the Czech Republic, Estonia, Lithuania, the Netherlands, Poland, Portugal, and Slovakia. Denmark, France, Germany, Ireland, and Spain allow for limited discounts in medicines not listed for mandatory health care systems reimbursement. The Netherlands, the UK, and Spain mandate some clawbacks to the national health systems that get back from pharmacies part of the discounts obtained from wholesalers and manufacturers.

In general in the EU, pharmacists are paid a fixed but regressive margin with respect to the price of each medicine sold. In Ireland, the Netherlands, and Slovakia pharmacists are paid a fixed fee for service, and in the case of the UK their purchase costs are reimbursed and there is a separate fixed fee for service.

In the US and Canada, pricing is free but it is agreed with Health Maintenance Organizations and the federal programs (Medicare and Medicaid) for reimbursed medicines.

Australia, New Zealand, and Syria have also regressive margin with respect to final prices to consumers. By contrast, in most low- and middle-income countries, pricing is regulated as a fixed margin over prices to consumers.

Benefits and Risks from Regulating Pharmacies

Having reviewed entry and price regulation in the world-wide industry, now it is time to turn to some economic analysis of the rationale for such regulations and review the practical experience with such regulations.

To start with it is required to assess how quality deterioration can result both in the presence and absence of information asymmetries. When quality deterioration is present, all four regulations outlined before (professional licensing, ownership restrictions, prescribing, and dispensing separation and management/supervision/assistance mandates) may help the industry to reduce it at a reasonable cost. Then the pros and cons of entry and pricing regulations are reviewed.

Regulating the Quality of Service

Quality deterioration

Professional services, in general, involve the application of professional human capital in order to judge individual cases. As a result of this peculiarity of the professional trades, the quality of the service provided is difficult to assess objectively. In the case of the pharmacy business, in general, a medical doctor makes the judgment as to which drug treatment is appropriate for each patient. Among the pharmacists' duties, the most important is to fill the prescription correctly, to advise the patient how best to comply with the treatment, and to prevent undesired drug interactions. In the case of OTC drugs, the pharmacist also assumes the duty of advising the patient regarding her decision on which drug is better for her specific minor ailment.

In contract theory terms, the patient is the principal and pharmacist is her agent. However, the agent has an information advantage. The pharmacist decides whether to invest effort in providing a high-quality service, or shirk and provide a low-quality service. The patient knows, but only imperfectly, about the quality of the service received after the purchase. This is a simple and typical hidden information situation usually termed a 'screening problem.' It can lead to quality deterioration phenomenon through adverse selection of service providers in the market. Patients would like to screen the high and low quality of service pharmacies out of the pool of available pharmacies. The separating equilibrium with full information is the one that allows the patients to pay for the high- or low-quality service, depending on their preferences and willingness to pay. For example, patients receiving new treatments may prefer to pay for a high-quality service, whereas patients with chronic diseases receiving repeat prescriptions may prefer to pay less for a low-quality service.

When patients cannot distinguish the pharmacy type due to the lack of information regarding the provider, and only uniform pricing is available as arbitrage is almost costless, the well-known problem of adverse selection is encountered. All pharmacies would end up by serving only at the low price and low-quality level.

Professional licensing may be used so as to regulate minimum quality standards: it screens the more able providers and deters shirking. Almost any licensing requirement imposes a fixed cost of entry that may drive out the least able providers: the entry cost is simply not affordable for the potential entrants with lower abilities.

Quality deterioration may also occur when there are no information asymmetries, when providers serve the marginal consumer having the lowest willingness to pay for quality. Minimum quality standards can help to avoid such outcomes. Such regulations are supposed to drive quality up, lifting it closer to the willingness to pay by the average consumer. In doing so, these regulations can raise welfare. Professional licensing and rules restraining management, supervision, and advising only to professional pharmacists may help out to monitor minimum standard quality regulations.

Licensing, externalities, and public goods

Professionals in general, and pharmacists in particular, provide services not only to their consumers but sometimes also to the public at large in the form of externalities or by providing public goods. When pharmacists dispense vaccines, it is not only their customers that gain some surplus but also the public at large thanks to an externality in the form of a reduced probability of other people contracting the disease in question. Likewise, when pharmacists dispense narcotics or antibiotics only with a proper prescription, it is not only that their patients benefit but also the public in general. Avoiding drug dependence or antibiotic resistance is something that improves other people's health.

The regulation of professional behavior is usually justified by rationales such as these. Codes of conduct for filling

prescriptions, checking for drug interactions, serving chronic patients, and so on, contribute to the benefit of both patients and the public at large. Pharmacists also produce positive externalities in their role of gatekeepers.

Pharmacists are paid for these services. Some argue that, while performing these duties, the pharmacist should keep the associated rents. The threat of being expelled from the licensed profession would then also entail loss of rents that further restrains them from underperformance. Entry restraints and professional body oversight is thus viewed as a mechanism for encouraging compliance with their professional obligations.

Alternatively, the pharmacist might be paid for preventive services.

Regulating Entry through A Needs Test and Pricing

Theory is ambiguous as to whether there is scope for welfare enhancing entry regulations in markets of differentiated products in the retail pharmacy industry. The literature has identified instances in which restrictions in the number of suppliers or price controls, or a combination of both, may be welfare increasing.

In each locality, entry brings the benefits of greater price competition and better local availability (access). Consumers gain from both, receiving cheaper medication and having outlets closer to where they are located. Each new entrant steals business from the pharmacies located in their catchment area but at the same time brings benefits to consumers. In industries in which there are no fixed costs of entry (an up-front cost not related to the volume of business), free entry is welfare enhancing.

Free entry may, however, be excessive. Entry is excessive whenever differentiation by location is low (consumers are not willing to pay for having more pharmacies at different locations in any catchment area). In this case, entrants add less to the consumer surplus when they enter the market than the amount they reduce the profits to incumbents by stealing their business from them. So, if the fixed costs of servicing patients are large, there is scope for a welfare enhancing regulation that restricts entry. At the same time, competition in each catchment area will take the form of an oligopoly game, so there might be also scope for pricing regulation. The equilibrium pricing game drives pricing in oligopoly well above marginal or average costs.

The more general models of pricing in oligopoly games with product differentiation show that the Nash equilibrium implies prices that differ from marginal costs. Additionally, when pharmacies have dominant positions in their catchment areas, there is also room for policy to avoid what it is known as double marginalization, when the upstream market of manufacturers or wholesalers is not competitive enough. Double marginalization appears when prices turn out to be excessive not only because manufacturers or wholesalers have upstream market power to overcharge in equilibrium but also retailers can overcharge as they can profit from their dominant position downstream. Limiting this double marginalization by allowing manufacturers or the government to set the final price can be welfare enhancing.

However, private interests might lobby for entry restrictions and pricing regulations to ensure that pharmacists obtain excess profits or pure regulatory rents. The European Commission has initiated infringement proceedings against countries that operate over tight entry and ownership regulations on the grounds that restricting freedom of establishment is neither an adequate nor a proportional public interest policy. On the contrary, the EC argues that it is a way of guaranteeing rents for incumbent pharmacies. The European Commission has initiated infringement proceedings against Austria, Bulgaria, France, Germany, Italy, Portugal, and Spain, though the results of infringement proceeding have been modest.

Attempts to reform entry and pricing regulations are problematic as the incumbents occupying the upper tail of the distribution and who gain most from the restrictions will invest heavily in lobbying to avoid policy reforms.

Concluding Remarks

Countries should choose the combination of policy options according to their efficiency and feasibility in any constrained environment.

Getting prices right (i.e., close to average costs, whether by competition, regulation, or contractual arrangement) is key to approaching an optimum. It makes entry regulation unnecessary. However, it should be borne in mind that any good price regulation or contractual arrangement has to get the number of pharmacies right when setting the price.

Adjusting the price to the costs in different localities is also important and difficult but it is essential to make sure that pharmacies are available and open to the public throughout the territory.

When countries do not have the institutions, the human capital or the technology to get the pricing right or to get the right number of pharmacies, they have to evaluate whether capture drives regulation toward undesirable outcomes and whether free pricing and entry is an attainable and reasonable second best.

Acknowledgment

The authors acknowledge the unconditional research grants from the Spanish Ministry of Science and innovation (ECO2009-06946), the Catalan Government (SGR2009-1066), and RecerCaixa (unconditional research grant from Fundació La Caixa, the philanthropic arm of the savings bank, La Caixa).

See also: Health Services in Low- and Middle-Income Countries: Financing, Payment, and Provision. Infectious Disease Externalities. Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity. Markets in Health Care. Markets with Physician Dispensing. Occupational Licensing in Health Care. Pharmaceutical Marketing and Promotion. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Primary Care, Gatekeeping, and Incentives

Further Reading

Abood, R. (2007). *Pharmacy practice and the law* (5th ed). Sudbury, MA: Jones and Bartlett Publishers.

Alderighi, M. and Piga, C. A. (2012). Selection, heterogeneity and entry in professional markets. Available at SSRN: http://ssrn.com/abstract=2194391 or http://dx.doi.org/10.2139/ssrn.2194391

Anderson, S. (2002). The state of the world's pharmacy: A portrait of the pharmacy profession. *Journal of Interprofessional Care* **16**(4), 391–404.

Arruñada, B. (2006). Managing competition in professional services and the burden of inertia. In Claus-Dieter, E. and Isabela, A. (eds.) *European competition law annual 2004: The relationship between competition law and the (liberal) professions* 2006, pp. 51–71. Oxford and Portland Oregon: Hart Publishing.

Borrell, J. R. and Fernández-Villadangos, L. (2009). Assessing excess profits from different entry regulations: The case of pharmacies in Spain, Xarxa de Recerca en Economia Aplicada (XREAP) Working Papers, 2009–3. Barcelona.

Borrell, J. R. and Fernández-Villadangos, L. (2010). Clustering or scattering: The underlying reason for regulating distance among retail outlets, Xarxa de Recerca en Economia Aplicada (XREAP) Working Papers, 2010–12. Barcelona.

European Commission (2008). Pharmaceutical sector inquiry. *Preliminary Report, DG Competition Staff Working Paper*, Brussels.

Mankiw, N. G. and Whinston, M. D. (1986). Free entry and social inefficiency. *RAND Journal of Economics* **17**, 48–58.

ÖBIG (2006). Surveying, assessing and analyzing the pharmaceutical sector in the 25 EU Member States. *Report Commissioned by the DG Competition – European Commission, Office for Official Publications of the European Communities*, Brussels.

Roberts, M. J. and Reich, M. R. (2011). *Pharmaceutical reform: A guide to improving performance and equity*. Washington, DC: The World Bank.

Schaumans, C. and Verboven, F. (2008). Entry and regulation. Evidence from health care professions. *RAND Journal of Economics* **22**, 490–504.

Waterson, M. (1993). Retail pharmacy in Melbourne: Actual and optimal densities. *Journal of Industrial Economics* **41**, 403–419.

WHO/HAI (2011). Competition policy. In: Review Series on Pharmaceutical Pricing Policies and Interventions, WHO/HAI Project on Medicine Prices and Availability, Working Paper 4. Geneva: World Health Organization and Health Action International.

Physician Labor Supply

H Fang, University of Colorado Denver, Denver, CO, USA
JA Rizzo, Stony Brook University, Stony Brook, NY, USA

Introduction

Labor supply has been a well-studied topic in the labor economics literature (Killingsworth, 1983; Pencavel, 1986; Killingsworth and Heckman, 1986; Blundell and MaCurdy, 1999). As the key clinical decision-makers, physicians provide an essential input into the production of health care for their patients. Understanding the determinants of physician labor supply has important implications for the production and cost of care and for health care access. Hence, physician labor supply is a topic of considerable interest in health economics as well.

This article examines the factors affecting physician labor supply. Labor supply measures are considered broadly, and may include annual hours worked, numbers of surgeries performed, office visits, and so on. (Strictly speaking, physician labor supply curves must include earnings as a determinant of labor supply. Many studies, however, focus on other aspects affecting labor supply, such as the competitive conditions in the market. These studies typically do not include earnings as a separate factor affecting labor supply. A likely reason is that both measures of competition and earnings would have to be treated as endogenous, prohibitively challenging data estimation issues. However, these studies are included because from a policy perspective, it is important to understand a variety of factors that may affect the physician's decision to offer services.) However, the related topic of aggregate forecasts of physician manpower is not addressed. The authors begin with a general framework discussing the competing goals of the physician in choosing labor supply and then review studies that have examined a number of key issues affecting the labor supply decision.

This article is organized into seven parts. The Section Introduction discusses the conceptual issues in the physician labor supply decision. The Section Conceptual Issues considers studies of the relationship between physician earnings and labor supply. The effects of competition and physician fee schedules on labor supply are described in the Section Earnings and Labor Supply. This section also reviews labor supply responses as possible evidence of demand inducement. The Section Competition, Fee Schedules, and Labor Supply examines the roles of the so-called 'target income hypothesis,' and more recently reference incomes as they pertain to the labor supply decision. Physician labor supply under managed care is discussed in the Section Income Targets, Reference Incomes, and Labor Supply and the effect of malpractice liability on labor supply is examined in the Section Managed Care and Labor Supply. The Section Malpractice Liability and Labor Supply concludes this article.

Conceptual Issues

The physician is assumed to maximize his utility with respect to income and leisure, both of which are 'goods'. Income includes labor income and nonlabor income. In altruistic variants of this basic framework, patients' health enters as an argument into the utility function as well. (In altruistic models, the physician plays the role of agent for his patients, with patients' health entering as an argument into the physician's utility function. As the physician has competing objectives including the desire to obtain more income and leisure, he may be an imperfect agent.) The physician chooses labor supply so as to maximize this utility function. A complex constellation of factors, including earnings potential, altruistic goals, competitive conditions, incentives provided by insurers, the regulatory environment, and the threat of malpractice litigation may affect this decision.

A long and controversial debate has centered on the notion that income targets affect the provision of physician services. The so-called target income hypothesis asserts that physicians set income targets and, when their actual incomes fall below these targets, will increase the volume of their services to offset, in whole or in part, their perceived income shortfall. In critiquing this hypothesis, McGuire and Pauly (1991) have argued that there is no conceptual basis as to how income targets are set. Moreover, they present a theoretical model which demonstrates that increasing volume of services in response to fee restrictions need not depend on income targets, but may occur in the context of a standard model of profit maximization, provided that income effects from fee reductions are sufficiently strong.

Although the standard model underlying the physician's labor supply decision is neoclassical, more recent treatments that have considered the role of target or reference incomes on labor supply have adapted models from prospect theory (Kahneman and Tversky, 1979). Borrowing from prospect theory, these models incorporate the notion that physicians set reference incomes and compare these benchmarks to their actual earnings in deciding on whether to adjust their labor supply (Rizzo and Zeckhauser, 2003, 2007). These models also incorporate the notion of loss aversion (Tversky and Kahneman, 1991; Goette *et al.*, 2004), a phenomenon in which individuals strongly prefer avoiding losses to equivalent-sized gains. Loss version posits a kink in the physician's utility curve for different marginal utilities of income below and above this reference income. When the physician's actual income is less than the reference income, the marginal utility of income is very steep in that range. The relevance of reference points and loss aversion has been well-established in the experimental psychology literature (Rabin, 1998; Heath *et al.*, 1999; Schmidt and Traub, 2002; Fellner and Maciejovsky, 2007; Rizzo and Zeckhauser, 2003, 2007). Their inclusion in models of physician labor supply provides an empirically validated framework for understanding the relevance of reference of target or reference incomes for physician labor supply.

 doi:10.1016/B978-0-12-375678-7.01113-5

Earnings and Labor Supply

Perhaps the most salient factor affecting the physician's decision to supply labor is the economic return that labor will generate. Hence, a number of studies have examined this relationship empirically. Changes in physician earnings may exert both income and substitution effects on the labor supply decision. If the substitution effect dominates, an increase in earnings will lead to an increase in labor supply. At sufficiently high earnings levels, however, the income effect may dominate, in which case one would observe a backward-bending labor supply. Early studies examined the effects of nonlabor income on physician labor supply by focusing on exogenous variations in nonpractice income, typically finding insignificant nonlabor income effects (Sloan, 1974; Hurdle and Pope, 1989).

In terms of labor income, however, most physicians are not paid by wages or salary; hence, their hourly earnings are endogenous. Sloan (1975) examined the relationship between hourly earnings and hours worked employing a two-stage estimation treating earnings as endogenous. The effects of earnings on labor supply varied somewhat according to specification, but were generally modest and in some cases statistically insignificant. Using a nationally representative database of young physicians under the age of 40 years, Rizzo and Blumenthal (1994) estimated a model of physician labor supply, treating physician earnings as endogenous and employing two-stage least-squares estimation. They obtain separate estimates of the income and substitution effects of a change in hourly physician earnings for male physicians, finding a significantly negative income effect with an elasticity of -0.26, and a significantly positive substitution effect elasticity of 0.49. The total effect of wage increase is to raise labor supply, with an elasticity in the range of 0.2–0.3. Small sample size precludes obtaining separate estimates for female physicians. A more recent study from the UK also reports a modest positive association between earnings and physician labor supply, with elasticities in the range of 0.09–0.12 (Ikenwilo and Scott, 2007).

Brown (1994) contrasts the use of aggregate versus physician-level data in estimating the earnings/labor supply relationship. Using aggregated data he finds no effect (Brown and Lapan, 1972; Brown *et al.*, 1974), but a negative relationship using physician-level data, with an elasticity of -0.2. He argues that physician-level data are preferable for estimating labor supply. Bradford and Martin (1995) also find evidence of a backward-bending labor supply curve. Thornton and Eakin (1997) estimate a model of a utility-maximizing solo practitioner. They find that an increase in nonlabor income will lead solo practitioners to allocate fewer hours to medical practice activities. In addition, they also report both income effect and substitution effects of labor income changes. Consistent with earlier research, they find that the net effect of physician service fee reductions leads physicians to reduce their labor supply, so the substitution effect dominates the income effect.

Competition, Fee Schedules, and Labor Supply

Studies relating competition and fee schedules to physician labor supply have typically sought to provide evidence of demand-inducing behavior by physicians. If physicians increase their supply of services in response to greater competitive pressures or reductions in fees, then it is taken as indirect evidence of demand-inducing behavior. In fact, even if such relationships exist, it is unclear whether these behaviors reflect demand inducement. One is on firmer ground simply interpreting them as labor supply responses to changing financial incentives.

Competition and Labor Supply

Studies of the effect of competition on labor supply have employed metrics for competition such as per capita physicians in a market area, relating this measure to various types of physician services. An early study by Fuchs (1978) examines the supply of surgeons on the number of surgeries performed, treating physician supply as endogenous, with variables measuring the appeal of a market (e.g., urban, hotel receipts) serving as instruments. The results suggest that increased competitive pressure as measured by physician supply leads to an increase in surgeries, with an elasticity of 0.3. Subsequent studies employing a similar strategy but with more complete control variables produce similar results (Rossiter and Wilensky, 1983, 1984; Cromwell and Mitchell, 1986; Birch, 1988; Grytten *et al.*, 1990; Scott and Shiell, 1997; Baltagi *et al.*, 2005).

A limitation with these studies is the choice of instruments. Employing an instrumental variable approach typical in these studies, Dranove and Wehner (1994) produce the seemingly bizarre result that greater competition among obstetrician/gynecologists (OBGYNs) increases childbirths. As it is highly unlikely that physicians induce demand for children, this result is taken as evidence that the instrumental variable approach employed in these studies is suspected. In fact, the likely explanation for their result reflects border crossing, for example, that pregnant women travel to locations where there are ample supplies of OBGYNs. Dranove and Wehner (1994) calls into question any causal interpretation between competition and labor supply or demand inducement using the instrumental variable strategy described earlier.

Gruber and Owings (1996) employ an alternative approach to address the endogeneity problem. They study the relationship between physician financial incentives and cesarean-section delivery. Between state and intertemporal variations in fertility rates are used as exogenous measures of competitive pressures facing OBGYNs. They hypothesize that, in response to declining fertility in the US, OBGYNs will substitute vaginal childbirths for the more lucrative and physician-intensive cesarean deliveries. Using nationally representative data from the period of 1970–82, they find that a 10% drop in fertility leads to a 0.6% increase in cesarean sections.

Fang and Rizzo (2009) argue that the relationship between competition and physician labor supply depends on the nature of third-party reimbursement. Using data from the Community Tracking Study Physician Survey 2000–01, they find that physician volume increases with more competition under fee-for-service reimbursement, but decreases with greater competition under managed care.

Fee Schedules and Labor Supply

A number of studies have examined physician labor supply responses to fee restrictions (Lee and Hadley, 1981; Rice, 1983; Mitchell *et al.*, 1989; Hurley *et al.*, 1990; Hurley and Labelle, 1995; Escarce, 1993, Rochaix, 1993; Nguyen, 1996; Nguyen and Derrick, 1997; Yip, 1998; Tai-Seale *et al.*, 1998; Gruber *et al.*, 1999; Mitchell *et al.*, 2000; Kantarevic *et al.*, 2008). In contrast to the competition studies, fee reductions may be considered exogenous to the individual physician. A number of these studies have found evidence of a volume offset effect; that is, physicians respond to real declines in their fees by increasing the volume of their services. However, the volume increase is not sufficient to fully recoup the income losses from the fee cuts.

Thus, Nguyen (1996) studies physician volume responses to reductions in the Medicare fee schedule reduction. The results indicate that physicians will increase their service volume by 3.7% in response to a 10% reduction in the Medicare fee schedule. Nguyen and Derrick (1997) examine the impact of Medicare fee cuts for certain 'overpriced procedures,' finding that for physicians who experience the largest fee reductions, a 10% decline in price lead to a 4% increase in volume. In a comprehensive assessment of volume offset effects for multiple specialties and payers, Tai-Seale *et al.* (1998) find that the point estimates for volume responses to fee restrictions varied across specialties, but these responses were statistically insignificant in most cases. Yip (1998) studies the effects of reductions in fee schedules on the volume of coronary artery bypass grafting (CABG) procedure using a longitudinal panel of physicians in New York and Washington States. For physicians in both Medicare and private markets, she finds that physicians whose incomes are cut by reduced fee schedules exhibit a large volume response, recouping 70% of their lost income. Gruber *et al.* (1999) investigate the effect of Medicaid fee differentials on cesarean delivery. They find that cesarean delivery for Medicaid patients is significantly less likely than for privately insured patients, reflecting that the fee differentials between cesarean and vaginal deliveries are smaller under the Medicaid program than the private insurance. Mitchell *et al.* (2000) analyze physician labor supply responses to Medicare fee reductions during the period of 1991–94 using pooled cross-sectional time series data. They focus on ophthalmologists and orthopedic surgeons performing cataract extractions and major joint repair/replacement procedures, respectively. In contrast to most previous research, they find that fee reductions are associated with fewer surgeries.

Income Targets, Reference Incomes, and Labor Supply

The target income hypothesis asserts that physicians set income targets and will attempt to reach or get closer to this target by increasing their services in response to increased competitive pressures or cutbacks in their fees. Most studies of this issue have been indirect, relating measures of competition and fee cuts to the volume of physician services. Such studies have already been addressed in Section Earnings and Labor Supply.

Few studies have related direct measures of income targets to physician labor supply. Rizzo and Zeckhauser (2003) use a unique panel data of physicians under the age of 40 years that includes a physician-specific measure of target or reference income (in particular, physicians were asked: "Considering your career stage, what do you consider to be an adequate income after expenses but before taxes from your professional activities?" The response to this question was taken as the physician's reference income (Rizzo and Zeckhauser, 2003, 2007)). They find that incomes increase substantially in response to higher reference incomes for physicians who are below their reference incomes, but not for those who are at or above. However, they also note that physicians appear to raise their incomes, not by increasing labor supply as measured by hours worked, but by performing activities that generate a higher hourly return (e.g., performing more lucrative services). A subsequent study also finds no evidence that hours worked respond to reference incomes (Rizzo and Zeckhauser, 2007).

Managed Care and Labor Supply

Managed care has grown rapidly in the US since 1980 and is the dominant form of health insurance (Robinson, 1999). Intended to control the rapid growth in health care costs in the US, managed care may exert strong effects on physician practice patterns, including labor supply, because the physician's decision-making process is likely to respond to the financial incentives and restrictions created by managed care.

Hirth and Chernew (1999) discuss two fundamentally different labor market regimes for physicians: fee-for-services and managed care, noting that physicians practicing in managed care environments are less likely to enhance their income by providing more services compared with those mainly reimbursed on the basis of fee-for-service. Libby and Thurston (2001) examine the impact of managed care contracting on physician labor supply by extending the standard labor supply model to incorporate managed care incentives. They find that managed care contracting generally reduces the number of hours that physicians practice, but the net effects become small and insignificant after accounting for the endogeneity of physician managed care contracting behavior.

Malpractice Liability and Labor Supply

Physician labor supply may also be affected by other factors such as the threat of malpractice liability. Thornton (1997) shows a significant income effect of a change in malpractice premiums on physician labor supply. In particular, higher malpractice premiums lead primary care physicians to increase their practice hours, possibly to recoup some of the income losses associated with these premiums. An alternative interpretation is that physicians regard malpractice premiums as a tort signal and attempt to work more hours to reduce the possibility of malpractice liability. Either effect leads physicians to increase labor supply. Thornton (1999) compares the magnitudes of the income effect and tort signal effect in

response to malpractice premiums. He finds that the income effect dominates, with the tort effect being much smaller.

In contrast, Helland and Showalter (2009) analyze the effect of state-level malpractice reforms during the period of 1983–88 on physician behavior. They find that an increase of 1% in the probability of incurring a malpractice suit will reduce the weekly hours worked by 0.29%. The magnitude of this elasticity increases to -1.224 for physicians aged 55 years or older. Kessler and McClellan (1996) analyze malpractice reforms designed to reduce the threat if physician liability to examine whether physicians practice defensive medicine; for example, increasing the provision of services to ward off the threat of malpractice suits. They find that malpractice reforms are associated with a significant reduction (5–9%) in physician expenditures for Medicare patients after the malpractice reforms without substantial changes in mortality or medical complications.

Conclusion

As key players in medical decision-making, understanding factors affecting the physician's decision to supply care will remain an important topic in health economics research. Much effort has been devoted to this issue already and with fruitful results. Most studies suggest that greater earnings potential and fee restrictions both lead to increased labor supply, though the response is typically fairly small and inelastic. Greater competition also appears to increase labor supply, although econometric challenges associated with much of this literature suggest that these results should be viewed with more caution. Perhaps not surprisingly, managed care has had a restrictive role on physician labor supply. Less certain are the effects of target or reference incomes and the role of medical malpractice liability.

Although the notion that reference points are used in decision-making has considerable empirical support (Rizzo and Blumenthal, 1996; Rabin, 1998; Heath *et al.*, 1999; Schmidt and Traub, 2002; Fellner and Maciejovsky, 2007; Rizzo and Zeckhauser, 2003, 2007), McGuire and Pauly (1991) rightly assert that there remains no theory as to how and why physician income targets are set and why they should affect the physician labor supply decision. Understanding these issues is an important direction for further research.

Increasingly, patients are playing a more proactive role in the care they receive from physicians (Fang *et al.*, 2008). This consumerist orientation may have implications for physician labor supply as well. Fang and Rizzo (2009) introduce the notion of 'physician demand enablement' as physician labor supply responses to patient-initiated requests for services. The effects of 'consumerist' patients on physician's willingness to supply care also warrant further study.

See also: Income Gap across Physician Specialties in the USA. Medical Malpractice, Defensive Medicine, and Physician Supply. Organizational Economics and Physician Practices. Physician Management of Demand at the Point of Care. Physician-Induced Demand. Specialists

References

Baltagi, B. H., Bratberg, E. and Holmas, T. H. (2005). A panel data study of physicians' labor supply: The case of Norway. *Health Economics* **14**(10), 1035–1045.

Birch, S. (1988). The identification of supplier-inducement in a fixed price system of health care provision: The case of dentistry in the United Kingdom. *Journal of Health Economics* **7**(2), 129–150.

Blundell, R. and MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In Ashenfelter, O. C. and Card, D. (eds.) *Handbook of labor economics*, vol. 3, pp. 1559–1695. North-Holland: Elsevier.

Bradford, D. and Martin, R. E. (1995). Supplier-induced demand and quality competition: An empirical investigation. *Eastern Economic Journal* **21**(4), 491–503.

Brown, D. M. (1994). The rising price of physicians' services: A correction and extension on supply. *Review of Economics and Statistics* **76**(2), 389–393.

Brown, D. M., Feldstein, M. and Lapan, H. E. (1974). The rising price of physicians' services: A clarification. *Review of Economics and Statistics* **56**(3), 396–398.

Brown, D. M. and Lapan, H. E. (1972). The rising price of physicians' services: A comment. *Review of Economics and Statistics* **54**(1), 101–105.

Cromwell, J. and Mitchell, J. B. (1986). Physician-induced demand for surgery. *Journal of Health Economics* **5**(4), 293–313.

Dranove, D. and Wehner, P. (1994). Physician-induced demand for childbirths. *Journal of Health Economics* **13**(1), 61–73.

Escarce, J. J. (1993). Effects of lower surgical fees on the use of physician services under Medicare. *Journal of American Medical Association* **269**(19), 2513–2518.

Fang, H., Miller, N. H., Rizzo, J. A., Zeckhauser, R. J. (2008). Demanding customers: Consumerist patients and quality of care. *NBER Working Paper W14350.* Cambridge, MA: National Bureau of Economic Research, Inc.

Fang, H. and Rizzo, J. A. (2009). Competition and physician-enabled demand: The role of managed care. *Journal of Economic Behavior and Organization* **72**(1), 463–474.

Fellner, G. and Maciejovsky, B. (2007). Risk attitude and market behavior: Evidence from experimental asset markets. *Journal of Economic Psychology* **28**(3), 338–350.

Fuchs, V. R. (1978). The supply of surgeons and the demand for operations. *Journal of Human Resources* **13**(supplement), 35–56.

Goette, L., Huffman, D. and Fehr, E. (2004). Loss aversion and labor supply. *Journal of the European Economic Association* **2**(2–3), 216–228.

Gruber, J., Kim, J. and Mayzlin, D. (1999). Physician fees and procedure intensity: The case of cesarean delivery. *Journal of Health Economics* **18**(4), 473–490.

Gruber, J. and Owings, M. (1996). Physician financial incentives and cesarean section delivery. *Rand Journal of Economics* **27**(1), 99–123.

Grytten, J., Holst, D. and Laake, P. (1990). Supplier inducement: Its effect on dental services in Norway. *Journal of Health Economics* **9**(4), 483–491.

Heath, C., Larrick, R. P. and Wu, G. (1999). Goals as reference points. *Cognitive Psychology* **38**(1), 79–109.

Helland, E. and Showalter, M. H. (2009). The impact of liability on the physician labor market. *Journal of Law and Economics* **52**(4), 635–663.

Hirth, R. A. and Chernew, M. E. (1999). The physician labor market in a managed care-dominated environment. *Economic inquiry* **37**(2), 282–294.

Hurdle, S. and Pope, G. C. (1989). Physician productivity: Trends and determinants. *Inquiry* **26**(1), 100–115.

Hurley, J. and Labelle, R. (1995). Relative fees and the utilization of physicians' services in Canada. *Health Economics* **4**(6), 419–438.

Hurley, J., Labelle, R. and Rice, T. (1990). The relationship between physician fees and the utilization of medical services, in Ontario. In Scheffler, R. and Rossiter, L. (eds.) *Advanced in health economics and health services research*, pp. 49–78. Greenwich, CT: JAI Press.

Ikenwilo, D. and Scott, A. (2007). The effects of pay and job satisfaction on the labour supply of hospital consultants. *Health Economics* **16**(12), 1303–1318.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2), 263–291.

Kantarevic, J., Kralj, B. and Weinkauf, D. (2008). Income effects and physician labour supply: Evidence from the threshold system in Ontario. *Canadian Journal of Economics* **41**(4), 1262–1284.

Kessler, D. and McClellan, M. (1996). Do doctors practice defensive medicine? *Quarterly Journal of Economics* **111**(2), 353–390.

Killingsworth, M. R. and Heckman, J. J. (1986). Female labor supply: A survey. In Ashenfelter, O. C. and Layard, R. (eds.) *Handbook of labor economics*, vol. 1, pp. 103–204. North-Holland: Elsevier.

Killingsworth, M. R. (1983). *Labor supply.* Cambridge: Cambridge University Press.
Lee, R. H. and Hadley, J. (1981). Physicians' fees and public medical care programs. *Health Service Research* **16**(2), 185–203.
Libby, A. M. and Thurston, N. K. (2001). Effects of managed care contracting on physician labor supply. *International Journal of Health Care Finance and Economics* **1**(2), 139–157.
McGuire, T. G. and Pauly, M. V. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics* **10**(4), 385–410.
Mitchell, J. B., Wedig, G. and Cromwell, J. (1989). The Medicare physician fee freeze: What really happened? *Health Affairs* **8**(1), 21–33.
Mitchell, J. M., Hadley, J. and Gaskin, D. J. (2000). Physicians' responses to Medicare fee schedule reductions. *Medical Care* **38**(10), 1029–1039.
Nguyen, N. X. (1996). Physician volume response to price controls. *Health Policy* **35**(2), 189–204.
Nguyen, N. X. and Derrick, F. W. (1997). Physician behavioral response to a Medicare price reduction. *Health Services Research* **32**(3), 283–298.
Pencavel, J. (1986). Labor supply of men: A survey. In Ashenfelter, O. C. and Layard, R. (eds.) *Handbook of labor economics*, vol. 1, pp. 3–102. North-Holland: Elsevier.
Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature* **36**(1), 11–46.
Rice, T. H. (1983). The impact of changing Medicare reimbursement rates on physician-induced demand. *Medical Care* **21**(8), 803–815.
Rizzo, J. A. and Blumenthal, D. (1994). Physician labor supply: Do income effects matter? *Journal of Health Economics* **13**(4), 433–453.
Rizzo, J. A. and Blumenthal, D. (1996). Is the target income hypothesis an economic heresy? *Medical Care Research and Review* **53**(3), 243–266.
Rizzo, J. A. and Zeckhauser, R. J. (2003). Reference incomes, loss aversion, and physician behavior. *Review of Economics and Statistics* **85**(4), 909–922.
Rizzo, J. A. and Zeckhauser, R. J. (2007). Pushing incomes to reference points: Why do male doctors earn more? *Journal of Economic Behavior and Organization* **63**(3), 514–536.
Robinson, J. C. (1999). The future of managed care organization. *Health Affairs* **18**(2), 7–24.
Rochaix, L. (1993). Financial incentives for physicians: The Quebec experience. *Health Economics* **2**(2), 163–176.
Rossiter, L. F. and Wilensky, G. R. (1983). A reexamination of the use of physician services: The role of physician-initiated demand. *Inquiry* **20**(2), 162–172.
Rossiter, L. F. and Wilensky, G. R. (1984). Identification of physician-induced demand. *Journal of Human Resources* **19**(2), 231–244.
Schmidt, U. and Traub, S. (2002). An experimental test of loss aversion. *Journal of Risk and Uncertainty* **25**(3), 233–249.
Scott, A. and Shiell, A. (1997). Analyzing the effect of competition on general practitioners' behaviour using a multilevel modeling framework. *Health economics* **6**(6), 577–588.
Sloan, F. A. (1974). A microanalysis of physicians' hours of work decisions. In Perlman, M. (ed.) *The economics of health and medical care*, pp. 302–325. New York, NY: John Wiley and Sons.
Sloan, F. A. (1975). Physician supply behavior in the short run. *Industrial and Labor Relations Reviews* **28**(4), 549–569.
Tai-Seale, M., Rice, T. H. and Stearns, S. C. (1998). Volume responses to Medicare payment reductions with multiple payers: A test of the McGuire–Pauly model. *Health Economics* **7**(3), 199–219.
Thornton, J. (1997). Are malpractice insurance premiums a tort signal that influence physician hours worked? *Economics Letters* **55**(3), 403–407.
Thornton, J. (1999). The impact of medical malpractice insurance cost on physician behaviour: The role of income and tort signal effects. *Applied Economics* **31**(7), 779–794.
Thornton, J. and Eakin, K. (1997). The utility-maximizing self-employed physician. *Journal of Human Resources* **32**(1), 98–128.
Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics* **106**(4), 1039–1061.
Yip, W. C. (1998). Physician response to Medicare fee reductions: Changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors. *Journal of Health Economics* **17**(6), 675–699.

Physician Management of Demand at the Point of Care

M Tai-Seale, Palo Alto Medical Foundation Research Institute, Palo Alto, CA, USA

Introduction

Many perspectives can be taken to look at physician practice behaviors. Other articles in this section of the Encyclopedia and in the literature (e.g., by McGuire, Chandra, Cutler, and Song) provide extensive information on multiple approaches to studying physician practice. In general, empirical studies of physician practice behaviors build on administrative data which contain the information necessary for billing but no details on what happened in the encounter. They do not contain detailed information on how patient demand was expressed and how physician responded while they were engaged in an office visit, i.e., at the point of care. Survey data have its own problems, and can be subject to recall, social desirability, or self-perception biases. Physician management of patient demand at this microlevel needs special data to shed light on the exchanges between patient and physician. Video or audio recordings of office visits are two such data sources. This article focuses on interactions between primary care physicians and patients at the point of care, where patient demand is managed through conversations. By analyzing recordings in a detailed way, this methodological approach enables the authors to closely observe what Kenneth Arrow notes as the activities of producing medical care that are unobservable through the lens of administrative data.

Video or audio recordings of the visits allow the authors to examine the length and content of visits. They provide a comprehensive representation of the patient–physician encounter, unlike chart review which can be influenced by physicians' charting patterns and their tendency to underreport delivery of some services or overreport other services. The following two excerpts from transcripts of two visits illustrate the nuanced information about the demand for bypass surgery (Example 1) and uncertainty about incidence of disease and efficacy of treatment, and the use of heuristics in decision making under uncertainty (Example 2).

Example 1

An elderly female patient seeing her primary care physician for a dry cough, before undergoing bypass surgery that was scheduled to take place in a few weeks.

Patient: (Clasping her hands) and really, why I agreed to the surgery was because I thought that they would be able to fix the damage the heart attack had done to my heart. And they said no.
Physician: No, you can't fix that. It helps to reestablish the circulation so that you don't have any further heart attacks. … so I will see you back here after your surgery.
Patient: Yes, if I don't cancel it again.

It is seen that the patient had expressed some reservations about the operation. The physician did not address her concerns or allowing her to change her mind about undergoing the surgery.

Example 2

An elderly female patient seeing her primary care physician, after having had 3 visits during which multiple new psychotropic medications were started and stopped in the last 3 months. She complained of unsteadiness and off balance, anxiety, and forgetfulness. The following was the exchange at 17 min into the visit.

Physician: The girls are going to set up a follow-up appointment in two weeks and we will see how we're doing. You're going to stop the Lorazepam, stop Lorazepam, take Vitamin E, water pill, …
Patient: (Raising her hand as though to signal she has something to say) Now, …
Physician: (Taking her hand, shaking it, and continuing to talk) … everything else stays the same, including the Wellbutrin and we're going to see you back in two weeks.
Patient: But now, you said on that Vitamin E, 1000 twice a day, 2000?
Physician: Yes, ma'am.
Patient: Ok.
Physician: That's what the study states. It's written down here. Ok?
Patient: Yeah, sure.
Physician: (Moving to help patient down from exam table and starts walking towards the door) There you go. We'll try a little 'addition by subtraction' and hope that by stopping the Lorazepam that will stop your coordination difficulties and maybe the Wellbutrin we can continue.

Stopping Lorazepam suddenly instead of tapering it off slowly could exacerbate the patient's anxiety. The exceedingly high daily dose of vitamin E is also not indicated. From the video data it is noticed that the physician was not willing to hear any more from the patient but wanted to bring her back for another visit 2 weeks later.

Questions That Have Been Informed by Microlevel Interaction Analysis

Direct Observation Analysis

A number of studies that have shed some light on the 'blackbox' of patient–physician exchanges at the point of care using microlevel interaction analysis of video and audio recordings of office visits have been undertaken. They encompass three general areas: (1) time allocation in primary care office visits, (2) time management practice in office visits that resemble the use of a behavior rule, and (3) management of diverse demand with heterogeneous level of professional and personal uncertainties.

Time allocation in primary care office visits

Time is a scarce resource in a physician's office practice. How physicians use clinic time has important implications on

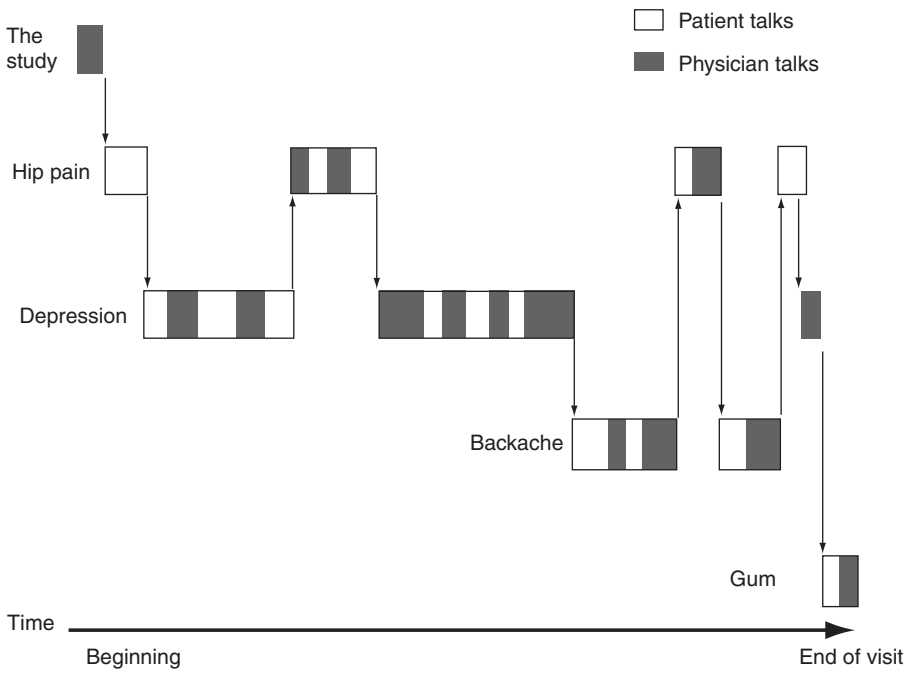


Figure 1 Flow of conversation during a visit.

quality of care, patient trust, and malpractice suits, and is one of the components of physician payments in the resource-based relative value scale. Primary care office visits are essentially communication events between patient (demand) and physician (supply) on which data and research methods from other scientific disciplines have made extensive efforts. Social psychologist Mishler views patient–physician conversations as complex, multidimensional, and multifunctional exchanges. Health services researchers recognize the unique and critical role of primary care physicians in providing patients with an 'advanced medical home' where complex comorbidities are diagnosed and treated. Despite previous research efforts on patient–physician interactions, however, the literature was silent on how physicians allocate time within a visit. The point of care exchanges that occur behind the closed door were considered hidden and noncontractible.

To examine how clinic time was actually spent during patients' visits to primary care physicians and to identify the factors that influence time allocations, a novel approach was developed to analyze video recordings of routine office visits in primary care practices. Currently, audio recordings of period health examinations (PHEs) are being analyzed in an integrated health delivery system which provides supplemental data on service utilization before and after the recorded visits. The findings have been rather thought provoking.

Specifically, not only the length of visits but also, more importantly, the content of visits in terms of units of clinical decision making referred to as 'topics,' operationalized as clinical issues raised by either participant was examined. An interaction was coded directly from an audio or video recording of the visit, along with transcripts of the interaction, based on topics sequentially introduced by patient or physician. After partitioning a visit into topics, the amount of time spent on each topic by patient and physician was further recorded. In the PHE study, the quality of communication on each topic was also measured. Figure 1 illustrates the flow of conversation in one visit, from topic to topic, over time. It is evident that the exchange took a rather free flow form, consistent with general conversation patterns in casual conversations, despite training in medical school and residencies on how to structure an office visit.

This approach of using microlevel data collected at the point of care allows the authors to examine how much time is dedicated to specific topics, the cognitive and emotional efforts invested in the exchanges across topics, and the factors that influenced how clinical time and efforts are allocated. It has been found that primary care office visits vary not only in length but also in the division of time among topics. Patients typically present multiple complaints during an office visit requiring physicians to divide time and resources during a visit to deal with competing demands. Very limited amount of time was dedicated to specific topics. In the video study, it was found that the median visit length was 15.7 min covering a median of six topics. Approximately 5 min were spent on the longest topic, whereas the remaining topics each received 1.1 min. Although time spent by patient and physician on a topic responded to many factors, length of the visit overall varied little even when contents of visits varied widely. Macrofactors associated with each site (e.g., academic medical center where physicians are paid by salary vs. physicians in fee-for-service solo practices or in a managed care group practice) had more influence on visit and topic length than the nature of the problem patients presented.

Time management practice resembling the use of a behavioral rule

The New York Times published an article about the above study entitled "The Ticking Clock in the Doctor's Office" along

with a cartoon which depicts a clock wedged between a patient and a physician. The ticking clock resembles an upward sloped shadow price of time that rises as time elapses once a visit starts. Although patient and physician both initiate discussions on topics during the visit, it is the physician who decides when the visit needs to end. The authors wanted to examine how physicians decide when to end a visit. They were most interested in the order of topics in terms of 'seriousness,' or of the physician's assessment of the benefits of spending time on the topic.

One hypothesis was that potential topics are ranked by importance and the most important topics are covered first. The efficient allocation of the physician's time can be described by a threshold value or shadow price, call it λ, such that any topic with value greater than λ is dealt with by the physician, and any topic with value less than λ is not. The value of λ is set so as to just use all of the time the physician has available. Another interpretation is that the physician has another activity with a constant value of λ. This other activity might, for example, be 'administrative work,' that can be done during the day or handled at the end of the day.

An alternative hypothesis was that physicians have a 'target' amount of time to spend with each patient, like a tennis coach who gives equal amount of time to each of his students. One way to model a 'target' is to regard the shadow price of time to be zero up to the target and infinite after the target. Under the alternative hypothesis, physician admits a new topic if and only if the value of the new topic is greater than or equal to λ given how much time has already elapsed in the visit. The target can be set by a norm that is dictated by protocols at the practice regarding the number of patients a physician needs to see each day to reach productivity goals. Under the influence of such productivity goals, the decision to end the visit is determined by a behavioral rule rather than maximization of the expected net benefit of their time with patients.

Empirically, the probability of a topic being the last topic of a visit was modeled. The key right-hand-side variables are four binary indicators of time elapsed when a topic was introduced: within 5 min of the beginning of the visit, between 5–10 min, between 10–15 min, and after 15 min. Multiple measures of the 'seriousness' of a topic were incorporated. The empirical findings support the alternative hypothesis: The likelihood of a topic being the last increased successively and significantly with each increment in the block of time for topic introduction. The results are robust to various specifications.

This example resonates with time accounting and targeting heuristics observed in behavioral economics. New York City cab drivers have been documented to quitting earlier on high-wage days and driving longer on low-wage days. Physicians are found to work under a similar behavioral rule, spending time patiently until the target is reached, then quickly closing the visit. The second example presented at the beginning of the article was from a visit in which the physician was by turns patient and inquisitive (at 5 min into the visit) and brisk and dismissive (at 17 min). He transformed the patient's raised hand for inquiry into a good-bye handshake, and escorted her from the room while telling her that he would bring her back for another visit in 2 weeks. Another approach used to close a visit was referring patients to social workers who would spend more time listening and figuring out how social services might offer support. Giving prescriptions for medications (sometimes even unindicated medications) is yet another approach to close a visit. Although fee-for-service environment clearly rewards such approach to increase demand, they are not necessarily what a perfect agent would have done. It is the opposite of what one would do to maximize the effectiveness of each visit which entails covering multiple issues at one sitting. Managing demand by limiting the number of issues addressed or not addressing some of them effectively – as evidenced in the findings – may actually contribute to backlog in access to office visits because more demand for return visits has been created to address unresolved issues.

Management of diverse demand with heterogeneous level of professional and personal uncertainties

Unlike specialty care visits in which patients usually seek service for one particular condition, for example, rotator cuff injury, carpel tunnel syndrome, that is within the expertise of the specialist, primary care office visits routinely involve multiple patient complaints that could reach beyond the expertise and comfort zone of primary care physicians. Professional and personal uncertainties pose additional layers of complexity in physician's micromanagement of heterogeneous patient demand. Nevertheless, generalists are expected to address more complex issues. Whether what they do meets this expectation is an empirical question. Direct observation study can facilitate better understanding of how well they perform to the standards and what system modifications or policy changes are needed to complement or substitute some work of the primary care physicians to reduce inefficiencies in the agency relationship.

Whether and how to respond to demand for treatment for mental illnesses

Depression is the most common mental illness that is encountered by primary care practitioners who deliver most of its treatment, especially for elderly patients. Depression treatment practice guidelines call for at least four office visits, with counseling on mental health problems lasting for at least 5 min. They also advocate educating patients about treatment options, including medications' mechanisms of action, costs, risks, and benefits. Although treatment guidelines have been developed based on clinical research and expert opinion, the 'meaning' of these standards in terms of routine medical practice is not well understood. Detailed aspects of guidelines are rarely applied in quality assessment studies. The video and audio data enabled the authors to explore in detail how primary care physicians managed patient demand for treatment for depression in routine office visits.

It was found that the median length of time spent addressing depression was only 2 min, during which the patient spoke for 1.2 min and the physician spoke for 0.8 min. Furthermore, it was found that just because a physician has seen a patient with a mood disorder for an appropriate number of visits and prescribed a psychotropic agent or even multiple agents, the appearance of adherence to current guidelines does not necessarily mean that the patient received good mental health treatment. The authors explain in details below with findings from qualitative analyses of video recordings.

Qualitative analysis of the video recordings of visits during which mental health was addressed revealed three themes which characterized how physicians managed patient's demand for depression treatment at the point of care. The first theme was taking the time to investigate the disease and the patient as a person. The visit with the longest time on mental health discussion (17 min) in the whole study sample was a visit by a 69-year-old white male who broke into tears when his physician asked him how things were. The physician explored carefully and confirmed that the patient was depressed and suicidal, with a plan to use a revolver (already loaded) in the bathtub to end his life. Although the physician was very thorough in his assessment, the treatment plan was inadequate when compared to guidelines on addressing suicidal patients. He asked the patient to give a 'no suicide contract' and asked him to call a psychiatrist.

The second theme was allocating some time to gathering information, recognizing depression, but giving inadequate treatment. A case in point of this theme was a series of three recorded visits between a female patient and a male physician (Example 2 at the beginning of the article). The time allocated to mental health in each visit was 9, 5, and 11 min, respectively. Over approximately 7 months, this patient was sequentially prescribed paroxetine hydrochloride (10 mg for 6 weeks), fluoxetine hydrochloride (10 mg for 2 weeks), venlafaxine hydrochloride (37.5 mg for 6 weeks), and bupropion hydrochloride (unknown dose for 4 weeks) and taken off of Lorazepam, which she had taken for a long time. In one of the visits, the physician turned her raised hand for inquiry into a good-bye handshake. The management of her depression and anxiety had deviated from guidelines. For instance, low dose and short course of the antidepressants could have rendered these efficacious medications ineffective. Furthermore, stopping Lorazepam abruptly could increase withdrawal symptoms, potentially compounding anxiety. Despite the deficiencies in how her conditions were managed, research or quality improvement efforts based on claims data would have characterized these visits as guideline concordant, because only visit frequency was observed.

The third theme was physician dismissing patient's cue and indications of emotional distress. Five consecutive visits between a female physician and a female patient were seen, in which perfunctory and dismissive treatment of a patient's emotional distress was apparent. This patient was hospitalized to receive a stent after percutaneous transluminal coronary angioplasty. 2 min and 40 s were spent on the patient's emotions:

Physician: What you been up to?
Patient: I have just been crying my eyes out. (Crying…)
Physician: Why?
Patient: I don't know. I can't help it. (Crying…)
Physician: Why?
Patient: And then people ask me how I am, I just cry. (Crying…)
Physician: Oh (pause). Well I am not going to ask you that anymore.

This physician's paternalistic model of medical practice did not alleviate the patient's suffering. This is a case in which the 2-min mental health care clearly failed, because the patient left the visits with her depression neither evaluated nor treated. Such omission could impede her healing from the heart disease. It was somewhat surprising to see that, in a postvisit survey, the patient was satisfied with her visit and continued to return to the same physician for her care.

Whether responding to patients' clues would lengthen visits

Patients often give clues of distress and invite their physicians to respond. Although some physicians respond immediately, others choose not to respond for fear of sinking too much time if they were to respond. Communication researcher Levinson *et al*. examined the length of visit and patients' presentation of clues and found no evidence that responding to clues lengthens visits. Actually, visits in which a physician responded to a patient's clue were shorter than when the physician missed the opportunity. For primary care, visits without clues were a mean 15.7 min. Those with one clue were 12.7 min. Visits were longer (20.1 min) when there was a missed opportunity, compared to visits where the physician demonstrated at least one positive response to a clue (17.6 min). Visits in which patients repeatedly brought up emotional issues after the physician missed an opportunity to respond to a clue were longer than those with a positive response (18.4 min).

Whether discussion on a topic ends with an explicit decision

From the perspective of clinical communication, a decision can be defined as a verbal commitment to an explicit action. A clearly stated decision can facilitate a cognitive closure in the minds of the patient and physician that the discussion on a particular topic has reached an end. Communications research repeatedly documented a deficit in informed decision making in routine office visits and the lack of clear understanding patients have about what they needed to do after they leave their visits. There was a gap in knowledge on how often explicit decisions are actually made when discussion on a topic ends. The proportion of topic discussions in the sample that ended with an explicit decision was examined. The findings suggested, while the majority of topics ended with a decision (77%), there were variations related to the content and dynamics of interactions. Topics in which patients spoke more (67 s) were more likely to end with an explicit decision. Larger number of topics in a visit was associated with lower probability of a topic ending with a clear decision.

In summary, Arrow and colleagues have commented on the challenges in monitoring agents when their actions are unobservable. When the authors studied patient–physician interactions captured in video or audio recordings, they had an invaluable opportunity to observe agent actions. Combined with data on patient behaviors, a more nuanced understanding of how physicians manage patient demand at the point of care was gained. It could be seen that physicians have been observed to be habitual in their management of time during office visits, subject to influence of other demands presented by patients, and their own familiarity with the issues. In observing longitudinal visits between the same dyads of patient and physician, it is also noticed that, in this 'repeated game' context, patients return to the same physicians even though their previous clues of their desire to receive mental health services were overlooked or dismissed. It

appears that Albert Hirschman's encouragement to individuals to exercise their abilities to exit or use their voice in order to show their dissatisfaction is challenging for patients to carry out. The assumption of full information rarely holds true in medical decision making. Physician behavior often deviates from profit maximization expected of a firm. Simple extensions of the profit or utility maximization models may not produce satisfactory explanations of principal and agent behaviors. Simon's 'Satisficing' under constraints model offers a more plausible explanation of some of their behaviors. How behavioral economics may offer promising perspectives to study these behaviors are briefly discussed below.

Perspectives from Behavioral Economics

Health care exchanges, physician practice in particular, are fertile grounds for behavioral economics research. Yet the bulk of the application of behavioral economics to issues in health economics has been on patient behaviors, particularly addictive behavior around cigarettes, drugs and alcohol, and unhealthy lifestyles. Physician behavior has just begun to be subject to investigations guided by behavior economics perspectives. Some examples of physician behaviors that can be explained by behavior economics perspectives are elucidated here.

Use of Heuristics

An important finding from behavioral economics is the use of heuristics by decision makers that works reasonably well over a broad array of circumstances, but can be far off the mark in others. Following a norm or what Frank and Zeckhauser refer to as a 'ready-to-wear' treatment would be one such heuristic. The choice of heuristics implies less attention to purposeful optimization which is consistent with Simon's satisficing behavior. Humans, as opposed to Econs, have frailties. Often, the cognitive resources to maximize is lacking: the relevant probabilities of outcomes is usually not known, all outcomes with sufficient precision can rarely be evaluated, and the memories are weak and unreliable. Wennberg told the story of physicians recommending tonsillectomy for certain percentage or number of recently seen patients. Approximately 40% of children previously deemed not needing surgery were recommended for surgery at each subsequent waves of examination by additional physicians. The findings from direct observation data are consistent with this notion.

Attribution Bias

Context under which decisions are made needs to be taken into consideration. Loewenstein argues against 'context free' thinking because visceral and emotional factors can affect decisions in unexpected ways. For example, people may overattribute other people's behavior to personal dispositions whereas overlooking situational causes or transient environmental influences on behaviors. In doing so, the decision maker falls prey to attribution bias. Case study findings suggest that some physicians overlook the effects of inaccessibility of healthy food choices and walking paths in low-income neighborhoods, or other social determinants of health and overattribute the obesity problem to obese individuals being lazy.

Groopman told a gut-wrenching story of an elderly African-American patient being labeled as noncompliant who suffered from congestive heart failure, diabetes, hypertension, coronary artery disease, and advanced rheumatoid arthritis. She had been repeatedly admitted to a major academic medical center. None of her previous physicians knew that she was unable to read the labels on the medicine bottles until an African-American internist recognized what the other physicians had overlooked. This physician paid attention to the social context of severe disadvantages of being a black woman in the rural Mississippi of the 1930s and was able to arrange for the patient's daughter to be present at discharge and be informed of plans for care at home. The patient's recovery was remarkable afterwards.

Such attribution bias appears to be fairly common. Patients' weights significantly affect how physicians view and treat them. Patients with higher body mass index are also less likely to be perceived by physicians as medication adherent. Physicians ordered more tests for obese patients, spent less time with them, and viewed them with more negativity than nonobese patients. In a study of patient–physician communication over management of chronic pain, a physician was observed to be telling an elderly female African-American patient with disabling knee pain: "all you need to do is to lose 50 pounds. So you won't be hobbling around with all the extra weight on your knees".

Anchoring and Availability Bias

Frank and Zeckhauser termed the 'My Way Hypothesis' for situations in which physicians would regularly prescribe a therapy that was quite different from the choice that would be made by their peers. Although it is possible that the physician chose that therapy because she had differential expertise, the My Way Hypothesis may also apply when a physician has had personal 'good luck' with it, a plausible heuristic, but one that falls prey to the availability heuristic, namely overweighting evidence that one can bring easily to mind. Groopman recounted an emergency department (ED) physician using a 'studied calm' approach to avoid anchoring and availability biases when a patient presented with symptoms suggestive of a kidney stone. Rather than going along with the kidney stone diagnosis made by the triage nurse, the ED physician asked what might be the worst-case scenario thereby avoiding these cognitive biases and correctly diagnosed a dissecting abdominal aortic aneurysm, a far more life-threatening emergency.

Therefore, there is systematic evidence that heuristics can frequently lead decision makers astray, particularly when probabilistic outcomes are involved, as is almost always the case with medicine. The power of the field of behavioral economics has developed from the broad insight that heuristics can lead to significantly suboptimal behavior. Application of behavior economics perspectives can help advance the understanding of micromanagement of patient demand.

What Is the Value of This Research in Improving The Functioning of Patient–Physician Interaction?

Rather than accepting the notion that agent behaviors are unobservable and noncontractible, research efforts using direct observation data available in other fields has shed some light on agent behaviors at the point of care were applied. Building on the insights from behavior economics, coupled with empirical evidence from direct observation of physician–patient interaction, a more informed understanding of how physicians manage demand at the point of care, more like Humans with frailties, rather than Econs who have full information and can do probabilistic decision analysis on the go can be attained. Better point-of-care clinical decision support systems, redesign staffing structure to provide effective support, implement incentives that are conducive for escaping the lull of the norm may be designed. Shared decision making has been shown to lead to better honoring of patient's wishes and lower procedure-based service use. This line of research might continue to contribute to improving patient–physician interaction.

Coding communications within visits at the topic level is time consuming. Establishing inter- and intrarater coding reliability takes much effort and time. Some may question if the effort is worthwhile. It has been asserted that the findings have value in improving the functioning of patient–physician interaction. For example, it is noted that many topics compete for visit time, resulting in small amount of time being spent on each topic. A highly regimented schedule might interfere with having sufficient time for patients with complex or multiple problems. Efforts to improve the quality of care need to recognize the time pressure on both patients and physicians, the effects of financial incentives, and the time costs of improving patient–physician interactions.

Where This Research Area Should Go?

To understand how physicians manage patient demand, Simon and Fuchs' admonishment about using good data must be adhered to. Direct observation using video or audio data has offered unique insights and can continue to do so. Multidisciplinary collaboration with researchers in other fields (e.g., health communication and medical education) for data and communication analysis empirical approaches can continue to be a fruitful endeavor. It would be important to have large enough sample size to enable the examination of causal relationship between communication characteristics and downstream patient-reported outcomes.

Increasingly, patients are communicating with physicians asynchronously via secure messaging through the electronic health records (EHR) and personal health record. The method for studying patient–physician communication must evolve accordingly to take advantage of the EHR as an additional source of data for data mining. Some EHR, for example, EpicCare(EPIC) EHR, offers an unobtrusive portal to study time use through analysis of EPIC access log, a feature in EPIC EHR designed for monitoring access to patient's EHR for security and privacy concerns. The EPIC access log tracks the user of the EHR, time of access, device from which the access was made, and EHR functional location of the access, for example, progress note, medication list, phone encounter, and secure messaging. EHR enables the authors to leverage existing 'behind-the-scenes' data to study how much time clinicians spend on performing tasks. Natural language processing software is making progress in harvesting useful information from this data source to inform research on physician behaviors. Continued effort can bring promises to the field.

Sensitivity to institutional changes taking place in health care is essential. The health care delivery system is undergoing fundamental changes. To be relevant, one must be mindful of the institutional context and delivery system characteristics – dynamic rather than static – in which physicians manage patient's demand at the point of care. The redesign of clinical care processes and payment incentives can be informed by this type of research. For instance, it has been observed that quite a bit of time is being spent on listening to the lung, doing the traditional litany of system review where old information was rehashed with no apparent value. For example, if a patient's grandparent died of cancer 20 years ago, repeating this information at every periodic health exam offers no new information. How much value do these clinical routines offer? Should reimbursement continue to be triggered by following these routines? Substituting tradition-based medicine with evidence-based medicine may free up some time for physicians to do more shared decision making on more important issues in patient's view at the point of care to maximize the benefit of time.

Even the definition of point of care needs to be expanded to accommodate new models of care delivery. For example, the emergence of team care makes the function of care managers and other providers on the team important forces that might affect how patient demand is managed in the world of team care. The literature is silent on how team communicates among its members and with patients, let alone the impact on demand.

Tools designed to reduce information asymmetry and uncertainty in decision making are being developed, tested, and prescribed for patients. These decision aids – for example, for prostate cancer, breast cancer, depression treatment, and end-of-life care preferences – are making significant changes in patients' understanding of options and assisting them to alter their demand for health care services. How physicians use these decision aids and how they respond to patient's modified preferences can be important areas of research. As the field pays more attention to patients with multiple chronic conditions, more refined clinical decision support systems that can be tuned to accommodate demands from multiple morbidities will be a welcome addition to practicing physicians. Direct observation data can also provide unique insight on how demand from each condition is managed at the point of care with multimorbidity patients. For those interested in accessing the video recordings that have been used in the research, they may access them if they are medical educators or researchers working to improve the doctor–older patient relationship.

See also: Medical Decision Making and Demand. Physician-Induced Demand. Rationing of Demand

Further Reading

Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. (1997). Labor supply of New York City cabdrivers: One day at a time. *Quarterly Journal of Ecnomics* **112**, 407–442.

Frank, R. G. and Zeckhauser, R. J. (2007). Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices. *Journal of Health Economics* **26**(6), 1101–1127.

Groopman, J. (2007). *How doctors think*. New York: Houghton Mifflin Company.

Hirschman, A. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Cambridge, MA, USA: Harvard University Press.

Hsiao, W. C., Yntema, D. B., Braun, P., Dunn, D. and Spencer, C. (1988). Measurement and analysis of intraservice work. *Journal of American Medical Association* **260**(16), 2361–2370.

Huizinga, M. M., Bleich, S. N., Beach, M. C., Clark, J. M. and Cooper, L. A. (2010). Disparity in physician perception of patients' adherence to medications by obesity status. *Obesity* **18**(10), 1932–1937.

Levinson, W., Gorawara-Bhat, R. and Lamb, J. (2000). A study of patient clues and physician responses in primary care and surgical settings. *Journal of the American Medical Association* **284**(8), 1021–1027.

McGuire, T. G. (2000). Physician agency. In Newhouse, J. and Culyer, A. (eds.) *Handbook of health economics*, vol. 1A, pp. 462–517. Amsterdam: Elsevier.

Mishler, E. G. (1984). *The discourse of medicine, dialectics of medical interviews*. Norwood, NJ: Abex Publishing Corporation.

Murray, M. and Berwick, D. (2003). Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association* **289**(8), 1035–1040.

Simon, H. A. (1958). Theories of decision-making in economics and behavioral science. *American Economic Review* **49**, 253–283.

Tai-Seale, M. and McGuire, T. (2012). Time is up: Increasing shadow price of time in primary-care office visits. *Journal of Health Economics* **21**(4), 457–476.

Thaler, R. and Sunstein, C. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. New York, USA: Penguin Books.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157), 1124–1131.

Wennberg, J. (2008). Commentary: A debt of gratitude to J. Alison Glover. *International Journal of Epidemiology* **37**(1), 26–29.

Relevant Website

http://www.nytimes.com/2007/02/06/health/06insi.html
New York Times.

Physician Market

PT Léger, HEC Montréal, Montreal, QC, Canada
E Strumpf, McGill University, Montreal, QC, Canada

Glossary

Accountable care organizations Networks of physicians, or physicians and hospitals, that contract with insurers.
Any willing provider laws Restrictions on insurers' ability to form exclusive networks by requiring them to contract with any provider who is willing to join and meets the network requirements.
Board certification A process by which physicians demonstrate their knowledge and skills in a particular specialization.
Capitation A payment mechanism where physicians receive a predetermined fixed fee for each patient they enroll in their practice in exchange for care for a fixed period of time without any marginal reimbursement.
Cournot oligopoly A market dominated by a small number of sellers where each firm makes quantity decisions taking the optimal strategies of all competitors into account.
Credence good A good whose value is difficult for the consumer to judge even after consumption (in contrast to an experience good).
***Ex post* moral hazard** The situation where patients use excessive amounts of care because they face a subsidized price due to insurance.
Fee-for-service A payment mechanism where physicians receive a predetermined fixed fee for each billable service they provide, which is generally at or above the marginal cost of production for that service.
Independent practice association A network of independent physicians that contracts with managed care organizations and employers.
Managed care organizations Organizations that integrate health insurance and healthcare service delivery and also place restrictions on which providers can been seen and the use of some services.
Monopolistic competition A market where firms or providers offer differentiated products, allowing the suppliers some market power in the short run.
Numerus clausus Refers to the limit on the number of students who may enter a program of study (e.g., medical school).
Physician-Hospital Organizations Joint ventures between hospitals and physicians.
Preferred provider organization A form of managed care where enrollees face lower prices when they see in-network providers.
Scope of practice The procedures, actions, and processes that different healthcare practitioners are licensed (or permitted) to perform.
Supplier-induced demand A situation in which physicians convince patients to demand excessive amounts of care in the presence of asymmetric information.
Switching costs Any cost, psychological, financial, or otherwise, associated with leaving one's current physician and finding an alternative source of care.

Introduction

A market is generally defined as a set of firms or individuals selling similar (or at least partially substitutable) goods or services to a given set of consumers. Under some basic conditions, competitive markets yield Pareto optimal outcomes (The First Optimality Theorem). As a result, policymakers and regulators have generally favored competition and prohibited and punished anticompetitive behavior. However, many researchers have argued that: (1) many of the basic conditions required for the existence, and Pareto efficiency, of a competitive equilibrium are unlikely to be satisfied in the healthcare market and (2) in the presence of deviations from the necessary conditions, competition may actually be welfare decreasing. Important deviations from the model of perfect competition in the physician market include imperfect and asymmetric information, differentiated products, administratively set prices, insurance, and barriers to entry (Table 1). Simply put, conditional on being in a second-best world, the effect of competition on quality and cost of healthcare may not necessarily be positive.

Although clearly being different from markets that fit the requirements for perfect competition, the market for physician services does resemble other markets involving asymmetric information, expertise, or credence goods. Markets for car repair, legal advice, and taxicab rides are classic examples with the potential for inefficient outcomes and scope for government intervention. However, the market for medical care suffers from a larger number of imperfections than most, moving quite far from the first-best being a relevant welfare benchmark. In addition to the information asymmetries seen in these other markets, insurance and administratively set prices are notable attributes of the market for healthcare and physician services which complicate the welfare implications of competition and many of the conclusions drawn from nonhealthcare analyses.

Consistent with not using the first-best as the main point of reference, physicians have been seen by many (including lawmakers and courts) to be outside the scope of competition policy (in part because of their role as 'learned professionals' in the presence of imperfect information). In 1975, the US Supreme Court ruled that learned professionals (including

 doi:10.1016/B978-0-12-375678-7.01107-X

Table 1 Deviations from first optimality theorem assumptions

Assumption	*Deviations/examples*
Full and symmetric information	1. Patients may have partial information on: a. their need for care (or illness severity) b. treatment options c. quality of care/provider d. prices 2. Information on (a) through (d) may remain partial even posttreatment in part because of an uncertain treatment-outcome relationship 3. Physicians may have partial information on: a. their patient's illness severity b. their patient's history c. their patient's preferences, income 4. Insurers may have partial information on the patient's illness severity and need for care 5. Insurers may have partial information on the care received
Differentiated products	1. Physicians provide heterogeneous services: a. diagnostic skills b. treatment style-skills c. bedside manner d. hours of operation e. location 2. Also differentiated by virtue of the patient–physician relationship: a. trust 'capital' b. knowledge of medical history
Absence of search and switching costs	1. Patients may face important search costs: a. collect information on physicians b. find out who is taking patients c. inferring or predicting alternative physicians' quality 2. Patients may face important switching costs when leaving one physician for another: a. rebuilding trust b. informing the new physician regarding medical history c. potential monetary costs or penalties
Competitive prices	1. Administratively set prices are by construction a violation of this assumption, i.e., prices cannot adjust to equate demand and supply 2. Price setting varies from one jurisdiction to another
Consumers face true prices	1. Presence of insurance an obvious violation of this assumption 2. Insured patients face either a fully or partially subsidized price depending on the deductible and copayment
No barriers to entry or capacity constraints	1. Barriers to entry in the medical market include: a. medical school admission restrictions b. licensing and other regulations (especially important for outsiders) which can include retraining, testing, internships, and language requirements c. limits on nonphysician competition through scope of practice regulation (nurses, midwives, and pharmacists)

physicians and the entire healthcare market) were indeed subject to antitrust laws, while nonetheless recognizing the particularities of the environment and the potentially perverse role of competition. In other countries, including Canada, Germany, and the Netherlands, lack of price competition among physicians is institutionalized as prices are explicitly negotiated between the government and physician associations.

Understanding several aspects of the economics of the physician market is necessary to address the aforementioned debate. First, what is the level of competition in the physicians' marketplace? Second, what factors contribute to its relative strength? Third, what form does it take (e.g., price vs. quality competition)? The idea of competition in the physician market is relevant to all contexts and healthcare systems but its form and level are dependent on the institutional context (e.g., payment mechanisms and regulations regarding balance billing) and the economic environment (e.g., the relative supply and demand of healthcare services). Although little empirical work exists investigating the market power of physicians or the welfare implications of reduced competition in the physicians' market, examining the characteristics of the physicians' market can provide a sense of the form and level of competition present in different jurisdictions.

Beyond describing competition in the physician market, its normative implications must be considered. In the context of the second best, it is unclear whether competition in the physician market will improve or harm social welfare. Theoretical work (Allard *et al.*, 2009) has shown that, in a mixed-payment system, competition may provide appropriate incentives even in the presence of noncontractible effort, information asymmetry, unobserved heterogeneity, switching costs, uncertainty, and regulated prices. More specifically, competition makes a patient's 'threat' to seek care elsewhere

more credible, and may serve as an important incentive to provide appropriate care on both contractible and noncontractible dimensions. However, it may also lead to treatment heterogeneity, overtreatment (i.e., defensive medicine) and unstable physician–patient relationships. In this article the authors examine some of the aforementioned necessary conditions for the existence and Pareto efficiency of a competitive equilibrium as a first step to understanding the broader issue of competition and welfare in healthcare markets. The authors also review conclusions from the empirical research in this area.

To address these different issues, the authors start by defining the 'players' in the physicians' market (i.e., who competes against whom), a key first step in any study of the level, form, or welfare impacts of competition in the physician market. The market is defined simultaneously by the set of suppliers and consumers in a geographic area, but geographic considerations and 'firms' will be discussed sequentially for ease of exposition. The authors first identify the set of consumers over which physicians (or other healthcare providers) compete (generally defined by a geographic area such as a zip code or Metropolitan Statistical Area (MSA)). Next, to determine the level of competition within a given market, the services over which different types of providers might compete will be considered. Not surprisingly, physicians compete with providers of their own specialty but also across specialties and even with healthcare providers who are not physicians. Once the competitors and their potential customers are identified, the authors then separately examine the different elements that may limit competition. Although most of the empirical evidence discussed here is set in the North American context, this is primarily due to greater interest in physician competition there by researchers and policymakers. The general issues and theoretical frameworks discussed here are, however, common in many jurisdictions and the underlying economic principles remain the same.

The remainder of this article is organized as follows. In Section Who Are the Patients/Consumers That Physicians Compete over?, the physicians market in terms of customers is examined, i.e., who do physicians compete over? In Section Who Competes against Whom?, the market in terms of sellers is discussed, considering the possibility that physicians may compete against nonphysician healthcare providers. In Section The Competitive Model, the authors discuss the necessary conditions for the existence of a Pareto efficient competitive equilibrium and corresponding welfare implications, and how the physicians' market may deviate from such conditions. Conclusions are drawn and remarks made in Section Conclusion.

Who Are the Patients/Consumers That Physicians Compete Over?

Physicians compete over a set of potential patients, generally defined by their geographic location. Although such geographical areas may be simple to define (e.g., all physicians within a city limit are assumed to form one market), they may not accurately reflect the actual market. That is, the set of physicians that a patient will consider likely depends on several additional factors including: (1) the type (e.g., specialty) and quality of services provided, (2) the distance to travel, (3) whether a given provider is covered by the patient's insurance, and (4) prices. For example, a patient may be willing to travel longer distances for nonemergency care than emergency care in order to get lower prices and/or greater quality and therefore considers seeking care from a wider set of physicians.

The question of prices that patients face is intricately tied to that of insurance for healthcare services. Because widespread, fairly comprehensive insurance coverage for physician services is the norm throughout developed countries, consumers of and payers for services are often not the same party. Physicians' incomes can come from a mix of public and private payers (both insurance companies and individuals), one that varies significantly across countries, regions, and medical specialties, but rarely is comprised mostly of direct, out-of-pocket payment by patients. Insurers often restrict services they pay for to those provided by providers they contract with. In Canada, this means any physician participating in the provincial health insurance plan (nearly all physicians), whereas in the US, a managed care plan may have a more restrictive network of providers. As a result, the number of insurers in a market and the ways in which they steer patients to certain providers (if at all) can have implications for whether and how physicians compete for patients directly or whether they compete for contracts with insurers. Although the presence of insurance in general may make patients less responsive to price or quality differences across physicians (Section Insurance), payers can also play important roles in reducing information asymmetries between patients and physicians and can create direct incentives for patients to seek out higher quality physicians.

Relatively few empirical studies that quantify physician markets exist, but this question can be informed by the larger literature on competition in hospital markets which faces similar issues. Studies that use geographical areas like Primary MSAs or actual consumption patterns to define markets (i.e., competitors) suffer from creating barriers that patients often traverse and from generating bias if individuals choose a particular hospital based on unobservable characteristics like quality, which may also be correlated with market share. Take, for example, a physician who is of such low quality that they only attract patients in their immediate vicinity (e.g., in the immediate zip code). By calculating the market shares (and thus corresponding level of competition via a measure like HHI) using zip codes as markets, this physician may appear to face little competition (and thus benefit from a large market share). This conclusion, however, neglects the fact that many patients in neighboring zip codes may consider this particular physician as part of the choice set but choose not to seek care because of his or her poor quality. To deal with this potential bias, recent studies have used predicted market shares based on exogenous factors rather than actual shares.

Researchers who wish to define physicians' markets and measure their degree of competition should consider similar potential biases. The authors are not, however, aware of any papers that estimate the level of competition in actual physicians' markets in such a manner or which estimate individual physician's market share. As a result, the discussion relies on

geographical markets rather than estimated potential markets, although recognizing their limitations. Although not being the sole indicator of competition, the extent to which market characteristics (i.e., provider density) vary across jurisdictions can be instructive regarding the degree of competition.

Who Competes against Whom?

Even once the geographic boundaries of a market have been defined, identifying who competes against whom in healthcare (and thus which providers constitute a particular market) is complicated by the fact that: (1) different types of providers will provide (im)perfect substitutes for each other and (2) different types of providers may have overlapping scopes of practice. Obviously, general practitioners (GPs) compete against other GPs, whereas obstetricians compete against other obstetricians (i.e., within-specialty competition). However, GPs may also compete with other specialists (from gynecology to psychiatry) as they have considerable overlap in their scope of practice.

Physicians may also compete across specialties because different types of specialists provide alternative treatments for the same health problem (e.g., 'stenting' by a cardiologist or bypass surgery by a cardiothoracic surgeon). As a result, when considering the market for a particular health problem or medical service, one may have to consider different types of physicians. Physicians may also compete with allied health professionals (AHPs) such as physician assistants (PA) and advanced practice nurses (APN) (which include nurse practitioners (NP), certified nurse–midwives, clinical nurse specialists and nurse anesthetists). Although the allowable scope of practice and level of independence of each of these groups varies from one jurisdiction to another, both have increased over the years. Dueker *et al.* (2005) found that in states where they are given a lot of professional independence, APNs have lower salaries (whereas their PA counterparts have higher salaries). They argue that this is because physicians in states where APNs are granted greater professional independence substitute away from APNs toward PAs in their hiring decisions for fear that they will take on greater roles in hospital settings. This is not surprising as the authors also find that greater independence of APNs negatively impacts physicians' salaries. Finally, physicians are likely to compete with other 'nonmedical' healthcare providers such as chiropractors, osteopaths, and acupuncturists. As a consequence, a more complete vision of the market should consider these alternatives.

The legal context that frames the physician market includes licensing and scope of practice regulations. These largely determine the market, both geographically and with respect to specific services, and establish the 'rules of the game' in which competition flourishes (or doesn't). These frameworks vary across jurisdictions, which can include an entire country (physician licensing and registration by the General Medical Council in the UK) or encompass states or provinces (provincial-level licensing in Canada). Although potentially helping to disseminate information and maintain or improve quality, these regulatory mechanisms are also likely to reduce competition, issues that will be discussed further in Section Limits to Supply and Barriers to Entry.

The Competitive Model

As noted previously, competitive environments yield Pareto efficient outcomes under specific conditions. If they are met, deviations from the competitive equilibrium lead to losses in social welfare. Because physicians provide differentiated services, entry into the market is strictly controlled, prices are often set administratively and patients generally do not face true prices because of insurance, physicians may have considerable market power. However, because the conditions underlying the First Welfare Theorem are unlikely to be met in the healthcare setting (Arrow, 1963), it is not obvious that such market power decreases welfare relative to greater competition.

In the context of concerns about physicians' market power, accusations of collusion and anticompetitive behavior in the physicians' market have become more prevalent and the US Department of Justice concluded that physician collusion when negotiating with third-party payers was in violation of the Sherman antitrust rules. Recent empirical work suggests that in spite of these prohibitions, physicians benefit from considerable market power. More specifically, Wong (1996) found that the GP and family physicians' market is consistent with monopolistic competition (and not with monopoly or perfect competition), whereas Gunning and Sickles (2010) found that the medical and surgical specialists' market is consistent with Cournot oligopoly.

In the following subsections the authors discuss each of the conditions required for a competitive market to reach efficient outcomes in greater detail. Beforehand, however, the authors present some summary statistics on the density of different types of physicians across different jurisdictions in order to get a sense of the level of competition (or at least, how it may vary across different specialties and jurisdictions). Given that physicians are unlikely to vary greatly in terms of individual market shares, market-level HHIs are unlikely to provide much more information than simple concentration ratios.

Across Organization for Economic Co-operation and Development (OECD) countries, the number of physicians averaged 3 per 1000 population in 2007 and the average number of GPs was 0.9 per 1000 and 1.8 specialists per 1000. Considerable variation also exists across countries in each of these measures (**Table 2**). The number of physicians ranged from 1.5 per 1000 in Turkey to 5.5 per 1000 in Greece. Specialists greatly outnumbered generalists in central and eastern European countries and in Greece. Other countries, notably Belgium, France, and Portugal, have maintained a more equal balance between specialists and generalists. This variation in provider densities suggests differences in the degree of competition in physician markets across these jurisdictions.

The number of nurses averaged approximately 9 per 1000 population across OECD countries, ranging from 1.3 per 1000 in Turkey to 16 per 1000 in Denmark. The number of midwives averaged 72 per 100 000 women, ranging from 1 per 100 000 women in the US to 178 per 100 000 women in Australia. Such variation in the densities of providers that are

Table 2 Physician supply across selected OECD countries, 2007

	Practicing physicians per 1000 population	*General practitioners (GPs) per 1000 population*	*Specialists per 1000 population*	*Ratio of GPs to specialists*
Greece	5.4	0.3	3.4	0.1
Belgium	4.0	2.0	2.0	1.0
Switzerland	3.9	0.5	2.8	0.2
Sweden	3.6	0.6	2.6	0.2
Germany	3.5	1.5	2.0	0.8
France	3.4	1.6	1.7	0.9
OECD	3.1	0.9	1.8	0.5
Australia	2.8	1.4	1.4	1.0
UK	2.5	0.7	1.8	0.4
US	2.4	1.0	1.5	0.7
Canada	2.2	1.0	1.1	0.9
Japan	2.1	na	na	na
Mexico	2.0	0.7	1.3	0.5
Korea	1.7	0.6	1.1	0.5
Turkey	1.5	0.5	1.0	0.5

Notes: Data is from 2007, or the latest year available.

Practicing physicians are defined as the number of doctors who are providing care directly to patients. In many countries, the numbers include interns and residents (doctors in training). The numbers are based on head counts, except in Norway which reported full-time equivalents before 2002. Ireland, the Netherlands, New Zealand, and Portugal report the number of physicians entitled to practice (resulting in an overestimation). Data for Spain include dentists and stomatologists (also resulting in a slight overestimation).

Not all countries are able to report all their practicing physicians in the two broad categories of specialists and generalists. This may be due to the fact that specialty-specific data are not available for doctors in training or for those working in private practice.

Source: Reproduced from OECD (2009). *Health at a glance 2009: OECD indicators*. OECD Publishing. Available at: http://www.oecd-ilibrary.org/about/about;jsessionid=cb6gobckacpip.x-oecd-live-02 (accessed 13.08.13).

complements and substitutes to physicians can also have implications for competition in the physician market.

Studies show that demographic factors such as age and gender can also have important implications for the supply of providers and level of competition, because those with different characteristics often practice in different locations (i.e., hospitals vs. private practice), work fewer or more hours, and select different specialties. In 2007, on average 40% of physicians across OECD countries were women, ranging from more than half of physicians in central and eastern European countries and Finland to less than 20% in Japan. The share of women physicians has increased from 29% in 1990 across the OECD and from 20% to 30% in the US between 1990 and 2007, and now account for nearly half of all medical students there (NCHS, 2009). Characteristics such as gender, age, and race are important predictors of specialty choice for some specialties but variation in expected income across specialties is also important.

A third factor that is linked to the competitiveness of the physician market is the issue of adequate physician supply, which remains an important policy topic especially with respect to urban/rural differences. For example, in the US the number of practicing physicians per 10 000 civilian population was 25.7 in 2008, but largely rural states like Wyoming and Idaho had only 18.7 and 17.0, respectively, whereas Massachusetts had 39.7 and Maryland had 35.3 (NCHS, 2011). Low physician-to-population ratios in rural jurisdictions mean that those providers face less competition from other physicians. Furthermore, the range of explicit policies and incentives that encourage physicians to locate in 'underserved areas' (e.g., the US National Health Service Corps, Quebec's Differential Remuneration Program) clearly demonstrate that the market is not competitive in the sense that wages do not freely adjust to equilibrate supply and demand. Some of the factors that prevent such a wage adjustment are discussed in the following section.

Information and Physician Payment Methods

One of the defining characteristics of healthcare markets and the patient–physician relationship is the presence of incomplete and asymmetric information. Patients may have greater knowledge about their symptoms, habits, medical history etc., whereas physicians are likely to have privileged information about the patient's illness and alternative treatment options. It is precisely because of their medical expertise that physicians act as their patients' agent (i.e., they act on their patients' behalf regarding medical decisions).

Under the traditional fee-for-service (FFS) system, physicians receive a fixed payment for each service they provide that is above the marginal cost of production. If physicians are paid via a blended payment (partial capitation plus FFS), or if there are multiple insurers across which physicians can cross-subsidize, FFS payments need not be above marginal cost (e.g., the US Medicaid program).

As a result, the FFS payment system rewards the volume of services. This incentive may lead physicians to manipulate information that they provide to their patients in order to encourage them to consume more care than they would chose themselves if they were fully informed. This implies patients would consume units of care whose marginal benefit is less

than its marginal cost. The physician may manipulate the patients' information, and thus their choice, by either exaggerating the illness severity or the need for care. This phenomenon is known as supplier-induced demand (SID). If patients' demand is responsive to the quality of services, then FFS may also reward quality via volume. Of course, the distinction between quality that improves the value of healthcare services versus amenities that are perceived by patients as higher quality but do little to improve value beyond patient satisfaction is an important one.

Several theoretical models and empirical analyses tackle this issue. In a classic article, McGuire and Pauly (1991) presented a model of healthcare provision and costly inducement of care. In a setting with only one type of care (or patient), they show that physicians will wish to induce care in the presence of a decreased fee (in order to recoup lost earnings) only if the income effect dominates the substitution effect. In the thoracic surgery setting, one study found that surgeons whose fees were reduced under Medicare recouped some of the lost earnings by increasing the number of surgeries they performed.

McGuire and Pauly's model has two types of patients, with private or Medicare health insurance coverage, each with their respective reimbursement rate. They show that a decrease in one reimbursement rate (i.e., the Medicare fee) can lead to an increase in the amount of care provided to the now more generously reimbursed patient (i.e., those with private insurance), again if the income effect dominates. Another study found that obstetricians who faced more competition (through falling birth rates) responded to the potential reduction in their income by inducing demand for cesarean deliveries (which are more lucrative than vaginal births).

To reduce this incentive for volume, alternative forms of physician payments such as capitation and salary have been introduced. In a capitation system, physicians receive a fixed payment for each patient they enroll in their practice in exchange for care for a given period of time without any marginal reimbursement. By not tying a physician's income to volume, the capitation payment system completely eliminates the incentive to provide inefficiently high levels of care. Although this may reduce or eliminate SID, it has been associated with reduced quality and quantity of care (i.e., physicians may wish to manipulate information not to encourage care but rather discourage it). One study found that the increase in managed care has limited some of physicians' ability to induce care via supply-side constraints. Similarly, salaried physicians have no incentive to induce care but may have little incentive to provide the appropriate quantity and quality of care (as their income is invariant, at least in the short run, to such decisions). Although less competitive markets may allow physicians to exploit their market power (say, through SID) their willingness to do so will depend, in part, on the way that they are paid.

Differentiated Medical Services and Switching Costs

Although physicians may provide similar sets of services, they are likely to be different along several dimensions like quality (including time spent with the patient, diagnostic skills, treatment recommendations, bedside manner). Even in the presence of a common specialty and types and levels of quality, patients are unlikely to see their physicians as perfect substitutes by simple virtue of the patient–physician relationship. More specifically, patients may have developed a sense of trust with their regular physician. Furthermore, the physician is likely to have privileged information regarding the patient's health history and need for care. Because of these aspects, patients may face important switching costs when seeking care from an alternate provider and physicians are likely to exert some degree of market power. One way physicians may exploit such market power would be to price discriminate across patients. It is precisely the potential for this behavior that is at the root of uniform administrative pricing for physician services. Although administratively set prices eliminate physicians' ability to price discriminate, they may nonetheless exploit their market power through other means (which in turn will depend on the level of competition). The authors discuss this issue in the next section.

Fee Setting and Organizational Response

The presence of administratively set prices is by definition a deviation from a competitive environment. Although administratively set prices may limit physicians' ability to exploit their market power through price discrimination (and thus, may in fact be welfare enhancing), it does not necessarily prevent physicians from exploiting the heterogeneity and nontradability of their services or the presence of switching costs (and this even in the presence of perfect information). By simply giving patients a 'take-it-or-leave-it' offer on the quantity of care to be provided, the physician may convince the patient to consume excessive amounts of care (i.e., yield higher profits for the physician and lower patient welfare; McGuire, 2000). For example, consider a patient who can consume care at p dollars per unit q. Further suppose that the patient's net benefit from consuming the utility maximizing amount of care is given by $B(q) - pq$. Now suppose that the patient faces a switching cost k if they decide to leave their current physician for an alternative who would be willing to provide q units (also at p). Switching costs could be monetary if tests or procedures need to be redone, or psychic in the sense of building a relationship with a new physician. If the patient leaves their current physician for an alternative one, they then receive a net benefit of $B(q) - pq - k$. Thus, the current physician can persuade the patient to consume q' units of care (greater than q) as long as $B(q') - pq' > B(q) - pq - k$. Thus, in the presence of administratively set prices, nontradable goods, and switching costs, physicians may hold important market power even in the presence of full information. The above suggests that increasing the patient's ability to move from one physician to another (i.e., decreasing the switching costs) is likely to yield important implications on physicians' market power and patient welfare (which is consistent with previous work (Allard *et al.*, 2009)). As a result, one could imagine that the presence of portable medical records and increased patient education may have important benefits in encouraging an efficient provision of care.

How the fees are actually set and how closely they reflect marginal cost pricing is also an important issue. In several systems, payment rates are primarily based on relative weights that reflect the costs of inputs used to provide physician services and a conversion factor which translates these weights into dollar amounts. The weights reflect factors such as physician work (time, effort, skill, etc.), practice expenses (rent, equipment, and staff), and professional liability insurance expenses and the medical profession often exerts significant control in determining them. Governments can use (or attempt to use) adjustments to conversion factors to limit spending on physician services and to constrain spending growth (e.g., US Medicare's Sustainable Growth Rate system).

In many jurisdictions, physicians have played an important role in setting these rates. In Canadian provinces, provincial physician associations negotiate directly with their respective governments to set the fees – fees which then apply to all physicians. In Germany, decisions made both by a government committee at the federal level and by regional physicians' associations affect the amounts physicians are paid for the services they provide to patients. In the US, however, reimbursement rates paid to the same physician can vary between insurance providers. For example, a physician might receive different reimbursement rates for the same procedure depending on the patient's (governmental or nongovernmental) insurer (from Medicare, to Medicaid, to private insurers). Furthermore, physicians are prohibited of collectively bargaining when negotiating reimbursement fees.

One of the ways that insurers have reduced the fees paid to physicians is through Managed Care Organizations (MCO). MCOs control costs through a variety of tools including selective contracting. Selective contracting refers to limits imposed on the pool of physicians that enrolled patients can consult with (i.e., a network of providers). By limiting their enrolees' access to providers, MCOs gain bargaining power relative to physicians. Although very strict forms of restrictive contracting were popular in the early years of managed care, less restrictive contracts such as Preferred Provider Organizations are more common now. These forms of managed care contracting also coexist with 'any willing provider' laws, which require managed care plans to explicitly state their evaluation criteria and to accept any qualified provider who is willing to accept the plan's terms and conditions. Selective contracting and 'any willing provider' laws are clearly designed to have opposing effects on competition in the physician market.

In response to the potential negative effects of MCOs on physician earnings, physicians have introduced different organizational structures in order to reduce costs or increase their bargaining power. Independent Practice Associations are networks of independent physicians that contract with MCOs and employers. Physician-Hospital Organizations, however, are joint ventures between hospitals and physicians. Emerging Accountable Care Organizations are networks of physicians or physicians and hospitals that contract with insurers. Although such networks may provide important efficiency gains, they run the risk of antitrust violations. Accordingly, 2010 legislation in the US prevents new physician-owned hospitals from being built and puts strict limits on already existing ones given the physician's obvious conflict of interest (and corresponding evidence that they may cherry-pick customers, avoid costly services and provide low quality of care).

Insurance

In many markets, consumers face a price that reflects the cost of production when making purchasing decisions. This is usually not the case in the medical services market however because of the presence of insurance. With insurance, consumers face a partially or fully subsidized price and they will wish to consume care until the marginal benefit of care is equal to their subsidized cost (i.e., their out-of-pocket cost) and not the actual marginal cost of production – a phenomenon known as *ex post* moral hazard. As a result, insurance can cause important deviations from the Pareto efficient outcome. To reduce this problem of *ex post* moral hazard (and thus, limit also its perverse effect) one can simply increase the share of the expense the patient pays for (i.e., increase the copayment). By doing so, however, one reduces the risk-spreading benefits of insurance.

Limits to Supply and Barriers to Entry

The condition of free entry is clearly not met in the case of the physician market, with significant barriers existing in training, licensure, and migration across jurisdictions. Nearly all OECD countries exercise some form of control over the number of medical school students – often in the form of a numerus clausus, or limit on the number of medical school slots – or residency positions. This is motivated by a variety of factors including restricting medical school entry to the most able applicants, controlling the total number of physicians for cost-containment reasons, and limiting the direct cost of training. Although countries vary in their approaches to these limits (i.e., ministerial decisionmaking vs. financial incentives) and the extent to which control is devolved to subnational governments (state/provincial/canton) or medical schools themselves, the net result is that the number of physicians is largely determined by policy.

Beyond the total number of physicians, the mix of physicians in different specialties is a function of the number of residency slots, which is largely determined by organized medicine. In the US, Residency Review Committees, consisting primarily of physicians from a particular specialty, control the number of residents who train in, and therefore the number of physicians who flow into, each specialty. Nicholson (2003) showed that these committees could set the flow of residents in order to achieve their desired combination of licensed physicians and physician rents and that minimum wage regulation and scarcity of teaching material (i.e., patients) could also result in the observed persistent excess rates of return to specialization.

After medical education and residency training are completed, entry into the market for both physicians and AHPs is restricted via licensing and practice regulations that can limit competition geographically and in terms of specific services provided. Three primary mechanisms are used to regulate physicians and AHPs, licensure, certification, and registration, and in the US this occurs at the state level. Licensure involves a

mandatory system of state-imposed standards that practitioners must meet to legally practice within the state. Nongovernmental boards, dominated by members of the regulated profession but often also including political appointees or members of the public, determine applicants' eligibility requirements, develop standards of practice, and enforce disciplinary actions. Certification refers to a voluntary system of standards, usually set by nongovernmental agencies or associations. Physicians may become board certified within a specialty to establish they have the appropriate level of knowledge and skills in that area. Registration is the least restrictive mechanism, requiring practitioners to file their name, address, and qualifications with a government agency in order to practice, but not to meet any additional educational or experience requirements.

As discussed in Section Who Competes against Whom?, competition in the market for physician services is also affected by nonphysician providers, the extent to which they are complements to or substitutes for physicians, and the laws and regulations that govern the range of services they provide and how they are remunerated. Feldstein (1988) describes how organized medicine has influenced the demand for, and supply of, physician services, notably by encouraging insurance coverage for their services, disallowing price competition among members, and restricting (expanding) the scope of practice of substitute (complementary) providers allowed under state practice acts. He also describes similar efforts by the American Dental Association, American Nurses Association, and American Hospital Association on behalf of their memberships. For example, the American Medical Association has favored the use of foreign medical graduates to serve as interns and residents (complements, because the physician bills for services provided) and also their return to their home country once their residencies are completed and they become substitutes. Similar tensions are seen between physicians and PA, NP, and chiropractors. Beyond limiting the legal scope of practice of substitute providers, efforts to exclude their payment by a third party has often occurred in the context of public insurance programs (US Medicare, Canadian provincial plans, etc.).

The last barrier to entry considered here is migration, or physical entry into the physician market. Internationally, OECD countries have adopted specific policies designed to stimulate immigration of foreign physicians although respecting the constraints of licensing requirements, protection of vested interests by domestic physicians, and minimization of any negative impacts on the home country. Even within countries, state, or provincial-level licensing requirements mean that physicians may face considerable barriers to practicing in another jurisdiction.

Conclusion

In this article the authors provided an overview of the physicians' market and the role of competition within it. Specifically, the authors noted that although the idea of competition is relevant in all contexts, the observed level and form of competition in any market for physician services will depend on the institutional context and the economic environment. Although the authors empirically observe indicators of the degree of competition in the physician market to vary across contexts, whether a lack of competition decreases social welfare remains an open question. The uncertainty regarding competition's impact on welfare in this market stems from the numerous market failures discussed in Section The Competitive Model. Although some of these deviations from the competitive model are likely inherent to markets for 'expert' services (information asymmetries, search and switching costs), others are more amenable to change via health policy and regulation (insurance, limits on entry, price competition, and provision of services by other providers).

In this second-best environment, the value of competition as a tool to increase welfare is unclear. Whether competition is welfare decreasing depends on the context, specifically how many market imperfections exist and how big are they? Put another way, how far are we from the first best? Competition may be a positive if it encourages the provision of high-quality care (e.g., heterogeneous physicians provide an appropriate mix and levels of care in the presence of information asymmetry and insurance). Competition may be undesirable if it encourages the wasteful use of resources in attracting and treating patients or increases the quantity of services without corresponding reductions in administered prices. For example, Kessler and McClellan (2000) empirically examined the impacts of competition in US hospital markets and found that, in that specific context and time, competition improved welfare. The authors expect that similar work to better understand the implications of physician competition, taking into account the specific institutional and economic contexts, will also be valuable.

Acknowledgment

Strumpf gratefully acknowledges funding support from a Chercheure Boursière Junior 1 from the Fonds de la Recherche du Québec – Santé and the Ministère de la Santé et des Services sociaux du Québec. Léger thanks HEC Montréal for funding.

See also: Advertising Health Care: Causes and Consequences. Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Empirical Market Models. Health Services in Low- and Middle-Income Countries: Financing, Payment, and Provision. Heterogeneity of Hospitals. Income Gap across Physician Specialties in the USA. Interactions Between Public and Private Providers. Internal Geographical Imbalances: The Role of Human Resources Quality and Quantity. International Trade in Health Workers. Learning by Doing. Managed Care. Market for Professional Nurses in the US. Markets in Health Care. Medical Malpractice, Defensive Medicine, and Physician Supply. Monopsony in Health Labor Markets. Nurses' Unions. Occupational Licensing in Health Care. Organizational Economics and Physician Practices. Physician Labor Supply. Physician Management of Demand at the Point of Care. Physician-Induced Demand. Physicians' Simultaneous Practice in the Public and Private Sectors. Preferred Provider Market. Primary Care,

Gatekeeping, and Incentives. Public Health Profession. Specialists. Switching Costs in Competitive Health Insurance Markets. Waiting Times

References

Allard, M., Léger, P. T. and Rochaix, L. (2009). Provider competition in a dynamic setting. *Journal of Economics and Management Strategy* **18**, 457–486.

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5), 941–973.

Dueker, M. J., Jacox, A. K., Kalist, D. E. and Spurr, S. J. (2005). The practice boundaries of advanced practice nurses: An economic and legal analysis. *Journal of Regulatory Economics* **27**(3), 309–329.

Feldstein, P. J. (1988). *The politics of health legislation: An economic perspective*. Ann Arbor: Health Administration Press.

Gunning, T. S. and Sickles, R. C. (2010). Competition and market power in physician private practices. Mimeo. Rice University.

Kessler, D. P. and McClellan, M. (2000). Is hospital competition socially wasteful? *Quarterly Journal of Economics* **115**(2), 577–615.

McGuire, T. G. (2000). Physician agency. In Culyer, A. J. and Newhouse, J. P. (eds.) *The handbook of health economics*, pp. 461–536. Amsterdam: Elsevier.

McGuire, T. G. and Pauly, M. V. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics* **10**(4), 385–410.

National Center for Health Statistics (NCHS) (2009). Health, United States, 2008. Hyattsville, MD: National Center for Health Statistics.

National Center for Health Statistics (NCHS) (2011). Health, United States, 2010: With special feature on death and dying. Hyattsville, MD: National Center for Health Statistics.

Nicholson, S. (2003). Barriers to entering medical specialties. *NBER Working Paper #9649*. Cambridge, MA: National Bureau of Economic Research.

Wong, H. S. (1996). Market structure and the role of consumer information in the physician services industry: An empirical test. *Journal of Health Economics* **15**, 139–160.

Further Reading

Blevins, S. A. (1995). The medical monopoly: Protecting consumers or limiting competition? Policy Analysis No. 246. Washington, DC: Cato Institute.

Federal Trade Commission and Department of Justice (2004). Improving healthcare: A dose of competition. *A report by the Federal Trade Commission and The Department of Justice, 2004*. Available at: http://www.ftc.gov/reports/healthcare/040723healthcarerpt.pdf (accessed 20.08.13).

Gaynor, M., Haas-Wilson, D. and Vogt, W. (2000). Are invisible hands good hands? Moral hazard, competition and the second best in health care markets. *Journal of Political Economy* **108**, 992–1005.

Gaynor, M. and Town, R. (2011). Competition in health care supply. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *The handbook of health economics*, vol. 2, ch. 9. North Holland: Elsevier Science.

Gruber, J. and Owings, M. (1996). Physician financial incentives and caesarean section delivery. *RAND Journal of Economics* **27**, 99–123.

Yip, W. (1998). Physician responses to medical fee reductions: Changes in volume of supply of coronary artery bypass graft (CABG) surgeries in the medicare and private sector. *Journal of Health Economics* **7**, 675–700.

Relevant Website

http://www.oecd.org/health/
Organization for Economic Co-operation and Development.

Physician-Induced Demand

EM Johnson, Massachusetts Institute of Technology, Cambridge, MA, USA

Introduction

Physicians are often blamed for the high cost of healthcare in the US. Physicians dupe patients into consuming too much care, the story goes, driving up costs without producing commensurate gains in health.

This line of reasoning derives from the physician-induced demand (PID) hypothesis, which is a long-debated topic in health economics. Under the PID hypothesis, physicians influence patient demand to suit their own interests. They are able to do this because their patients know relatively little about the type or quantity of treatment they need. Faced with payment systems that reward quantity of care on the margin, the inducing physician provides care beyond the level that objective clinical judgment and patient preferences would dictate. In short, inducing physicians create their own demand rather than reacting to market demand.

The idea that doctors create their own demand is often used to make the case for healthcare reform, particularly changes to provider payment systems. Peter S. Orszag, the director of the Office of Management and Budget from 2009–10, was paraphrased by the New York Times as saying, 'the supply of hospitals, medical specialists, and high-tech equipment appears to generate its own demand' in June of 2009 (Pear, 2009). Induced demand is also a leading explanation for the geographic variation in utilization that has been documented across the US (Fisher *et al.*, 2003a, b). Atul Gawande has argued in the New Yorker magazine that induced demand combined with differences in the 'culture of money' across areas explains regional variation in Medicare costs per capita (Gawande, 2009). Although there has been no rigorous test of this relationship (as noted in Fuchs (2004)), policymakers have latched onto the idea that altering physician incentives in high cost areas can reduce costs without sacrificing quality of care. 'The Economic Case for Health Care Reform' of the White House states, "large variations in spending suggest that up to 30 percent of health care costs (or about 5 percent of GDP) could be saved without compromising health outcomes" (http://www.whitehouse.gov/administration/eop/cea/The Economic Case for Healthcare Reform).

This article reviews the empirical evidence on PID from the health economics literature. In the next section, induced demand is defined and the evidence on the topic reviewed. The following section brings evidence to bear from related literatures. Finally, the concluding section discusses policy implications and areas for future research.

Empirical Evidence

The concept of induced demand is first attributed to Evans (1974). The precise definition of McGuire (2000) follows:

> Physician-induced demand exists when the physician influences a patient's demand for care against the physician's interpretation of the best interests of the patient.

Under induced demand, a physician takes an action to shift the patient's demand curve in the direction of the physician's own interests. Physicians can affect such a shift, because they have more information regarding the patient's condition and treatment options than the patient, an example of the market failure known as asymmetric information. In theoretical models of induced demand, the action taken by the physician is unobserved and is limited at the margin by its costs. Typically inducement is itself costly for the physician, but in some models inducement negatively impacts patient flows (e.g., Pauly, 1980; Rochaix, 1989) or physician reputation (e.g., Dranove, 1988).

Two aspects of this definition merit clarification. First, induced demand does not include actions that influence demand in the best interest of the patient. Indeed, moving demand toward the patient's optimum is a responsibility of physicians. Second, the definition leaves room for treatment to vary across patients and providers. The benefits and risks of treatments vary with patient characteristics, and it is the job of the physician to tailor care to individual patients. Moreover, differences in physician practice styles, practice environments, and experience mean the true costs and benefits of treatments vary across physicians.

In the researcher's ideal world, the quantity of care the physician views as being optimal from the patient's perspective would be observable. The econometrician could then compare actual treatment with this benchmark, taking any difference as evidence of inducement. By comparing inducement across incentive environments, she could then estimate the physician's objective function. However, many of the characteristics of patients and doctors that determine appropriate treatment are unobserved. For this reason, empirical work on induced demand has used alternative identification strategies. This review groups papers according to the empirical approach and reviews each group sequentially. First are studies that use shocks to physician incomes, and especially physician-to-population ratios, to test for induced demand. Next are studies that use changes in physician fees or variation in patient information to identify inducement.

Before turning to empirical results, it is helpful to briefly clarify the predictions of PID models and compare them with alternative models of physician behavior (see McGuire (2000), for more detail). Under PID quantity is determined in equilibrium by physicians equating the marginal cost of induction with its marginal benefit. Physician incomes, fees, and patient information are all predicted to affect the quantity of care. Physician incomes affect the quantity of care through the income effect. A negative (positive) income shock increases (decreases) the marginal utility of income and increases (decreases) the returns to induction. If quantity of care is reimbursed at the margin, physicians then respond by increasing the quantity of care. An important caveat is that this prediction applies only when there is an income effect; otherwise inducing doctors induce equally at different levels of income (McGuire, 2000).

Changes in physician fees also affect quantities under PID. As in the case of an income shock, a fee reduction increases care quantities through the income effect. In addition, if the fee reduction differentially affects one area of the physician's practice (e.g., certain treatments or patients) relative to others, then there is a substitution effect (McGuire and Pauly, 1991). The relative returns to inducing decrease in the more affected area, and quantities are shifted to less affected areas. Thus, in response to own-fee reductions quantity can either increase or decrease, depending on the relative strength of the income and substitution effects. The prediction for less affected areas is unambiguous: quantity will increase. A final prediction that has been tested empirically is that physicians should have less ability to induce demand among more informed patients, where the asymmetry of information is lessened. This arises as long as the costs of inducement to informed patients are higher.

All of these predictions are in contrast with the perfect agent benchmark, in which only patient preferences and clinical factors matter for treatment, but it is more interesting to contrast the predictions of PID with models of physicians under symmetric information. When patients are informed, profit-maximizing physicians cannot shift demand, but they can affect healthcare quantities by making take-it-or-leave-it offers of nonretradeable services or by altering their choice of quality or effort (McGuire, 2000). This means observing quantities that depart from the patient's optimum is not sufficient evidence of PID. More relevant for evaluating the existing empirical work on PID, observing substitution in response to a fee change is not informative on PID, as physicians with informed patients can also be expected to shift quantities in this manner (McGuire, 2000).

However, profit-maximizing suppliers differ from inducers in that they do not adjust quantity in response to income shocks. This explains the focus of the empirical literature on using shocks to physician incomes and large fee changes to test for PID, though this approach amounts to jointly testing PID and the income effect. Because profit-maximizing physicians have no reason to treat informed and uninformed patients differently, studies using variation in patient information are also informative as to the underlying model of physician behavior. With this in mind, for some policy questions, one only needs to understand the reduced form relationship between physician incentives and utilization, so evidence in this category is considered as well.

Income Shocks

Many empirical studies of induced demand use variation in physician incomes to test for inducing behavior. The earliest studies of induced demand fall into this category. For the most part, these studies examined the relationship between market-level physician-to-population ratios and utilization. This approach is rooted in the idea that an exogenous increase in the number of physicians in a local practice area should spread patients more thinly, lowering physician incomes. Healthcare utilization is then increased if inducing physicians respond through the income effect and treat patients more intensively. Termed the 'availability effect' by Pauly (1980), these studies come the closest to directly testing the proposition that healthcare supply creates its own demand.

The first paper in this vein, Fuchs (1978), runs cross-sectional two-stage least squares (TSLS) regressions of surgeries on the number of surgeons per capita at the market level. To identify supply shifts, Fuchs instruments for the number of surgeons using characteristics of metropolitan areas that should affect surgeons' location decisions, but not local demand (e.g., metropolitan status, hotel receipts, and percent white). He finds a 10% increase in the surgeon-to-population ratio increases surgery by 3%, which he interprets as evidence of induced demand. Cromwell and Mitchell (1986) use a similar methodology with more data and finer geographic markets and find a 1.3% increase in elective surgery. Rossiter and Willensky (1983, 1984) relate healthcare utilization to physicians-per-capita using physician-level data and find even smaller effects.

These studies were highly influential, but there is concern that the instruments employed to isolate supply shocks do not satisfy the exclusion restriction. For example, Gruber and Owings (1996) suggest that results are biased toward inducement, because the average coinsurance rate, which is unobserved, is likely correlated with demand and with the included measures of attractiveness of an area to physicians. An additional concern is that supply shocks may reduce the price or time cost of services, causing patients to move down a static demand curve. Omitting these factors would bias the results toward finding inducement. Dranove and Wehner (1994) provide a powerful critique of the empirical methodology. Employing a method similar to Cromwell and Mitchell (1986), they show that increasing the number of obstetricians increases utilization on a dimension clearly out of the physician's control: the number of births.

Gruber and Owings (1996) avoid many of these problems and provide some of the best evidence to date on PID. In this paper, the authors instrument for state-level changes to the physician-to-population ratio, using the secular decline in the fertility rate from 1970 to 1982. They then look for evidence that physicians respond to the income shock by increasingly performing highly reimbursed Cesarean sections (C-sections) in lieu of less profitable vaginal deliveries, and they find a modest effect: obstetricians replaced approximately 10% of their income by increasing C-sections. By studying a plausibly exogenous shock to income, this approach is not subject to the criticisms of the previous literature. There are also fewer concerns about changes in time–cost in this context. However, it is difficult to compare the size of the estimated effect with previous studies, as obstetricians may have also recovered income on other margins.

Changes in Physician Fees

There is also a large empirical literature that uses changes in physician fees to identify inducement. The main advantage of this approach is the availability of large, exogenous fee changes for study, and most studies have used Medicare fee changes. Medicare fee changes are also appealing for testing PID because Medicare patients make up a significant fraction of physicians' practices. This is important because only fee changes that affect physician incomes have differential

predictions for utilization under PID and models with symmetric information (McGuire and Pauly, 1991).

Rice (1983) studies a large fee change enacted by Medicare in Colorado in 1976. Consistent with an income effect, Rice finds increases in Medicare volume in the Denver metropolitan area, where fees are lowered, relative to the surrounding areas where fees are raised. Point estimates suggest that a 10% decline in reimbursement led to a 6.1% increase in medical services and a 2.7% increase in surgery. However, it is possible that patient demand was affected by changes in patient responsibility over the time period. The short panel also prevents the author from assessing whether urban and rural areas were affected by differential trends over the period of study.

Nguyen and Derrick (1997) also study the impact of a Medicare fee change on Medicare volumes. They study the 1990 Medicare fee change, legislated in the Omnibus Budget Reconciliation Act of 1989, which reduced reimbursement for procedures deemed to be 'overpriced.' Using physician-level data, the authors find physicians experiencing fee reductions increase Medicare volumes. The volume response is similar in magnitude to Rice (1983), but it is significant only for the 20% of physicians who experienced the largest price reductions. While results are again consistent with PID, the study suffers from limitations similar to Rice (1983).

Yip (1998) also studies the 1990 reform, additionally considering the effect of Medicare fee changes on non-Medicare volumes. The paper focuses on thoracic surgeons, whose reimbursement rates were significantly reduced by the 1990 Medicare fee change. In this context, fee cuts led to increased volumes to both Medicare and private payers, with providers recouping, on average, 70% of income lost due to price reductions. The paper also convincingly demonstrates that the income effect is driving results by showing that physicians whose incomes were hit hardest by the reform have the largest volume responses. Jacobson *et al.* (2011) exploit a more recent Medicare fee change. The authors study the 2005 change in Medicare's reimbursement of outpatient chemotherapy drugs, and they find that physicians responded to reduced fees by increasingly administering chemotherapy. They also show that physicians substituted toward drugs that were lesss affected by the fee reduction in their prescribing behavior.

Gruber *et al.* (1999) is another important paper on fee changes and quantities. This paper studies Medicaid fee changes, specifically changes in Medicaid's reimbursement for C-sections. The authors expect the policy to have only a small income effect since Medicaid patients are a small fraction of providers' practices, and in fact, they find that the substitution effect dominates in this context: a 10% increase in the Medicaid fee led to an 8.4% rise in C-sections in the Medicaid population. While this result is consistent with models of PID, it is also consistent with physicians setting quantity under symmetric information.

Variation in Patient Information

There is also a long-established literature that uses variation in patient information to test for induced demand. These studies are motivated by the idea that informed patients should resist doctors' attempts at moving them away from their optimum consumption level. When inducing physicians are reimbursed for treatment on the margin, one expects utilization to be lower among more informed groups. The first study in this vein, Bunker and Brown (1974), compares rates of surgery for lawyers, businessmen, and ministers with those of physicians, who they view as informed consumers of medical services. Contrary to the prediction, they find self-reported surgery rates to be equal or higher among physician families when compared with other professional families in the same county. They conclude that physicians must have unobservably higher demand for medical services.

The conclusion of Bunker and Brown (1974) highlights the main weakness of the approach. When comparing utilization across patient groups, any omitted factors that are correlated with utilization will bias results. For example, prices, care quality, and health status may all differ across the professionals considered in this study. Hay and Leahy (1982) adopt the same approach using survey data with more extensive controls, including income, insurance coverage, and self-reported health status, but they also find higher use among physicians.

Domenighetti *et al.* (1993), in a more recent survey in Switzerland, find that the average person's probability of receiving one of seven major surgical interventions is one-third above that of a physician or a member of a physician's family. Ubel *et al.* (2011) survey physicians and find they want less intensive treatment for themselves than they would recommend to patients in two fatal disease scenarios. Again, results are difficult to interpret as patient characteristics influencing demand may differ across groups. Currie *et al.* (2010) address this weakness by conducting a patient audit study, which allows them to ensure comparability across informed and uninformed groups. Fake patients visited physician offices in China, where physicians have a financial incentive to prescribe medication. They then compare prescription rates of patients who verbally signal their understanding of appropriate prescription behavior with those who do not. It is found that prescription rates for the uninformed patient are higher by 25%. However, the physician could also have interpreted the information signal as a signal of patient preferences.

Related Literature

In this section, evidence on PID from related empirical literatures is considered. First, the empirical literatures on medical malpractice and defensive medicine are reviewed briefly, and results from the growing literature on physician incentives in managed care are summarized. The literature on physician self-referral, which considers whether physicians respond to the private incentive to use resources they partially own more intensively, is also reviewed. Finally, the literatures on pay-for-performance programs and studies of physician convenience factors are discussed.

First consider the literature on medical malpractice. So far, it has been assumed that the incentive physicians respond to in inducing demand is financial, but physician response to private liability risk is also consistent with the definition of

induced demand. Kessler and McClellan (1996, 2002) show that tort reform reduces medical expenditures on Medicare heart patients without affecting patient outcomes. They interpret this as evidence that doctors practice 'defensive medicine,' providing care that does not benefit patients in order to reduce their liability risk. More recently, Currie and MacLeod (2008) show that malpractice pressures increase the utilization of procedures that reduce liability risk, such as diagnostic testing, but decrease the use of risky treatments, such as the performance of C-sections in delivery. More research is needed to explore the relationship between financial and malpractice incentive systems.

The discussion has so far presupposed that it is financially rewarding for physicians to provide more healthcare to patients. Although this is true in fee-for-service payment systems, physicians paid by capitation have incentives to provide less treatment (Ellis and McGuire, 1986; McGuire, 2000). In fact, researchers have shown that physicians paid by capitation spend less time with patients (Mechanic *et al.*, 2001; Tai Seale *et al.*, 2007; Glied and Zivin, 2002; Melichar, 2005) and provide less care to each patient (Epstein *et al.*, 1986; Safran *et al.*, 2002; Stearns *et al.*, 1992; Greenfield *et al.*, 1992). Salaried doctors and doctors with bonuses tied to utilization measures also appear to respond to incentives for providing less care (see e.g., Hickson *et al.* (1997), Barro and Beaulieu (2003), Gaynor *et al.* (2004), and Hemenway *et al.* (1990)). These results cannot be interpreted as conclusive evidence of PID, however, as symmetric information models also predict this behavior. Further complicating interpretation is the fact that even perfect agents may reduce care if resources are rationed under managed care.

There is also a large empirical literature on self-referral practices by physicians. This literature studies treatment decisions when physicians have an ownership stake in some part of their practice. Reimbursement is typically higher when resources are owned by the physician, and studies find that physicians respond to this incentive by increasingly recommending patients for treatment. Mitchell (1992) and Hillman *et al.* (1992) study ownership incentives and referrals to diagnostic testing facilities, Yee (2011) studies ambulatory surgery centers, Barro *et al.* (2005) study specialty hospitals, and Baker (2010) studies the utilization of imaging devices. Iizuka (2012) also contributes evidence by showing that physician prescription behavior responds to pharmaceutical markups in Japan, where physicians dispense as well as prescribe drugs. Afendulis and Kessler (2007) studies a related conflict of interest. They observe that integrated cardiologists, who can both diagnose and perform interventional procedures to treat heart disease, have stronger incentives to recommend patients for intervention compared with non-integrated cardiologists, who must refer patients for treatment. They find that patients of integrated cardiologists are, in fact, more likely to receive percutaneous interventions.

Finally, the recent literature on pay-for-performance programs, which tie physician reimbursement to observable quality measures, suggests that performance incentives can affect care (Campbell *et al.*, 2007; Rosenthal and Frank, 2006; Mullen *et al.*, 2010). Studies of labor and delivery also suggest that obstetricians sometimes perform C-sections for their own convenience (Burns *et al.*, 1995; Spetz *et al.*, 2001). These results are inconsistent with the perfect-agent model. However, symmetric information models also predict substitution toward more highly reimbursed and away from more costly treatments. Therefore, it is again difficult to disentangle distortions due to financial incentives from those due to principal-agent concerns.

Suggestions for Future Research and Policy Implications

There is a large and growing body of empirical evidence that physicians' treatment decisions are influenced by factors beyond their patients' needs. Convincingly identified studies have shown that obstetricians do more C-sections in response to declining fertility, cardiac surgeons treat more intensively when their incomes are impacted by fee reductions, and physicians in China prescribe more medication to 'uninformed' patients. This evidence is inconsistent with the model of physicians as perfect agents, and it supports PID as one avenue through which physicians affect quantities of healthcare. In addition to this direct evidence of PID, physicians respond to private malpractice incentives and financial incentives for self-referral. Physicians also appear to respond to the incentives in managed care plans. Finally, there is some evidence that pay-for-performance programs and physician convenience factors affect healthcare choices. Taken together, these studies suggest that physician incentives, broadly defined, are important determinants of both healthcare costs and the distribution of health resources in the United States.

However, more work is needed before one can make statements about the economic importance of PID. Although empirical research has provided estimates in several contexts, there is reason to believe that the effect will differ across incentive environments, physician specialties, patient groups, and even across treatment categories within physician–patient pairs. Future research exploring this heterogeneity should also aim to bridge our current understanding of PID with claims made in the health policy arena. How much of the variation in utilization across geographic areas can be explained by demand inducement? Have physician incentives or constraints on inducement changed, such that PID has contributed to growth in health spending over time?

More theoretical work is also needed. Exploring the impact of competition on physician behavior is a promising area for research (and one that may produce new testable implications for PID), though this first requires refining our understanding of the sources of physician market power. It would also be interesting to theoretically explore the interplay of the various incentive systems that physicians face, for example, by studying physicians who are contracted with both HMOs and PPOs.

The PID research agenda is important for policy. The general direction of health policy in the US and other countries is to push some financial risk to physician groups, as accountable care organizations (ACOs) do in Medicare. If PID is pervasive and powerful, creating an interest among physicians in providing less care may work well to reduce healthcare costs, but not without raising concerns about access and quality. The extent of induced demand also has implications for physician workforce and training policies; and gaining

clarity into induced demand behavior has implications for health insurance design – inducement affects the interpretation of parameters that are central in the optimal insurance literature. Finally, the literature on PID can help us to understand the impacts of patient empowerment policies, for example, patient ownership of their own medical information.

See also: Demand for Insurance That Nudges Demand. Managed Care. Medical Decision Making and Demand. Medical Malpractice, Defensive Medicine, and Physician Supply. Organizational Economics and Physician Practices. Physician Management of Demand at the Point of Care. Rationing of Demand

References

Afendulis, C. and Kessler, D. P. (2007). Tradeoffs from integrating diagnosis and treatment in markets for healthcare. *American Economic Review* **97**, 1013–1020.

Baker, L. (2010). Acquisition of MRI equipment by doctors drives up imaging use and spending. *Health Affairs* **29**, 2252–2259.

Barro, J. and Beaulieu, N. (2003). *Selection and improvement: Physician responses to financial incentives*. Working Paper 10017. Cambridge, MA: National Bureau of Economic Research.

Barro, J. R., Huckman, R. S. and Kessler, D. P. (2005). The effects of cardiac specialty hospitals on the cost and quality of medical care. *Journal of Health Economics* **25**, 702–721.

Bunker, J. P. and Brown, Jr., B. W. (1974). The physician–patient as an informed consumer of surgical services. *New England Journal of Medicine* **290**, 1051–1055.

Burns, L. R., Geller, S. E. and Wholey, D. R. (1995). The effect of physician factors on the cesarean section decision. *Medical Care* **33**(4), 365–382.

Campbell, S., Reeves, D., Kontopantelis, E., et al. (2007). Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine* **357**, 181–190.

Cromwell, J. and Mitchell, J. B. (1986). Physician-induced demand for surgery. *Journal of Health Economics* **5**, 293–313.

Currie, J. and MacLeod, B. (2008). First do no harm? Tort reform and birth outcomes. *Quarterly Journal of Economics* **123**, 795–830.

Currie, J., Lin, W. and Zhang, W. (2010). Patient knowledge and antibiotic abuse: Evidence from an audit study in China. *Journal of Health Economics* **30**, 933–949.

Domenighetti, G., Casabianca, A., Gutzwiller, G. and Martinoli, S. (1993). Revisiting the most informed consumer of surgical services. *International Journal of Technology Assessment in Health Care* **9**, 505–513.

Dranove, D. (1988). Demand inducement and the physician-patient relationship. *Economic Inquiry* **26**, 251–298.

Dranove, D. and Wehner, P. (1994). Physician-induced demand for childbirths. *Journal of Health Economics* **13**, 61–73.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement. *Journal of Health Economics* **5**, 129–151.

Epstein, A. (1986). The use of ambulatory testing in prepaid and fee-for-service group practices: Relation to perceived profitability. *New England Jounal of Medicine* **314**, 1089–1093.

Evans, R. (1974). Supplier-induced demand: Some empirical evidence and implications. In Perlman, M. (ed.) *The economics of health and medical care*, pp. 162–173. London: Macmillan.

Fisher, E. S., Wennberg, D. E., Stuken, T. A., et al. (2003a). The implications of regional variations in Medicare spending. Part I: The content, quality and accessibility of care. *Annals of Internal Medicine* **138**(4), 273–287.

Fisher, E. S., Wennberg, D. E., Stuken, T. A., et al. (2003b). The implications of regional variations in Medicare spending. Part II: Health outcomes and satisfaction with care. *Annals of Internal Medicine* **138**(4), 288–298.

Fuchs, V. R. (1978). The supply of surgeons and the demand for operations. *The Journal of Human Resources* **13**, 35–36.

Fuchs, V. R. (2004). Reflections on the socio-economic correlates of health. *The Journal of Health Economics* **23**, 653–661.

Gawande, A. (2009). The cost conundrum. *The New Yorker*, June 1.

Gaynor, M., Rebitzer, J. B. and Taylor, L. J. (2004). Incentives in HMOs. *Journal of Political Economy* **112**, 915–931.

Glied, S. and Zivin, J. (2002). How do doctors behave when some (but not all) of their patients are in managed care? *Journal of Health Economics* **21**, 337–353.

Greenfield, S., Nelson, E. C., Zubkoff, M., et al. (1992). Variations in resource utilization among medical specialties and systems of care: Results of the medical outcome study. *Journal of the American Medical Association* **267**, 1624–1630.

Gruber, J., Kim, J. and Mayzlin, D. (1999). Physician fees and procedure intensity: The case of cesarean delivery. *Journal of Health Economics* **18**, 473–490.

Gruber, J. and Owings, M. (1996). Physician financial incentives and Cesarean section delivery. *RAND Journal of Economics* **27**, 99–123.

Hay, J. and Leahy, M. J. (1982). Physician-induced demand. *Journal of Health Economics* **2**, 231–244.

Hemenway, D., Killen, A., Cashman, B., et al. (1990). Physician response to financial incentives: Evidence from a for-profit ambulatory care center. *The New England Journal of Medicine* **322**, 1059–1063.

Hickson, G. B., Altemeier, W. A. and Perrin, J. M. (1997). Physician reimbursement by salary or fee-for-service: Effect on a physician's practice behavior in a randomized prospective study. *Pediatrics* **80**, 744–750.

Hillman, A., Olson, G. and Griffith, P. (1992). Physicians' utilization and charges for outpatient diagnostic imaging in a Medicare population. *Journal of the American Medical Association* **268**, 2050–2054.

Iizuka, T. (2012). Physician agency and adoption of generic pharmaceuticals. *American Economic Review* **102**, 2826–2858.

Kessler, D. and McClellan, M. (1996). Do doctors practice defensive medicine? *Quarterly Journal of Economics* **111**, 353–390.

Kessler, D. and McClellan, M. (2002). How liability law affects medical productivity. *Journal of Health Economics* **21**, 931–955.

McGuire, T. (2000). Physician agency. In Culyer, A. J. and Newhouse, J. P. (eds.) *The handbook of health economics*, vol. 1, pp. 462–536. Amsterdam: Elsevier.

McGuire, T. and Pauly, M. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics* **10**, 385–410.

Mechanic, D., McAlpine, D. and Rosenthal, M. (2001). Are patients' office visits with physicians getting shorter? *New England Journal of Medicine* **344**, 198–204.

Melichar, L. (2005). The effect of reimbursement on medical decision making: Do physicians alter treatment in response to a managed care incentive? *Journal of Health Economics* **28**, 902–907.

Mitchell, J. M. (1992). Physician ownership of physical therapy services. Effects on charges, utilization, profits and service characteristics. *Journal of the American Medical Association* **268**, 2055–2059.

Mullen, K., Frank, R. and Rosenthal, M. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *RAND Journal of Economics* **41**, 64–91.

Nguyen, N. X. and Derrick, F. W. (1997). Physician behavioral response to a Medicare price reduction. *Health Services Research* **32**, 283–298.

Pauly, M. (1980). *Doctors and their workshops: Economic models of physician behavior.* Chicago: University of Chicago Press.

Pear, R. (2009). Health care spending disparities stir a fight. *The New York Times*. Available at: www.nytimes.com/2009/06/09/us/politics/09health.html (accessed 18.04.12).

Rice, T. (1983). The impact of changing Medicare reimbursement rates on physician-induced demand. *Medical Care* **21**, 803–815.

Rochaix, L. (1989). Information asymmetry and search in the market for physician services. *Journal of Health Economics* **8**, 53–84.

Rosenthal, M. and Frank, R. (2006). What is the empirical basis for paying for quality in health care? *Medical Care Research and Review* **63**, 135–157.

Rossiter, L. F. and Wilensky, G. R. (1983). A reexamination of the use of physician services: The role of physician-initiated demand. *Inquiry* **20**, 162–172.

Rossiter, L. F. and Wilensky, G. R. (1984). Identification of physician-induced demand. *Journal of Human Resources* **19**, 231–244.

Safran, D. G., Wilson, I. B., Rogers, W. H., Montgomery, J. E. and Chang, H. (2002). Primary care quality in the Medicare program: Comparing the performance of Medicare health maintenance organizations and traditional fee-for-service Medicare. *Archives of Internal Medicine* **162**, 757–765.

Spetz, J., Smith, M. W. and Ennis, S. F. (2001). Physician incentives and the timing of Cesarean sections: Evidence from California. *Medical Care* **39**, 536–550.

Stearns, S., Wolfe, B. and Kindig, D. (1992). Physician responses to fee-for-service and capitation payment. *Inquiry* **29**, 416–425.

Tai Seale, M., McGuire, T. and Zhang, W. (2007). Time allocation in primary care office visits. *Health Services Research,* **42**, 1871–1894.

Ubel, P., Andrea, M. and Brian, Z. (2011). Physicians recommend different treatments for patients than they would choose for themselves. *Archives of Internal Medicine* **171**, 630–634.

Yee, C. A. (2011). Physicians on board: An examination of physician financial interests in ASCs using longitudinal data. *Journal of Health Economics* **30**, 904–918.

Yip, W. (1998). Physician responses to medical fee reductions: Changes in the volume and intensity of supply of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors. *Journal of Health Economics* **17**, 675–700.

Further Reading

Feldman, R. and Sloan, F. (1998). Competition among physicians, revisited. *Journal of Health Politics, Policy and Law* **13**, 239–261.

Jacobson, M., Earle, C., Price, M. and Newhouse, J. (2010). How Medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment. *Health Affairs* **29**, 1391–1399.

Mitchell, J. M. (2005), Effects of physician-owned limited service hospitals: Evidence from Arizona, *Health Affairs*, Supplementary Web Exclusives W5:481-490.

Mitchell, J. M. (2010). Effect of physician ownership of specialty hospitals and ambulatory surgery centers on frequency of use of outpatient orthopedic surgery. *Archives of Surgery* **145**, 732–738.

Phillips, K., Fernyak, S., Potosky, A., Schauffler, H. and Egorin, M. (2000). Use of preventive services by managed care enrollees: An updated perspective. *Health Affairs* **19**, 102–116.

Physicians' Simultaneous Practice in the Public and Private Sectors

P González, Universidad Pablo de Olavide, Sevilla, Spain

Glossary

Access The degree to which individuals are inhibited or facilitated in their ability to receive care and services from the healthcare system. Factors influencing this ability include geographic location, architectural features, availability of transport and financial considerations.

Asymmetry of information A situation in which the parties to a transaction have different amounts or kinds of information as when, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances.

Cost sharing An arrangement whereby the cost of healthcare is shared, typically between the patient and insurer.

Cream skimming A form of selection in private health insurance markets by which the insurer obtains a higher proportion of good risks (people with a low probability of needing care or who are likely to need only low-cost care – or both) in their portfolio of clients than is assumed in the calculation of the insurance premiums. Also called 'cherry-picking' and 'creaming'.

Dual practice A combination of public and private practice by doctors, sometimes even within the same hospital.

Efficiency A resource allocation is efficient if it is not possible to reallocate resources so as to increase one person's utility (or health or output) without decreasing another person's utility (or health or output). In health economics the entity maximized is generally assumed to be utility, health, or welfare.

Incentives Any factor (financial or nonfinancial) that influences providers' or consumers' behavior. In the health market, it often refers to financial or psychological rewards designed to motivate physicians to perform at or better than an established standard.

Physician-induced demand The effect that providers of services may have in creating more patient demand than there would be if they were to act entirely in the interests of their patients. Also called supplier-induced demand.

Prospective payment A method of reimbursing health service providers (especially hospitals) by establishing rates of payment in advance, which are paid regardless of the costs in actual individual cases.

Quality of healthcare It can refer variously to clinical process, hotel services, health outcomes, frequency of adverse events, or conformity with clinical guidelines and other authoritative standards of care.

Rationing Allocating resources according to their prices, a planning procedure, a set of rules, or other administrative arrangements.

Introduction/Background

Dual practice among doctors, which the literature often refers to as dual (or multiple) job-holding among health workers, is a phenomenon that can be observed in most countries with both public and private health care systems.

The term 'multiple job-holding' is defined as working simultaneously in more than one paid job. Any worker can hold several jobs at one time, and health professionals are no exception. Multiple job-holding in the health sector can take various forms. Doctors may combine medical practice with other related activities, such as research or teaching, for example, or with a totally unrelated activity, as in many developing countries. They may also hold more than one job within the public (or private) sector, or in both sectors simultaneously. This article focuses on the last option, that is, dual practice among doctors whose main job is in the public sector but who also do clinical work in the private sector.

Despite the prevalence of multiple job-holding among health workers, the phenomenon is largely undocumented. It is nevertheless allowed in most countries with mixed health systems, with two main exceptions, Canada and China, where it is forbidden by law. As a result, dual practice among doctors is widespread in European countries, prominent examples being Ireland, where 90% of the doctors employed in state hospitals also work in the private sector; and the UK, where 60% of state-employed doctors do so. Data also exist for countries outside Europe, for example, Australia and New Zealand, wherein according to the Royal Australasian College of Physicians, 79% and 43% of public sector doctors, respectively, hold some jobs in the private sector. In less developed countries, the exclusively state-employed doctor is a fading figure, due to low public-sector salaries. Dual practice is therefore widespread in Asian countries (Thailand, Vietnam, India, among others), Africa (Egypt, Zambia, Mozambique, etc.), Latin America (in Peru, almost 100% of doctors are dual practitioners), and also in Eastern Europe, wherein despite a rapidly growing private health market, doctors are reluctant to give up their public sector jobs entirely.

Just like many others, doctors may have various reasons for holding more than one job. Possibly, the most powerful being economic motivation. The intuition is that, what with overtime restrictions in the main job, people will be willing to earn extra income by taking on a second job, if it is sufficiently well paid. A recent empirical study for Norway by Godager and Luras, provides the evidence that, after a decrease in income due to shortage of patients, general practitioners have reacted by increasing the hours devoted to community health service.

Other reasons leading doctors to hold more than one job include: (1) complementarities between jobs, either in terms of income (public sector stability combined with private sector earnings that are higher on average, albeit more variable) or nonmonetary advantages (opportunities to expand professional contacts or obtain recognition and prestige within the profession), and also technological and training complementarities (a second job can enable doctors to widen their experience and/or learn new techniques); (2) professional and institutional factors relating to workload and work satisfaction, or ineffective organization, structural deficiencies, and unsatisfactory working conditions in the public health sector; and (3) personal characteristics (age, gender, household structure, etc.).

The economic theory underlying dual practice is still scant and relatively recent. Few theoretical models have been developed to analyze the issue (see Table 1 for a summary of the theoretical literature), and there is a notable lack of empirical studies on the subject.

Dual Practice: A Context for Discussion

Despite the prevalence of dual practice among doctors in the majority of mixed health systems, there is a surprising lack of evidence regarding the potential impact on the efficiency of health care resource management. Although allowing health professionals to hold more than one job may have some positive consequences, it may also give rise to a degree of opportunistic behavior that could compromise the efficiency and quality of public health provision. This is why dual practice among doctors is subject to social controversy. Detailed below are some of the arguments for and against dual practice, together with their theoretical underpinnings.

Costs of Dual Practice

It is generally acknowledged that dual practice may have three potentially negative consequences for public health provision: (1) poorer care quality, (2) longer waiting times and waiting lists, and (3) higher costs. Although these three problems are directly interlinked (e.g., longer waiting times can lead to poor quality of care), in the remaining part of this article they are presented as independent issues, along with the related, but limited, scientific literature.

The problem of the deterioration of the quality of care in public health services due to dual practice on the part of doctors is directly linked to incentives. Dual practice doctors may have incentives to dedicate as much time and effort as possible to their private work, and therefore fail to complete their hours effectively in their public jobs either by making less effort or by taking on a lower work load. Unjustified absenteeism and shirking by health professionals are quite common in many developing countries, although there is evidence, albeit anecdotic, that these are not unknown phenomena in western economies. However, the empirical studies that report this behavior are unable to separate the effect of dual practice from that of poor organization and general lack of motivation toward effort.

Potential shirking of professional responsibilities has been widely studied in the economic literature. Inherent in these research models is the notion of asymmetric information, relating to the fact that employers cannot observe the efforts of their workers (or the result of that effort) and supervision is costly. As contracts cannot be contingent on effort, doctors may have other incentives attributing to their less than the socially desirable level of effort into diagnosing and treating patients, with negative repercussions for health care quality.

When doctors are dual providers, the reduction in their care quality may result not from lack of dedication to the public job, but from their strategic undermining of patients' perception of public health care (a more subtle kind of physician-induced demand) so that more patients opt for private treatment. The motives for this are mainly economic, driven by the fact that the public sector pays a fixed salary, whereas in the private sector incentive schemes are more common. In a paper by Brekke and Sørgard (2007), dual practice leads doctors to skimp on their effort in the public sector in order to process fewer public sector patients and thus increase the demand for private treatment. In Biglaiser and Ma (2007), it is shown that only a fraction of doctors, who are classified as 'moonlighters', provide minimal service quality in the public sector and have incentives to refer patients out of the public system. Although the quality of public health care could still diminish, the authors show that this need not be the case, given that health authorities can utilize the savings in public health costs resulting from the diversion of patients to the private sector to improve the quality of service from 'dedicated' doctors working exclusively in the public health sector. Finally, Delfgaauw (2007) suggests that average public health care quality (interpreted in their model as the probability of receiving high quality care from an altruistic doctor) suffers because of dual practice, thus penalizing the poor who cannot afford private treatment.

A large part of the literature focuses on the design of contracts incorporating incentives to encourage desirable behavior. It is generally agreed that, for doctors working exclusively in one sector, to strike the optimum balance between cost and performance, it is necessary to combine a purely prospective payment system with partial cost reimbursement (i.e., adopting a mixed payment system). If doctors are dual providers, the effectiveness of incentives to discourage opportunistic behavior on the part of doctors does not depend entirely on the type of payment system used in the public sector. It also depends on how they would be paid in a potential second job. It has been suggested that a mixed payment system is also the best alternative for dual practitioners, although the cost reimbursement rule is likely to be more complicated. Depending on whether the relationship between public and private practice is one of substitution or complementarity, the optimal rule will yield a higher or lower return on costs than when doctors do not hold two jobs at the same time (Rickman and McGuire, 1999).

Another issue being closely linked to the deterioration of the quality of public health care is the fact that health care delivery by dual practitioners is sometimes associated with longer waiting times and longer waiting lists in the public sector. The origin of this problem may also be twofold. Firstly, it may be the result of incentives for doctors to shirk in the public sector and save their efforts for the private sector. Secondly, dual practitioners

Table 1 Overview of different theoretical models on physicians' dual practice

Authors	*Research question*	*Model description*	*Main results*	*Policy recommendations on dual practice*	*Reference*
Barros and Olivella (2005)	Cream skimming of patients when public treatment is rationed.	There are three groups of players: (1) a health authority which runs a public hospital and chooses a rationing policy; (2) a representative physician that treat patients in public sector, chooses the maximum level of severity he is willing to treat in his private practice, and offers those patients the possibility of resorting to a private treatment; (3) a set of patients that differ in their severity and choose whether to wait for free public treatment or pay for immediate private treatment.	Although doctors might have incentive to select highly profitable (low cost) patients for treatment in the private sector, full cream skimming is only compatible with intermediate rationing policies. If, on the contrary, rationing is either very lax or very stringent, then cream skimming is always partial.	None	*Journal of Economics and Management Strategy* **14**, 623–646.
Biglaiser and Ma (2007)	Welfare implications of dual practice when some public-service physicians may refer patients to their private practices.	There are three groups of players: (1) a regulator that decides physicians' payment in the public sector and whether dual practice is allowed or not; (2) two sets of doctors: dedicated (working only in the public sector) and moonlighters who choose a quality level to provide to their patients in the public system. Moonlighters decide also whether to refer their public patients to their private practices; (3) a set of patients that differ in income and who may be offered a private option by a moonlighter.	Allowing dual practice increases social welfare. But public care quality may decrease as a result of adverse behavioral reactions, such as moonlighters shirking more and dedicated doctors abandoning their sincere behavior.	Limiting private income through price ceilings limits adverse behaviors in the public system and improves consumer welfare.	*RAND Journal of Economics* **38** (Winter), 1113–1133.
Brekke and Sørgard (2007)	Interaction between public and private health provision in a National Health Service.	There are three different agents: (1) a health authority, which decides whether or not to allow private (out of plan) provision of health care alongside the National Health Service (NHS), and the public sector remuneration (wage level); (2) a set of physicians determining their labor supply in the public sector and, if allowed, in the private sector; (3) a set of individuals demanding medical treatment from the NHS and private (out of plan) providers, if this is an option.	Allowing physicians' dual practice 'crowds out' public provision and results in lower overall health care provision.	A ban on dual practice is desirable if private sector competition is weak and public and private cares are sufficiently close substitutes.	*Health Economics* **16**(6), 579–601.

(*Continued*)

Table 1 Continued

Authors	*Research question*	*Model description*	*Main results*	*Policy recommendations on dual practice*	*Reference*
Delfgaauw (2007)	Allocation of patients to physicians under different systems of health provision.	There are three agents: (1) the health authority, which enforces a minimum public treatment quality; (2) two groups of physicians (regular and altruistic), who are identical except for their attitude toward patients; (3) a population of patients who differ in income. In a purely public system, treatment is provided (free of charge) within the public sector, and patients and physicians are randomly matched. In a mixed system, there is also private provision of health care, and each physician chooses whether to work for the public or in the private sector, and each patient decides where to obtain treatment.	Allowing for private provision of health care, parallel to free treatment in a National Health Service, benefits all patients. But allowing dual practice is detrimental for the poorest patients.	None	Tinbergen Institute, Discussion Papers 07-010/1.
González (2004)	Design of contracts for dual providers who use their public practice as a source of reputation for their private activity.	There are three agents: (1) a health authority, which decides the physicians' payment contract aimed at providing incentives for an accurate diagnosis and treatment decision; (2) a representative patient who suffers from an illness whose severity is unclear and requires medical attention; (3) a representative physician in charge of diagnosing the illness and its severity, and deciding whether to treat or refer the patient to a specialist.	Physicians have incentives to overdo medical services when they use their public performance as a way of increasing their prestige as private doctors.	Exclusive contracts are the 'second best' choice. The desirability of limiting policies is ambiguous. On the one hand, they can help to mitigate the physicians' tendency to over-provide services. But, at the same time, they increase the cost for the health authority of inducing an accurate diagnosis on the part of the physician.	*Health Economics* **13**, 505–524.
González (2005)	Cream skimming of patients when health authorities reach agreements with private hospitals to have some of their public patients treated there.	There are three agents: (1) the health authority, which decides the payment to public physicians and the amount of patients that are referred to private hospitals to alleviate public sector congestion; (2) a set of patients differing in their degree of severity; (3) a representative physician, who selects the patients he wants to treat in the public sector. The remaining patients will be transferred to the private hospital.	Dual providers refer the most profitable (low cost) patients from public sector to private practice.	None	*Health Economics* **14**, 513–527.

González and Macho-Stadler (2013)	Optimal regulations on dual practice in developed and developing countries	There is a health authority that contracts physicians to provide public health and designs the regulatory regime regarding dual practice. Physicians have different skills and decide whether to work solely for the public sector, be dual practitioners, or work in the private sector exclusively. Private market rewards quality. The production technology of health differs for developed and developing countries, being more dependent on physicians' skills in the latter.	Banning dual practice increases the cost of retaining high skilled physicians in the public sector. Limiting dedication to private practice is more efficient than limiting dual practitioners' earnings. Whenever enforceable, a policy of limits is always more efficient than one of offering exclusive contracts.	In developing countries, the scope for regulation is generally limited due to the high health costs because of a brain drain of health professionals. In developed countries, when regulation is needed, limits to dedication emerge as the best policy.	*Journal of Health Economics* **32**(1), 66–87.
Iversen (1997)	Effect of a private sector on the waiting time associated with treatment in the public sector.	The model is a two-stage game. In a first stage, the hospital (physician) decides on hospital admissions and waiting time. Patients choose, then, between private treatment (costly) and public treatment (free but with rationing and waiting). In a second stage, the health authority decides the budget allocated to the hospital. When physicians are dual suppliers, part of their income coming from private practice is increasing in the public sector waiting time.	When waiting-lists admissions are rationed, the waiting time increases if public sector consultants are allowed to work in the private sector in their spare time.	None	*Journal of Health Economics* **16**, 381–396.
Rickman and McGuire (1999)	Optimal public sector reimbursement rule when physicians are dual providers.	A representative patient receives treatment from his/her physician in the public sector (e.g., the NHS) and the private sector. These treatments can be either substitutes or complements. Treatment in the public sector is free of charge but private treatment is costly. The physician derives utility from his/her private income and public sector work, and also from patient's welfare. The health authority designs the payment contract for the public sector in order to maximize social welfare.	Public sector cost-sharing remains socially efficient, but it is generally nonlinear. The precise details depend on whether public and private services are substitutes or complements and on the degree of social efficiency achieved in the private sector.	None	*Scottish Journal of Political Economy* **46**, 53–71.

may have strategic incentives for enabling public sector waiting times and waiting lists grow. Along these lines, Iversen (1997) shows that if admissions to public sector waiting lists are rationed, the waiting time for patients will increase when their doctors are allowed to work in the private sector in their free time. Doctors will be motivated by the desire to increase their private earnings, in this case by maintaining long waiting lists in the public sector so that more patients are willing to pay for private treatment.

Finally, dual practice by doctors can increase public sector direct costs in various ways. These include opportunistic behavior, such as the appropriation of material paid by the public sector or the use of public sector equipment to treat private patients. Although this kind of behavior is more frequent in developing countries, there is anecdotic evidence to show that it also takes place in western economies.

The cost of public health care may also increase if doctors have incentives to practice 'cream skimming.' This means selecting patients for treatment on the basis of either seriousness of their condition or their likelihood of recovery. In an environment of asymmetric information, cream skimming can often be the result of a prospective payment system, which encourages overtreatment of patients with minor ailments and undertreatment of serious cases. In a context with dual practitioners, cream-skimming may be explained as the result of doctors having incentives to refer less serious cases to their private practices, and leave more serious cases to the public sector. Barros and Olivella (2005) and González (2005) studied this phenomenon in a context of public sector waiting lists. The second of these studies shows that doctors holding more than one job have (purely economic) incentives to divert less serious (thus, less costly) cases to the private sector when the health authority decides to use private hospitals to alleviate congestion in the public sector, a policy that has recently been applied in many EU countries. Barros and Olivella (2005) stress that, when public treatment is rationed, the actual scope for cream skimming depends on the waiting list admission criteria used by the health authority. In fact, the scope for cream skimming is limited both by laxity, which results in a waiting list with a higher proportion of mild cases who will be less willing to pay for private treatment, and by extreme strictness, which reduces waiting times. Therefore, according to these authors, only an 'intermediate' policy will allow dual providers to divert less serious patients to their private practices, thus affecting public health costs.

Finally, it could be argued that dual practice is sometimes accompanied not by shirking but by precisely the opposite kind of behavior. Doctors may try to build their professional reputation and prestige through their work in the public sector and thus increase their private sector earnings. In this case, public health care would not suffer, but treatment costs might increase through doctors prescribing stronger (and more expensive) treatments (González, 2004).

Benefits of Dual Practice

Allowing doctors to work in the public and private sectors at the same time also carries some benefits for public health care provision.

In many countries, by allowing dual practice, governments are able to retain their best health professionals at less cost. If a high percentage of doctors abandon the public sector, care quality will be badly affected. This is potentially a serious problem in developing countries, where low salaries imply there is often a shortage of doctors in the public sector. An article by González and Macho-Stadler (2013) illustrates this possibility using a theoretical model where doctors differ in their levels of skill, which is understood as their capacity to deliver adequate health care. A ban on dual practice reduces the number of doctors working in the public sector. The impact this causes on the total amount of public health provision is particularly significant if the private market rewards quality and thus draws the more skilled doctors away from the public sector.

Also, allowing doctors to work in the private sector to supplement their public sector salaries can have the positive effect of reducing the prevalence of informal payments, which are so widespread in developing and transitional economies. Informal payments constitute an informal market for health care within the confines of public health care networks, and, compromises governmental efforts to improve efficiency and equity in the delivery of health services. It was to reduce this problem that Greece legalized dual practice by doctors in 2003.

When there are complementarities between the tasks performed in different jobs, working in different working environments can enrich the professional experience of dual practitioners. Sometimes, working in the private sector allows doctors to access new technologies and improve their knowledge and technical expertise. This benefits both their private and public sector patients. The latter could benefit not only directly, if the new techniques or technology are introduced in public hospitals, but also indirectly, through improvements in professional skill and experience.

Finally, the incentives of dual providers to direct their public sector patients into the private sector can be potentially beneficial in aggregate welfare terms. This is the case if the diversion involves higher income patients who are able to afford private treatment. Thus, the more altruistic kind of doctor may advise poor patients to seek free treatment in a public hospital, and refer only those who can afford it to their private practice. This is precisely the conclusion reached in the paper by Biglaiser and Ma (2007) mentioned earlier, in which allowing dual practice is found to result in welfare gains. The motivation is twofold. Firstly, there is a saving in public sector costs, where the numbers of patients being treated will decrease as some are diverted to the private sector. Secondly, efficiency will improve because patients in the private sector will receive medical treatment of quality for which they are willing to pay. This type of argument must be treated with some caution, however, because only if there are real differences in quality between public and private health provisions, will there be patients willing to pay for private treatment. This could mean that the rich and the poor would receive unequal medical treatment, that is, a clear two-tier system. There is also some empirical evidence to show that people from low income, poorly educated groups are often the most likely to respond to inducement, to use private services and pay for private treatment instead of using the subsidized public health service.

An Overview of Dual Practice Interventions

Whether physicians' dual practice should be regulated in some way, and how health authorities should best intervene are two difficult questions for which the literature has not yet found unanimous answers. As mentioned already, previous theoretical studies are rather ambiguous when it comes to identifying the ultimate impact of dual practice on the efficiency and quality of public health care provision. More importantly, there has been no solid empirical research to enable us to quantify the costs and benefits of allowing (or banning) dual practice.

The reality is that dual practice by doctors is regulated in a fair number of countries, although regulations differ enormously from one country to another. Canada and China are two examples of the very few countries where dual practice is prohibited. There is disagreement in the literature as to the effectiveness of prohibition. Brekke and Sørgard (2007) find that the banning of dual practice is an efficient policy in cases where the private health sector is not highly competitive, or patients perceive public and private health care as substitutes. In both these situations, the crowding-out effect of physicians' dual practice on public health care is so strong that dual practice is welfare-reducing and banning is desirable. González and Macho-Stadler (2013) nevertheless stress that a ban on dual practice does not appear to be a good strategy. In developed countries, professional ethics and self-regulation are well established and can reduce the incidence of unacceptable behavior. At the same time, dual practice enables governments to retain highly qualified professionals at a relatively low cost. In these countries, therefore, the gains from banning dual practice would not offset the costs. In less developed countries, although the institutional environment is weak and the work environment permissive, prohibition still appears undesirable. These countries often suffer from a heavy 'brain-drain' of doctors from the public sector in search of better job opportunities. If the prohibition of dual practice means that even fewer doctors are attracted to the public sector, the poorest patients, who cannot afford private treatment, will find quality health care frankly difficult to access. Finally, prohibition is not found to be a good strategy in Biglaiser and Ma (2007) because it cancels out the efficiency gains being generated by dual practitioners diverting those patients who are willing to the private sector where they can obtain better quality treatment in return for payment.

If prohibition is not the best solution, what other options are available? Governments worldwide have tackled this issue in a wide variety of ways. Some countries, such as Italy, Portugal, or Spain, among others, have taken the alternative of offering doctors some kind of premium (a salary bonus or promotion points) in exchange for voluntarily restricting their professional activity to the public sector. The limited theoretical literature on the subject suggests that this kind of policy is optimal only in certain circumstances: (1) when it is not possible to design incentive contracts to encourage desirable behavior for health professionals (González, 2005) and (2) when other regulations – particularly, dual practice restrictions – are difficult to implement (González and Macho-Stadler, 2013).

In developing countries, exclusive contracts are a still less desirable option. Leaving aside the financial constraints affecting these countries, exclusive contracts will inevitably attract more poorly skilled doctors with less chance of making money in the private sector. Poor countries sometimes lack sufficient clinical technology and clearly defined treatment protocols, and doctors often work in isolation. As a result, the quality of treatment in the public sector largely depends on the skill of the doctor. Thus, the efficiency gains from this policy will tend to be minimal (González and Macho-Stadler, 2013).

Another measure adopted by some countries is to place restrictions on dual practice. For instance, the UK has placed an upper limit on the amount of money dual practitioners are allowed to earn in the private sector. Biglaiser and Ma (2007) examined this option and show that, when the probability of doctors' shirking depends on their earning potential in the private sector, it is socially optimal to allow dual practice but cap doctors' private earnings, by means of price-ceilings. This enables the health authority to compensate for the loss in public sector health care quality due to dishonest doctors' shirking, with efficiency gains derived from dual practice, as mentioned above. González and Macho-Stadler (2013) show that it will always be more efficient to limit doctors' dedication to private practice rather than the earnings they make from it, because the latter will only affect highly skilled doctors, who will have to reduce their dedication to private practice to comply with earning constraints. Although this may be the case, it is also no less true that limits on dedication to private practice are much harder to implement and enforce than the ceiling on earnings.

Finally, regulations in countries such as Austria, France, Germany, Ireland, and Italy are oriented toward offering doctors incentives to enable their private practice from public hospitals, while specifying the maximum amount of private work that can be provided within public facilities. This kind of policy has the advantage of facilitating supervision, reducing opportunistic behavior, and easing the enforcement of restrictions.

Most dual practice control policies have been introduced in developed countries, wherein theoretically speaking, professional ethics and existing control mechanisms provide a guarantee against serious side effects. In the developing world, however, there is little control over dual practice, despite growing interest in the problem among public decision-makers in such areas. González and Macho-Stadler (2013) show that despite the risk of opportunistic behavior among dual practitioners in developing countries, direct regulation of dual practice is unlikely to be desirable because it carries the risk of pushing away the best doctors and thus reducing the population's access to quality health care. Furthermore, the implementation of any of these policies will require credible contracting institutions, which developing economies mostly lack. Therefore, in contexts such as these, regulations will only work subject to improvements in the contractual and institutional framework.

See also: Access and Health Insurance. Dentistry, Economics of. Moral Hazard. Physician-Induced Demand. Quality Reporting and Demand. Rationing of Demand. Risk Selection and Risk Adjustment. Waiting Times

References

Barros, P. P. and Olivella, P. (2005). Waiting lists and patient selection. *Journal of Economics and Management Strategy* **14**, 623–646.

Biglaiser, G. and Ma, C-t. A. (2007). *RAND Journal of Economics* **38**, 1113–1133.

Brekke, K. R. and Sørgard, L. (2007). Public versus private health care in a national health service. *Health Economics* **16**(6), 579–601.

Delfgaauw, J. (2007). Dedicated doctors: Public and private provision of heath care with altruistic physicians. Tinbergen Institute, Discussion Papers 07-010/1.

González, P. (2004). "Should physicians" dual practice be limited? An incentive approach. *Health Economics* **13**, 505–524.

González, P. (2005). On a policy of transferring public patients to private practice. *Health Economics* **14**, 513–527.

González, P. and Macho-Stadler, I. (2013). A theoretical approach to dual practice regulations in the health sector. *Journal of Health Economics* **32**(1), 66–87.

Iversen, T. (1997). The effect of a private sector on the waiting time in a national health service. *Journal of Health Economics* **16**, 381–396.

Rickman, N. and McGuire, A. (1999). Regulating providers' reimbursement in a mixed market for health care. *Scottish Journal of Political Economy* **46**, 53–71.

Further Reading

Eggleston, K. and Bir, A. (2006). Physician dual practice. *Health Policy* **78**, 157–166.

Ferrinho, P., Van Lerberghe, W., Fronteira, I., Hipólito, F. and Biscaia, A. (2004). Dual practice in the health care sector: Review of evidence. *Human Resources for Health* **2**, 1–17.

García-Prado, A. and González, P. (2007). Policy and regulatory responses to dual practice in the health sector. *Health Policy* **84**, 142–152.

García-Prado, A. and González, P. (2011). Whom do physicians work for? An analysis of dual practice in the health sector. *Journal of Health Politics, Policy and Law* **36**(2), 265–294.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization* **7**(Special issue), 24–52.

Socha, K. Z. and Bech, M. (2011). Physician dual practice: A review of literature. *Health Policy* **102**(1), 1–7.

Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes

C McCabe, University of Alberta, Edmonton, AB, Canada

Background

The past two decades have seen a revolution in the science that underpins new health technologies. Many new technologies offer hope for previously untreatable conditions and potential step changes in the outcomes of care for many others. Regulators committed to supporting the translation of the breakthroughs in biomedical knowledge into the clinic often approve new medicines on the basis of immature evidence, in terms of both the quantity and the quality of evidence for effectiveness and safety.

At the same time, the processes for manufacturing many of these new technologies are considerably more expensive than for conventional therapies such as small molecule pharmaceuticals. This, along with factors driving up research and development costs, means that many of these new technologies arrive at market with prices that are of magnitudes greater than previously encountered.

Improvements in population health status and the effectiveness of healthcare are also leading to increasing demand for treatment, as people live long enough to develop conditions associated with aging. As a result, not only the price but also the budget impact of introducing new technologies is much larger than was historically the case, with some technologies representing a genuine threat to the financial viability of smaller healthcare payers.

For reimbursement authorities charged with making protecting and promoting the health of the population they serve, the tension between the promise of these new technologies and the relative paucity of evidence that the promise will be fulfilled inevitably gives rise to the question 'What if the resources are consumed but the promise is not fulfilled?' The risk of making the wrong decision, the decision uncertainty, and how policy makers have responded are the focus of this article. The remainder of this article is constructed as follows. The Decision Uncertainty in Healthcare Resource Allocation section briefly reviews the concept of decision uncertainty and then outlines the factors that contribute to decision uncertainty in the context of value-based healthcare resource allocation decision processes. The Designing Patient Access Schemes section reviews the literature on their design and implementation, proposing a typology for differentiating them according to their objective and the mechanism for managing the decision uncertainty. The Evidence for the Success or Otherwise of Patient Access Schemes section considers the evidence for the success or otherwise of different types of scheme. The Linking Research and Reimbursement Decisions section considers recent developments linking research and reimbursement decisions, whereas Section Value-Based Pricing and Decision Uncertainty considers the degree to which a value-based pricing reimbursement framework might meet the needs of policy makers to manage the decision uncertainty around new and potentially innovative technologies. Eventually, Opportunities and Challenges in Emerging Decision Uncertainty Management Frameworks section considers emerging risk management frameworks from the perspective of health systems, manufacturers, and patients.

Decision Uncertainty in Healthcare Resource Allocation

Decision uncertainty can be thought of as the risk of making the wrong decision, the probability that the observed costs and outcomes will be sufficiently different from the expected costs and outcomes, that the option not chosen would have been better than the one chosen. Given that this is an unknown and one-off event, such probability is inherently a Bayesian rather than a frequentist concept. In the context of healthcare resource allocation decisions, the decision uncertainty is an accumulation of the uncertainty in the characterization of parameters in the decision problem, the characterization of the objectives of the decision process and the decision rules that flow from those objectives.

The extent to which decision makers need to take account of decision uncertainty is largely determined by the expected cost of making the wrong decision. The expected cost of making the wrong decision is determined by the probability of making the wrong decision given the currently available evidence and the expected health gains foregone and additional costs incurred due to making that decision. This is also known as the expected value of information (VoI). Where the expected cost of making the wrong decision is small, it is unlikely that investing in measures to reduce it will be justified, even when the decision uncertainty is high. Similarly, if the health gains foregone and costs incurred if the decision proves wrong are high, but the decision uncertainty is low, then investing in ameliorating mechanisms is unlikely to be efficient. However, with innovative technologies, healthcare systems are often facing a combination of high levels of decision uncertainty combined with high expected cost of uncertainty. In these circumstances, explicit consideration of mechanisms to reduce the expected cost of uncertainty is almost required for good stewardship of limited resources.

The simplest mechanism for reducing the cost of uncertainty is to reduce the cost of the technology; historically this has arguably been the most frequently used strategy. However, there are constraints on when this can be used and the degree of discount that is feasible. In pharmaceuticals, parallel imports, where drugs are bought in a low-price market and then sold in a higher price market, with the intermediary receiving the price premium rather than the manufacturer, are often cited as a reason for not using price discounts to manage risk. In some cases such as biologics, the cost of production of the technologies may mean that the scale of discount required

to reduce the cost of uncertainty to an acceptable level may be such that it would threaten its commercial viability.

If the expected cost of uncertainty cannot be reduced to acceptable levels through discounts, then there are a number of alternative drivers of the cost of uncertainty that can be addressed within the reimbursement decision process, some of which relate to the evidence for the technology, some to its clinical application, and some to the decision criteria used. Still others relate to the manner in which the technology is paid for by the health system.

Evidence and the Expected Cost of Uncertainty

The uncertainty in the evidence base for a new technology is the type of uncertainty that people consider most readily. It is recognized that for most new treatments, the evidence is provided by studies whose participants are very different from the patients who will be treated in everyday clinical practice; the length of follow-up in the studies is typically too short to provide any insight into the long-term effectiveness of the technology, and the number of participants is likely to be too small to uncover any rare but severe safety problems with the treatment. For technologies that treat previously untreatable conditions, it may even be the case that the health system will be substantially uncertain about how many patients have the condition of interest. To some degree the only way to be confident about how valuable a technology is, is to use it in a large number of 'typical' patients in routine clinical practice, and to do so for a reasonably long time. However, to do this simplistically entails the health system taking on the expected cost of the uncertainty. The development of Access with Evidence Development schemes, which are discussed in more detail in the section Decision Uncertainty and Innovative Payment Mechanisms, are an attempt to square the circle of generating real-world evidence on the value of a new technology while reducing a health system's exposure to the expected cost of the uncertainty.

Decision Uncertainty and the Clinical Application of a New Technology

Frequently the indications for the application of a new technology, as described in the license or the summary of product characteristics, tend to inclusivity. Where the value of the new technology is uncertain, reimbursement authorities will frequently seek patient subgroups within the licensed indication for whom there is evidence of a greater expected benefit than for the whole population. Even though, by definition, the uncertainty in the estimate for the effectiveness of this group is greater because the estimate is based on less data, the expected cost of uncertainty is reduced because the smaller number of patients reduces the budget impact and the expected value for the subgroup is more clearly below the decision threshold.

When reimbursement authorities are not in a position to identify a patient subgroup for whom to approve reimbursement, they may choose to impose a cap on the total budget impact of the technology. This strategy addresses the expected cost of uncertainty first by limiting the total expenditure directly and second by creating an indirect incentive for clinical practice to focus the utilization of the technology on those patients for whom it will be most beneficial. This strategy is particularly attractive where there is a significant risk of off-label use of the technology. In some cases, reimbursement authorities have linked manufacturers' payments to the achievement of predicted cost offsets in other areas of the budget, a form of financial risk sharing scheme that is distinct from the more widely known effectiveness-based risk sharing schemes. The advantage of this approach is that it creates an incentive for the manufacturers to discourage off-label usage as such use is less likely to generate the targeted cost offsets.

Decision Uncertainty and Reimbursement Decision Criteria

Healthcare resource allocation decisions are rightly subject to challenge by the patients, clinicians, and manufacturers who are affected by them. Arguably the most frequent challenge made to these decisions is that the value of the benefits of the technology has not been adequately captured in the evidence considered. Decision processes that assume the value of health gains are independent of the characteristics of the recipient, and are frequently challenged to take account of special factors such as the (lack of) alternative treatments, the severity of the condition, the imminence of death, the rarity of the condition, the age of the people affected, and even whether the health gain is produced in an innovative manner. All of these special factors attempt to shift the decision threshold and in doing so reduce the probability that the technology will prove not to be of good value, and thereby drive down the expected cost of uncertainty. The evidence base to specify decision criteria is both sparse and of variable quality. What evidence there is does not speak strongly to value premia for many of the proposed factors, but neither do they support a pure health gain maximization strategy. As a result the social legitimacy of these amendments to decision criteria frequently rests on the democratic legitimacy of the decision makers.

Decision Uncertainty and Innovative Payment Mechanisms

A final group of responses to decision uncertainty in reimbursement decision processes has been the development of new payment strategies. Conventionally, healthcare systems have paid for technologies in full prior at the time of their consumption, with the exception of large capital equipment such as magnetic resonance imaging (MRI) machines and surgical robots where leasing arrangements have been deployed. The effect of this is that all the risk associated with the uncertainty of the technology is transferred from the manufacturer to the health system before the outcome of treatment is known. Two distinct types of payment mechanism responses to this problem have been observed; the first operates at the individual patient level, whereas the second operates at the group level. Such schemes attempt to address decision uncertainty by reducing the expected budget impact and reducing the risk that payment will not produce the anticipated results.

Individual-level schemes – often referred to as payment by results, risk sharing, or Patient Access Schemes – link payment to outcomes for the individual patient. Beyond this basic shared characteristic, the specifics of the schemes vary. In

some, initial treatment is provided free of charge but only patients who respond to treatment continue with treatment that is paid for. With extremely expensive treatments, such as those for very rare diseases, the monitoring of response to treatment is sometimes a continuous process, so that if a patient stops responding, the funding can be stopped. In other schemes, patient treatments are funded up to a maximum number of administrations, after which the manufacturer provides the technology free of charge, the presumption being that only patients who are responding to treatment will remain on treatment beyond the maximum number of administrations.

Group- or population-level schemes tend to be referred to as Access with Evidence Development, Coverage with Evidence Development, or risk-sharing schemes. Under such schemes patients receiving the treatments will provide data on response to therapy as part of the scheme. These data are then used to inform a review of the reimbursement decision at a specified point in time. The review may lead to a change in the price or indeed a change in the reimbursement status. In principle, these group-level schemes offer a limit on the budget impact if the technology does not prove to be as valuable as hoped, and produces additional data to reduce the decision uncertainty.

Designing Patient Access Schemes

The range of policy responses to decision uncertainty in healthcare resource allocation has given rise to a relatively large number of labels in a remarkably short period of time: including risk sharing, coverage with evidence development, access with evidence development, patient access, and only with research (OWR). Behind all of these labels is a shared intention of achieving prompt patient access to the technology under consideration while attempting to ameliorate the expected cost of uncertainty associated with the reimbursement decision. Although the number of Patient Access Schemes is large, and the literature that comments around individual schemes is notable, substantial research on the principals that should inform their design and implementation is scarce.

The National Institute for Health and Clinical Excellence (NICE) has described the principals that will guide their assessment of a Patient Access Scheme, and the Commission for Medicaid and Medicare Services in the US is developing on guidance on the design of Coverage with Evidence Development schemes. In a similar vein, the International Society for Pharmacoeconomics and Outcomes Research is developing good practice guidance on the design of performance-based risk-sharing schemes. A Canadian group produced a consensus statement on the design of access with evidence development schemes in 2010, although this was based on a review of schemes that had worked well or not so well, rather than any clear theoretical framework.

Work on a theoretical framework for Patient Access Schemes is anchored in Decision Science, more specifically the VoI framework. Claxton and colleagues have focused on developing criteria to identify the efficient choice, for decision makers, between open and conditional reimbursement for a technology, and when the choice is for conditional reimbursement, to identify whether reimbursement should be only in research (OIR), i.e., only patients involved in the research have access to the technology, or OWR, which provides access for all patients as long as the research goes forward. Importantly, they have demonstrated that awaiting further research can be the correct decision even when the expected incremental cost effectiveness ratio is below the cost effectiveness threshold. It is the magnitude of the uncertainty, the budgetary impact of reimbursement, the feasibility undertaking the necessary research while the technology is generally available, and the reversibility of the investment that drive the value of further OIR or OWR Patient Access Schemes.

The work of Claxton and colleagues has tended to consider the burden of the uncertainty associated with specific parameters in a decision problem and not the details of the research that would be required to address that uncertainty. Their work is complemented by a series of publications from Willan and colleagues, who have developed methods for establishing the value of clinical trial research, taking due account of the time it takes for the research to report, costs incurred, and the value of any health gain foregone while the research is completed. More recently, Hall and colleagues placed this type of analysis in the decision framework used by Claxton, showing how to assess the expected value of an OIR and OWR Patient Access Scheme from the perspective of the healthcare payer. It is noteworthy that the work of Hall and colleagues indicates that OWR strategies are only likely to be an efficient use of health system resources when the expected cost of uncertainty is relatively low. These developments proffer real benefits to agencies interested in Patient Access Schemes as a means to reduce the expected costs of uncertainty by improving the evidence base for future decisions. However, this is only one component of the cost of uncertainty and only one of a number of possible objectives for a Patient Access Scheme.

Strictly, the VoI framework is not focused on Patient Access Schemes; rather it considers the most efficient means for generating additional evidence to inform a reimbursement decision. Achieving prompt patient access may be the result of such analyses, but it is not the primary concern. The primary concern is the risk that uncertainty in the evidence base may lead to the inappropriate reimbursement of an inefficient technology or the inappropriate rejection of an efficient one. The results of a well-designed and well-implemented VoI study may expedite or delay general patient access to a treatment. However, VoI does provide a framework for designing policy responses to uncertainty in the evidence base.

It may be useful, therefore, to differentiate between policy responses that have the reduction of uncertainty in the evidence base as their primary aim and policy responses whose primary aim is patient access to therapy. Within the former category, there are schemes that allow patient access to therapy while additional evidence is developed and schemes that constrain patient access in order to enable collection of additional evidence. In the latter category there are a range of schemes and they differ according to their secondary objective: (1) Patient Access Schemes that seek to reduce the cost per treated patient – in essence price discount schemes, (2) Patient Access Schemes that seek to limit the budget impact of the technology, (3) schemes that seek to target expenditure on those patients who respond to therapy, and (4) schemes that seek to develop evidence to inform future reimbursement decisions.

If policy responses that focus on the reduction of uncertainty are labeled Type 1, and responses that focus on patient access Type 2, six distinct categories of policy response to uncertainty in the evidence base can be defined: Type 1 OIR, Type 1 OWR, Type 2a, Type 2b, Type 2c, and Type 2d.

Understanding the specific type of scheme may help predict or explain observed policy responses to additional evidence. For example, Type 2d schemes may appear to be equivalent to Type 1 OWR schemes. However, the difference in the primary objective – reduced uncertainty versus patient access – is likely to lead to different policy responses to the same evidence. The UK multiple sclerosis risk sharing scheme is likely a Type 2d scheme. It was explicitly established to enable patient access to therapies that were not considered good value while at the same time collecting further evidence on the effectiveness of the therapies. Thus, accumulated evidence that might support changing the reimbursement status has received a very cautious policy response.

Evidence for the Success or Otherwise of Patient Access Schemes

The volume of Patient Access Schemes reported in the literature may well be the best evidence of their success. Decision makers keep returning to the schemes as a means of breaking the deadlock between patients and manufacturers on the one side and the limited resources of the healthcare system on the other. However, evidence that Patient Access Schemes have delivered affordable population health gain, or information to inform subsequent research decisions, is notable by its absence. Very few schemes have published reports of any data that have been collected and even fewer have reported estimates of the population health gain attributable to a scheme. By contrast, there is a substantial literature reporting problems with the process characteristics of Patient Access Schemes.

The Banff Workshop in 2010 identified problems with the process as one of the major impediments to success, where success was defined as observable changes in reimbursement and/or clinical practice in response to the evidence accumulated by the scheme. The same workshop, having reviewed published evidence and heard from experts involved in a number of different health systems, produced a consensus statement that emphasized the importance of governance in the establishment of schemes, if the intended objectives were to be achieved.

Linking Research and Reimbursement Decisions

Once a technology is licensed for use in a healthcare system, the typology of policy responses described in Section Designing Patient Access Schemes provides a useful framework for considering the linkages between further research and reimbursement. However, the reimbursement decision is also the mechanism that a healthcare payer has for signaling their willingness to pay for new technologies, and hence there is, conceptually at least, a link between reimbursement decision making and prelicensing research.

Considering postlicensing research and how it relates to reimbursement, the value of any research is dependent on the willingness of decision makers to change the reimbursement status in response to additional evidence. There is little, if any, convincing evidence that data collected as part of Type 2 Patient Access Schemes lead to changes in reimbursement status. The trial of lung volume reduction surgery, which many consider the first example of a Medicare Patient Access Scheme, produced strong evidence that the intervention was not effective and yet coverage of the procedure was not revoked. Similarly, the UK multiple sclerosis risk sharing scheme changed the rules in response to the first release of data, which indicated that the treatments were not effective, thereby avoiding a review of its reimbursement status.

There is an emerging interest in postlicensing trials of extremely expensive technologies such as Herceptin, rituximab, avastin, and lucentis (Table 1). Often these studies are at least partly funded by the healthcare payers and can often be cost saving in their own right, irrespective of the results that they provide, due to the cost of the technologies. However, there remains the challenge for decision makers of changing the reimbursement status when the studies report. In the absence of a commitment from decision makers to act on the evidence generated by such research, there must be questions about the ethics of enrolling patients to trials with a view to benefiting others, if the link to future reimbursement is fractured by the healthcare payers' reluctance to reverse a previous positive reimbursement decision. The recent decision by the UK NICE to reverse its positive recommendation for denosumab for prostate cancer is an encouraging development in this area.

Although postlicensing responses to uncertainty in the evidence base for new technologies are useful, they are a response to the problem that new technologies arrive at market that is often ill-suited to the needs of reimbursement decision makers. As market access is increasingly dependent on reimbursement rather than licensing and clinical use decisions, commercial considerations should lead companies engaged in developing new healthcare technologies to link their prelicensing research and development activities to the evidence needs of reimbursement decision makers. The VoI framework

Table 1 Examples of postlicensing trials of high-cost technologies

Study	*Technology*
FinHER study	Herceptin in early-stage breast cancer
PHARE trial	
Persephone study	
ARCTIC trial	Rituximab in CLL and in rheumatoid arthritis
SWITCH trial	
PARAMEDIC study	LUCAS CPR device
STAR trial	Sunitinib in locally advanced/metastatic renal cancer
IVAN trial	Ranibizumab in age-related macular degeneration
CATT trial	
OPTIMA study	Oncotype Dx for test-guided chemotherapy in early-stage breast cancer
RATPAC study	Point-of-care cardiac biomarkers
ELUCIDATE trial	Enhanced liver fibrosis test

that Claxton and colleagues have developed relatively fully for quantifying the value of research from the perspective of healthcare systems is increasingly being considered as a mechanism for evaluating prelicensing research investments.

The central observation that additional research, through reducing decision uncertainty, impacts on the probability of a positive or negative reimbursement decision is pertinent to both health systems and technology developers. However, the costs and benefits of positive and negative reimbursement decisions are very different. Specifying the cost and payoff functions for technology developers is arguably more complex than for healthcare systems, but it is likely to be at least as valuable. In a prelicensing context, additional research delays the time to licensing and thus the start of an income stream from the technology. It also 'burns' patent life and thus reduces the expected time from licensing to the onset of generic competition. These are costs that need to be captured in the evaluation of the investment. The characterization of benefit is also more complicated as evidence will inform decisions in a portfolio of healthcare systems. Each system is likely to operate different decision criteria and will differ in terms of the revenue stream to be expected conditional on a positive decision, reflecting different epidemiology and pricing policies. Willan and colleagues have started to examine these issues, but it remains a very immature literature. That said, it arguably has the greatest potential for matching research and development investment to technologies that will produce effects that patients and health systems will value and hence pay for.

A specific policy barrier to the wider utilization of these methods in the development of new technologies is the role of licensing authorities such as the Federal Drug Administration in the USA and the European Medicines Agency in Europe. Conventionally these organizations have focused on evidence of safety and efficacy rather than incremental value. Although the licensing authorities retain this focus and their approval continues to be a necessary but not sufficient condition for market access, technology developers will be understandably reluctant to adopt new strategies for designing and prioritizing prelicensing research and development. Although there are increasing communications between licensing and reimbursement authorities, the two communities appear to be still getting acquainted rather than developing coherent and complementary strategies focused on aligning research and the evidence needs of those who must decide whether and how much to pay for new technologies.

Value-Based Pricing and Decision Uncertainty

The concept of value-based pricing for health technologies is a relatively new one, but interest in it has grown rapidly. Initially, value-based pricing was thought of as a change in the mechanism for establishing the price of a treatment. Many healthcare systems are price takers. Reimbursement authorities consider whether a technology is of good value using the price that the manufacturer stipulates. The idea of value-based pricing is a simple one – why don't reimbursement authorities consider a technology and then specify the price at which it would represent good value. However, as consideration of how to operationalize this concept for real-world decisions has gathered pace, the debate about how to assess the value of a technology has intensified. In large part these discussions have covered the same arguments as the literature regarding the adequacy of the quality-adjusted life-year, as a measure of the effect of a technology, for use in cost effectiveness analysis. Equity arguments for value premia reflecting *inter alia* severity of ill-health, rarity, availability of alternative therapies, extensions of life at the end of life, and cause of disease have all been proposed as components of the assessment of value. The use of formal multicriterion decision-making processes has been proposed as a mechanism for capturing these disparate components of value. Although there is uncertainty if not outright ignorance about the relative and absolute value weights for these components of value, a multicriteria approach to resource allocation decisions is not a policy response to uncertainty.

Value-based pricing can be considered a policy response to uncertainty in that it allows decision makers to identify the price at which the expected cost of uncertainty does not support delaying reimbursement while further research takes place, or limiting reimbursed patient access during the research. However, in the context of the high cost and high levels of uncertainty associated with biologic, metabolomic, and genomic technologies, it is possible, even likely, that the price at which these conditions are met could have significant implications for the sustainability of private investment in health technology development and even the production of the technologies. In response to this concern, health policy makers are being encouraged to consider a premium for innovation and to allow prices to be revised upward if data from use of the technology in practice either reduce the uncertainty relating to its expected value or indicate that its actual value is greater than previously thought. The latter would represent a significant departure from current practice, where the price charged for a technology at launch is the highest price point, and subsequent developments will at best maintain the price and likely lead to price reductions. The data capture infrastructure required for the routine application of price adjustments based on observed effectiveness would be substantial, and it would be interesting to see whether there would be a symmetrical reluctance to increase price in the face of reduced uncertainty, reluctance to reduce price when evidence has suggested a technology has been less valuable than claimed. The incentives for gaming the system when the evidence of value is not derived from well-conducted randomized controlled trials will likely be significant, especially for therapies for common disorders.

The innovation premium is the added value attributed to a technology that does something that current technologies do not do, over and above the value attached to its effectiveness compared to currently available technologies. The justification for such a premium would likely rest in either the option value of subsequent alternative applications of the technology to meet other currently unmet needs or, when there is no evidence to support such an expectation, the value of the hope for such application. What is clear is that the innovation premium is a reward for 'newness' and thus likely to be highly and positively correlated with uncertainty. As such, the innovation premium works in the opposite direction to the expected cost of uncertainty and increases the likelihood of a positive reimbursement decision at any given level of decision

uncertainty. In the context of population health promotion, the magnitude of the innovation premium should depend on the option value of future potential applications. However, when healthcare organizations have implicit or explicit industrial policy objectives, such as the National Health Service in the UK and the Commission for Medicare and Medicaid Service in the USA, the magnitude of premium may partially reflect these considerations also.

A further potentially problematic characteristic of value-based pricing as a mechanism for addressing uncertainty is the creation of different prices for the same technology in different markets. Manufacturers are understandably concerned about differential pricing creating opportunities for parallel imports of their technologies acting as a downward pressure on their global revenue. Consumers in high-price environments such as the US, Germany, or France might procure their therapies in lower price markets such as Canada, Poland, and Spain. However, there are additional challenges associated with value-based pricing schemes for manufacturers. First, because knowledge is essentially a public good, health systems that are not engaged in a value-based pricing schemes may be able to free ride, even if the specifics of the additional evidence generated is kept confidential, as any change in price will be informative. Second, because healthcare systems differ in terms of budgets and the epidemiology of disease in the populations served, there will likely be much larger variation in prices between health systems than is currently observed, with an associated increase in the uncertainty in the expected return for investors.

Although value-based pricing is intuitively appealing as a response to uncertainty that will not require the reengineering of existing research and development processes, its operationalization for highly uncertain high-cost technologies may not be consistent with the sustainability of research and development investment. Further, its implementation may introduce new parameters into the decision problem about which there is substantial uncertainty from the healthcare payer perspective. At the same time, it may generate additional uncertainty regarding revenue flows for the manufacturers and investors, as prices arrived at through value-based pricing processes may differ markedly between healthcare markets even when the value criteria used are shared. New evidence generated through a value-based pricing mechanism in one system will likely influence prices in other systems; this may not be symmetrical; i.e., health systems may be more likely to 'free ride' on knowledge that supports a price cut than knowledge that supports a price rise. It may be that manufacturers would be more attracted to lower but more certain returns on their investment, compared to opening this Pandora's box.

Opportunities and Challenges in Emerging Decision Uncertainty Management Frameworks

Broadly there are four strategies for addressing the decision uncertainty facing reimbursement authorities driven by the mismatch between the evidence produced by conventional health technology development processes and the evidence required to inform efficient and equitable use of limited healthcare budgets: (1) Patient Access Schemes, which focus on achieving patient access alongside one or more secondary objectives such as per patient cost containment, total cost containment, and targeted use or evidence development; (2) research, which has the reduction of decision uncertainty as its primary objective. This may require that the technology is not available to patients except as part of the research (OIR) or may allow access to the treatment if that does not confound the required research study or is required for the research to proceed (OWR); (3) value-based pricing, which sets the price of the technology at a level that reduces the expected cost of uncertainty associated with reimbursement below the cost of requiring further research; and (4) reengineering the pre-licensing research and development process to meet the needs of reimbursement decision makers.

The first three strategies treat the focus of current research and development processes on the requirements of licensing authorities as immutable. They are more or less focused on the needs of the identified patients who will benefit from the newly licensed technology (1 and 3), or the needs of the unidentified patients who will bear the opportunity cost of reimbursing the new technology (2 and 3). The fourth strategy perhaps naively assumes that the structures within which technologies are developed can be redesigned and considers how it might be redesigned to match technology development investments to the objectives of healthcare systems.

The time it takes for investments in research and development to pay off means that policies that address problems with the conventional evidence development processes will be required for many years to come. However, this does not mean that work on developing a more efficient research and development process, focused on developing high-value technologies with 'reimbursable' evidence dossiers at the time of licensing, is not worth investment. That said, there are significant challenges to be addressed in developing the VoI framework to inform the design of research and investment processes.

VoI is predicated on a clearly specified payoff function – whether it be population health benefit or revenue from sales. For investors in mid- to late-stage clinical trials, the payoff function of interest is conditional on the objectives of the portfolio of healthcare systems that are the clients for the technologies they are seeking to develop. The methods for representing and combining these functions in assessing the value of alternative investments may not be intellectually trivial challenges.

Skeptics are likely to argue that much of the information that is pertinent to the decision problem cannot be known with any confidence so far in advance of the decision, and therefore early-stage VoI analyses are likely to involve as much guess work as knowledge. Although this is true, to use it as the basis for rejecting changes in the approach for designing research and development processes is to assume that a similar degree of guess work is not implicitly or even explicitly involved in the current processes. Given the high failure rate in the research and development process, and the problem with licensed technologies struggling to achieve reimbursement, it seems likely that the current process is based on at least an equally flawed assessment of the values and needs of future healthcare systems.

There are short-, medium-, and long-term challenges facing healthcare systems seeking to take a systematic approach to

managing the uncertainty in reimbursement decisions. In the short term, Patient Access Schemes are likely to be more not less prevalent and thus the total value of resources invested is likely to increase. Experience to date does not provide confidence that these schemes are automatically of good value to the health systems that enter into them. Careful design and governance may reduce the cost of uncertainty associated with these schemes. In the medium term, all these schemes rely to a substantial degree on capturing reliable evidence on the impact of therapies on patients in the typical clinical setting. Few health systems currently have routine data capture infrastructure fit for this purpose. The capacity for establishing successful Patient Access Schemes may well be among the more valuable, if less noticed, returns on investing in such infrastructure. On a longer term, all reimbursement authorities in healthcare – innovators, investors, regulators, clinicians, patients, and health systems – need to find mechanisms to align the research and development processes with the needs of all patients, signaling societies' willingness and ability to pay for health gain, so that the current incentives to invest large sums in high-risk candidates that may produce only marginal health gains are removed, leading to fewer marginal value and highly uncertain technologies being launched.

See also: Information Analysis, Value of. Primer on the Use of Bayesian Methods in Health Economics

Further Reading

Chen, M. H. and Willan, A. R. (2013). Determining optimal sample sizes for multistage adaptive randomized clinical trials from an industry perspective using value of information methods. *Clinical Trials* **10**(1), 54–62.

Claxton, K. (2007). OFT,VBP:QED? *Health Economics* **16**, 545–558.

Claxton, K., Palmer, S., Longworth, L., et al. (2012). Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. *Health Technology Assessment* **16**(46), 1–323.

Conti, S. and Claxton, K. (2009). Dimensions of design space: A decision theoretic approach to optimal research portfolio design. *Medical Decision Making* **29**, 643–660.

Department of Health (2011). A new value-based approach to the pricing of branded medicines: Government response to consultation. London. Available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_128404.pdf (accessed 12.04.13).

Desser, A. S., Gyrd-Hansen, D., Olsen, J. A., Grepperud, S. and Kristiansen, I. S. (2010). Societal views on orphan drugs: Cross sectional survey of Norwegians aged 40 to 67. *British Medical Journal* **341**, c4715.

Eckermann, S., Karnon, J. and Willan, A. R. (2010). The value of information: Best informing research design and prioritization using current methods. *PharmacoEconomics* **28**(9), 699–709.

Eckermann, S. and Willan, A. R. (2008). The option value of delay in health technology assessment. *Medical Decision Making* **28**, 300–305.

Eckermann, S. and Willan, A. R. (2009). Globally optimal trial design for local decision making. *Health Economics* **18**, 203–216.

Griffin, S., Claxton, K. and Welton, N. (2010). Exploring the research decision space: The expected value of information for sequential research designs. *Medical Decision Making* **30**, 155–162.

Grossman, G. M. and Lai, E. L. (2008). Parallel imports and price controls. *RAND Journal of Economics* **39**(2), 378–402.

Hall, P. S., Edlin, R., Kharroubi, S., Gregory, W. and McCabe, C. (2012). Expected net present value of sample information from burden to investment. *Medical Decision Making* **32**(3), E11–E21.

Linley, W. G. and Hughes, D. A. (2012). Societal views on nice, cancer drugs fund and value-based pricing criteria for prioritising medicines: A cross-sectional survey of 4118 adults in Great Britain. *Health Economics,* doi:10.1002/hec.2872.

McCabe, C., Claxton, K. and O'Hagan, A. (2008). Why licensing authorities need to consider the net value of new drugs – addressing the tension between licensing and reimbursement. *International Journal of Technology Assessment in Health Care* **24**, 140–145.

McCabe, C. J., Stafinski, T., Edlin, R., Menon, D. and Banff AED Summit (2010). Access with evidence development schemes: A framework for description and evaluation. *PharmacoEconomics* **28**(2), 143–152.

McKenna, C. and Claxton, K. (2011). Addressing adoption and research design decisions simultaneously: The role of value of sample information analysis. *Medical Decision Making* **31**, 853–865.

Menon, D., McCabe, C. J., Stafinski, T., Edlin, R. and Signatories to the Consensus Statement (2010). Principles of design of access with evidence development approaches: A consensus statement from the Banff Summit. *PharmacoEconomics* **28**(2), 109–111.

Mohr, P. E. and Tunis, S. R. (2010). Access with evidence development: The US experience. *PharmacoEconomics* **2**, 153–162.

NICE (2008). *Guide to the methods of health technology appraisal*, 3rd ed. London: NICE.

Sculpher, M. J., Claxton, K., Drummond, M. and McCabe, C. (2006). Whither trial-based economic evaluation for health care decision making? *Health Economics* **15**(7), 677–687.

Stafinski, T., McCabe, C. and Menon, D. (2010). Funding the unfundable: Mechanisms for managing uncertainty in decisions on the introduction of new and innovative technologies into healthcare systems. *PharmacoEconomics* **28**(2), 118–142.

Towse, A. (2010). Value based pricing, research and development, and patient access schemes. Will the United Kingdom get it right or wrong? *British Journal of Clinical Pharmacology* **70**(3), 360–366.

Towse, A. and Garrison, L. P. (2010). Can't get no satisfaction? Will pay for performance help? Toward an economic framework for understanding performance-based risk-sharing agreements for innovative medical products. *PharmacoEconomics* **28**(2), 93–102.

Wailoo, A., Tsuchiya, A. and McCabe, C. (2009). Why weighting must wait: Incorporating equity concerns into cost-effectiveness analysis may take longer than expected. *PharmacoEconomics* **27**(12), 983–989.

Pollution and Health

J Graff Zivin, University of California, San Diego, La Jolla, CA, USA
M Neidell, Columbia University, New York, NY, USA

Glossary

Ambient pollution Harmful contamination within a surrounding area, generally expressed as a concentration per unit of volume of air, for example, micrograms per cubic meter of air.

Avoidance behavior Actions an individual takes to reduce exposure to a pollutant.

Willingness to pay (WTP) The maximum amount of income a person would be willing to pay to reduce exposure to a certain amount of pollution.

Introduction

A primary objective of environmental policies worldwide is to protect human health. Optimal policy design, however, is typically hampered by limited information regarding both the benefits and the costs associated with regulation. Benefits assessments frequently rely on translating laboratory findings to uncontrolled settings, extrapolating from high- to low-concentration exposures within and across societies, and drawing inferences from observational analyses that do not account for the endogeneity of pollution. Economic assessments have typically focused on the costs of compliance to firms. Efforts to improve societal welfare clearly depend on a strong understanding of both elements. Although the health-pollution relationship largely remains the pursuit of epidemiologists, the focus of economics on casual identification along with valuation techniques consistent with utility maximization has helped to reframe these relationships in a manner that facilitates policy choice and environmental rule setting.

Early epidemiological investigations of the impacts of extreme pollution events were some of the first compelling studies to suggest a causal relationship, with one of the most famous focused on the 'killer fog' in London, England in December, 1952. A temperature inversion combined with windless conditions led to a sudden and dramatic increase in air pollution. Because residents were used to winter fogs, there was little, if any, changes in behavior, leading to a rather clean measure of pollution impacts in this case. The dramatic rise in mortality that precisely coincided with the timing of the fog had been a driving force behind the federal regulations aimed at air pollution control.

The pollution levels experienced under this and similarly studied extreme events, however, had been dramatically higher than those that nearly all people in developed countries face today. Moreover, most exposures do not conveniently arrive as a 'surprise' under which causal impacts can be easily assessed, and it is on this front that economists have made their most significant contributions. In particular, economic studies have typically focused on quasi-experimental settings in order to synthesize the 'surprise' of pollution. Besides improving the causal understanding of these relationships by minimizing threats from confounding, it has also identified important compensatory behaviors undertaken by individuals to mitigate exposure.

These behavioral responses are often nontrivial because many pollutants are observable, and even those that are not easily detectable by the public, such as ozone and particulate matter, are forecast and publicized through a broad range of media outlets. If optimizing individuals compensate for changes in ambient pollution levels by reducing their exposure, estimates that do not account for these responses will understate the biologic relationship between ambient pollution levels and health. This problem is potentially severe because the more an individual is likely to suffer under pollution, the more they have to gain from reduced exposure. Indeed, emerging empirical evidence finds that behavioral responses are largest for more vulnerable individuals (Neidell, 2009; Graff Zivin and Neidell, 2009; Graff Zivin *et al.*, 2011) and that individuals are more responsive to higher levels of pollution (Neidell, 2009; Mansfield *et al.*, 2006). Equally important, these behavioral responses are costly and thus ignoring them will also understate the welfare effects of pollution (or the regulation thereof). Although the costs of spending additional time indoors, rescheduling activities, or even relocating to areas with better environmental quality are often difficult to enumerate, they can represent a substantial fraction of the total costs of pollution.

In the remainder of this article, a basic economic framework for evaluating environmental health impacts is presented, followed by a discussion of the core empirical challenges that researchers face in estimating the relationship between pollution and health. A selective review of significant contributions from the literature that focus on the effects of air pollution is then provided, concluding with some suggestions for fruitful lines of future research.

Conceptual Framework

Estimation of the relationship between pollution and health is typically focused on the following health production function:

$$h = f(\mathrm{P}, \mathrm{A}, \mathrm{E}, \mathrm{S}) \quad [1]$$

where h is a measure of an individual's health, P is pollution levels assigned to the individual, and A is avoidance behavior. E are other environmental factors that directly affect health, such as weather and allergens, and S are all other behavioral, socioeconomic, and genetic factors affecting health. Given that

 doi:10.1016/B978-0-12-375678-7.00306-0

meteorological elements can play an important role in pollution formation and can also affect health (e.g., cold weather increases asthma exacerbation), E is defined separately because it represents an important source of environmental confounding.

Two main approaches are taken to eqn [1], with the difference stemming from the treatment of avoidance behavior. The first, or 'reduced-form' approach does not directly control for avoidance behavior. As health impacts will depend on ambient pollution levels and avoidance behavior that determines exposure to those pollution levels, the health relationship can be expressed as the following total derivative: $dh/dP = \delta h/\delta P + \delta h/\delta A * \delta A/\delta P$. The second, or 'production function' approach directly controls for avoidance behavior to obtain the partial derivative: $\delta h/\delta P$.

The importance in separating these two approaches is to relate each to the benefit calculation, or willingness to pay (WTP) for a reduction in pollution (Harrington and Portney, 1987; Cropper and Freeman, 1991; Deschenes and Greenstone, 2011). In the reduced-form approach, welfare is typically expressed as: $WTP = dh/dP * C_h + p_A * \delta A/\delta P$, where C_h is the 'full' cost associated with a change in health, and p_A is the price of avoidance behavior. In the production function approach, welfare is typically expressed as: $WTP = p_A * [(\delta h/\delta P)/(\delta h/\delta A)]$. Although the production function approach appears more data hungry because of the need to control for avoidance behavior when estimating eqn [1], the reduced-form approach must also control for avoidance behavior in order to estimate $\delta A/\delta P$, although this can be done separately from estimating eqn [1]. Furthermore, as these expressions demonstrate, all forms of avoidance behavior must be accounted for at some point in order to obtain a proper estimate of WTP.

One advantage of the reduced-form approach is that the econometrician does not need to properly specify the functional form of eqn [1] with respect to P and A. This is particularly helpful because data limitations often necessitate the use of proxy measures for avoidance behavior, and economic theory provides little guidance on how these proxy measures should enter into eqn [1].

The value of the production function approach is that it provides estimates of the biological effect of pollution. Because avoidance behavior is likely to vary across socioeconomic and cultural environments, but the biology is considerably less context specific, it facilitates generalizations across settings. Moreover, focusing on the biological effect enables one to potentially identify important nonlinear effects, such as threshold effects, and heterogeneous effects based on individual susceptibility, both of which can play an important role in defining the feasible set of policy interventions. Interested readers should consult Graff Zivin and Neidell (2013) for more elaboration on this framework.

Empirical Challenges

In this section, three primary challenges confronted by empiricists when estimating the relationship between pollution and health are outlined. Although weather is a potential confounder, this is not discussed at length because it is directly observable (often at a finer scale than pollution data), so that any threat can be obviated through the careful control of relevant variables.

Measurement of Health

The measurement of health outcomes and how to place monetary values on them is a persistent challenge. A frequently used measure is mortality, which is objectively measured and can be readily monetized using estimates of the value of a statistical life (VSL). One concern with using mortality is that it is an extreme outcome that misses more subtle outcomes that may be more commonplace. Furthermore, using VSL to monetize these impacts may be misleading if the loss only represents short-term mortality displacement, commonly referred to as 'harvesting.'

Measures of morbidity have also been examined using data on hospitalizations for various conditions, largely respiratory related. Although hospitalizations clearly capture events less severe than death, they may introduce sample selection. Those who have a relationship with a primary care physician (PCP) and receive regular care may never experience a hospitalization, and access to a PCP is clearly endogenous. Furthermore, the economic valuation of hospitalizations is particularly difficult as hospital charges (which are all that is typically available) do not capture the costs associated with the pain and suffering experienced by sickened individuals or their family members.

Birth outcomes are another metric that has some of the desirable properties of both mortality and morbidity endpoints, albeit for a select population. Like mortality and hospitalizations, birth outcomes are a census and not a sample, hence offering large sample sizes for analysis. Unlike mortality, birth outcomes can capture more subtle impacts, and unlike hospitalizations, they do not introduce sample selection because any birth that files for a birth certificate is reported. Valuation approaches can be used when the birth outcome studied has been linked to monetizable events – for example, birth weight has been linked with education and earnings (Black *et al.* 2007) – although these links may not capture all relevant costs.

An emerging area of focus is on indirect 'health' outcomes at school or the workplace, principally absenteeism and performance. Such outcomes offer terrific promise for capturing rather subtle health impacts that might be broadly disseminated throughout society. They are also generally straightforward to monetize, particularly for performance. Limited data availability, especially for representative samples, is a formidable obstacle to the conduct of credible empirical work in this area.

Assignment of Local Pollution Levels

Most studies focus on air pollution because of the availability of data from ambient air pollution monitors, which typically measure air concentrations at an hourly scale at a fixed location. Although this frequency of measurement generates data at a fine temporal scale, the limited number of monitor locations relative to the size of a country and the geographic distribution of the population leads to data that are rather

coarse on a spatial scale. As a result, studies often approximate contemporaneous pollution levels based on an individual's general location and the location of the monitor. This crude approach leads to measurement error that increases with an individual's distance from the monitor and the degree to which pollutants disperse nonuniformly. This measurement error will typically bias estimates downward, but with a large enough dataset, researchers can use data from multiple monitors and various weighting techniques to obtain more precise assignments of localized pollution levels. A finer level of geographic disaggregation for individuals, such as a residential address, also allows for better assignment of relevant pollution levels and hence is more likely to provide precise estimates.

The usual mobility of individuals in their everyday life (not in response to pollution, discussed below), both within a day and over time, can also present a measurement issue. Individuals spend their time not only at home, but at work, school, and other possible locations that are not typically recorded. Although the use of personal monitors is designed to overcome this, two issues remain: (1) the high costs of personal monitoring often result in the use of a small, unrepresentative sample without a clearly defined control group; and (2) the link to policy is less clear because indoor sources also contribute to pollution but are subject to different regulatory rules. Mobility over time also presents a significant measurement issue in assigning cumulative exposure over longer periods of time. Focusing on children, and in particular infants, whose parents are typically less mobile, can greatly limit this concern (Joyce *et al.* 1989; Chay and Greenstone, 2003).

Behavioral Responses to Pollution

Optimizing individuals may respond to pollution with permanent changes, such as relocating (i.e., sorting), and temporary changes, such as spending less time outside. As argued above, it is crucial to understand the role of these behavioral responses both to allow generalizations from one setting to another and to account for the full welfare costs of pollution. Although careful quasi-experimental designs can address permanent behavioral changes by exploiting exogenous shocks to pollution levels, short-run changes pose greater challenges because many of these responses involve nonmarket behaviors that are difficult to observe. For example, simply spending less time outside on a polluted day is a highly effective means for reducing exposure, but such an activity is rarely recorded. Clearly, the degree to which such short-run behavioral responses will be important depends on the 'visibility' of pollution, either literally, through information dissemination, or through health feedbacks that allow individuals to infer it on the basis of physiological responses.

Evidence

Rather than provide an exhaustive review of the economic literature examining the relationship between pollution and health, this section limits its attention to a selection of studies that offer key insights or introduce important methodological advances.

Primary Impacts

One of the earliest examples of a quasi-experimental approach to estimate an environmental health relationship in relatively recent times is found in a series of studies by Pope *et al.* (1992); Ransom and Pope (1992); Ransom and Pope (1995). The authors used changes in pollution that had resulted from the opening and closing of a steel mill, which was a major source of particulate matter, in the central Valley of Utah due to a labor strike. As the steel mill had closed due to a labor strike, the temporary changes in pollution were credibly exogenous and unlikely to lead to any immediate residential sorting. Furthermore, the authors selected a neighboring, unaffected community as a control group to account for time trends by estimating difference-in-differences models. When the steel mill was closed, the authors found significant declines in school absences, respiratory-related hospital admissions, and mortality. One potential concern with this study is that the steel mill closure has also led to a temporary change in income, which may affect one's use of time and services. This does not seem likely to be an issue for school absences, hence at least some of the findings are credibly causal. A more significant concern with the design is that, as an 'event study,' the pollution variable is common to all members in a group for a given time period (despite the availability of individual level health outcomes as dependent variables). As a result, their standard errors are likely to be nontrivially understated, making the appropriate statistical inference in this setting particularly challenging (Donald and Lang, 2007).

One important study by Chay and Greenstone (2003) overcame this problem by focusing on the recession of the early 1980s. The dramatic change in manufacturing that had resulted from this recession induced considerable spatial variation in total suspended particulates (TSPs) throughout the US in a short period of time, with some areas experiencing as large as a 35% decline in 3 years. These changes in TSPs are unlikely to be related to other factors affecting health. Importantly, although income changed considerably at the same time, it did not show comparable spatial patterns as with TSP. Using this exogenous variation in levels of pollution at the county-year level to identify environmental health effects, they estimate that a one-unit decline in TSPs associated with the recession yields benefits of roughly US$14 billion, recognizing that this captures only one health outcome and only for a specific group.

Although the Chay and Greenstone results are nontrivial, the continued improvements in air quality since then suggest that the results also apply to a time period when pollution levels in the US are considerably higher. Currie and Neidell (2005) turn their attention to infant mortality in California during the 1990s, a period that is much more reflective of contemporary pollution levels across much of the developed world. They use zip code fixed effects to account for residential sorting, thereby exploiting the strong temporal variations in pollution levels in the short-run due to changes in plausibly exogenous ambient conditions (rather than anthropogenic sources) to identify health impacts. They find that reductions in carbon monoxide over the 1990s saved approximately 1000 infant lives in California, which translates into benefits of roughly US$4.8 billion.

Currie *et al.* (2009), like Currie and Neidell (2005), focus on infant outcomes in a more recent time period, but use the exact address of the mother to improve pollution assignment and estimate sibling fixed effect models to control for differences in family background and genetics. They find that a one-unit change in mean carbon monoxide (CO) during the last trimester of pregnancy increases the risk of low birth weight by 8%, and a one-unit change in mean CO during the first two weeks after birth also increases the risk of infant mortality by 2.5% relative to baseline levels. The authors calculate that the 15-year decline in CO from 1989–2003 translates into US$720 million in lifetime earnings from improvements in birth weight and US$2.2 billion from the reduction in infant mortality for the 2003 birth cohort. The use of sibling fixed effects increases estimates, suggesting the importance of accounting for maternal characteristics within neighborhoods. And the better assignment of pollution by using the mother's exact address rather than zip code also increases point estimates, consistent with measurement error inducing a downward bias.

In a novel design, Lleras-Muney (2010) uses the relocation of military personnel to estimate the effect of various pollutants on children's health. The relocation of personnel is entirely based on 'the needs of the army', which explicitly rules out the possibility of sorting and offers a plausibly exogenous source of variation in pollution. Using this design, Lleras-Muney finds that a one standard deviation decrease in ground-level ozone exposure decreases the probability of a respiratory hospitalization for children by 8–23%. Her estimates suggest that lowering pollution levels nationwide to the levels experienced in 'low' pollution areas would save approximately US$928 million (US$1994) in direct medical expenditures alone.

All of the previously mentioned studies exploit 'natural' experiments that generate exogenous changes in ambient pollution in order to minimize concerns regarding residential sorting and other long-run behavioral responses to poor environmental quality. They generally ignore potential short-run adjustments that could also impact the environment-health relationship, and hence provide estimates of a reduced-form relationship between pollution and health. The key challenge in capturing these short-run behavioral responses is clearly the availability of data suited for the task, and researchers often follow creative paths for obtaining such data. One example is Neidell (2009), who uses attendance data from several outdoor facilities in Los Angeles to uncover significant behavioral responses to high ozone levels that are forecasted through smog alerts. As smog alerts are issued only when ozone is forecasted to exceed a particular threshold, he employs a regression discontinuity design to compare attendance on days just above the threshold to that just below. Although this paper does not provide estimates of the costs of avoidance behavior, in a closely related paper Graff Zivin and Neidell (2009) examine successive days of smog alerts to show that the costs of avoidance behavior, due to limited opportunities for intertemporal substitution, are increasing over time. Graff Zivin *et al.* (2011) identify substantial increases in the purchase of bottled water when local municipalities violate drinking water standards. As this type of avoidance behavioral is market-based, the authors have calculated the costs associated with it, and have found that water quality violations in 2005 induced roughly US$60 million worth of bottled water purchases nationwide.

Two notable studies attempt to produce estimates of the biological effect of ozone on health. In the paper discussed earlier, Neidell (2009) controls for smog alerts and ozone forecasts as a proxy for avoidance behavior when estimating the relationship between ozone and respiratory-related hospitalizations. Using zip code fixed effects and exploiting the strong daily temporal variation in ozone, he finds that including these proxies significantly increases the estimated impact of ozone on health. Moretti and Neidell (2011) use daily boat arrivals and departures into the port of Los Angeles as an instrumental variable (IV) for ozone levels, which deals with both avoidance behavior and measurement error in pollution assignment. Boat traffic represents a major source of pollution for the Los Angeles region and, because of the extended length of travel and unpredictable conditions at sea, daily variation in boat traffic is arguably uncorrelated with other short-run determinants of health and is not included in the ozone forecasts used to encourage avoidance behavior. Similar to Neidell (2009), they find that using boat traffic as an IV leads to significantly larger estimates for the impacts of pollution on health.

Although the short-run behavior literature has generally assessed the costs associated with avoiding exposure, this again represents a partial characterization of social welfare; a complete calculation requires an assessment of both avoidance costs as well as the costs of those adverse health effects that are not avoided. To our knowledge, the only attempt to bring both pieces together in a quasi-experimental setting is from Deschenes and Greenstone (2011), who focus on the health effects of extreme temperatures, which are forecast to increase under climate change. They construct a WTP estimate to avoid extreme heat that includes the costs due to excess mortality as well as expenditure on energy consumption as a proxy for air conditioning usage to buffer individuals from exposure to that heat. Using county fixed effects to exploit the plausibly exogenous variation in temperatures in an area within a given year, they find that the avoidance costs are roughly 25% of the mortality costs.

Secondary Impacts

Although most of the literature has focused on primary health endpoints, for example, mortality and hospitalizations, an emerging literature has begun to examine the manifestation of less visible health assaults on nonhealth outcomes. Although these impacts are referred to as secondary, it remains possible for them to exceed the costs of primary impacts depending on their prevalence. Almond *et al.* (2009) examine the impact from prenatal exposure to radioactive fallout from the 1986 Chernobyl accident on both birth and schooling outcomes for children in Sweden. Although Sweden is more than 500 miles away from Chernobyl, weather conditions forced some of the plume over Sweden, and local variation in rainfall levels led to stark geographic variation in the levels of fallout throughout the country. Their study reveals that radiation exposure exhibits latent effects that affect human capital development

later in life. Although they find little evidence of health effects as measured by birth outcomes and childhood hospitalizations, they find significant decreases in several schooling outcomes that correspond to roughly US$510 million in lost annual earnings.

Graff Zivin and Neidell (2012) also follow a nontraditional approach by examining the impact of ozone on worker productivity. They use a unique dataset on agricultural workers who are paid by piece rate and whose labor supply is highly inelastic in the short run, hence limiting the scope for avoidance behavior and the need to value it. Using models with worker fixed effects to exploit plausibly exogenous daily variation in ozone levels, they find that a 10 ppb decrease in ozone concentrations increases worker productivity by 5.5%, which translates into productivity benefits to the agricultural industry of approximately US$700 million.

Conclusion

Pollution affects a wide range of health outcomes, and these effects are nontrivial even at current emissions levels in the developed world. The optimal level of these pollutants is highly contested. For example, a proposed ozone standard issued by the EPA in 1997 was finally upheld by the Supreme Court in 2002, but only after endless appeals and lengthy lawsuits initiated by states and industry (Bergman, 2004). Better estimates of the relationship between pollution and health and society's WTP for improvements in pollution through the use of quasi-experimental research designs offers an important tool for informing this debate. Additional work on the measurement of avoidance behavior and its costs remains a critical piece of the puzzle.

Despite the growth of quality evidence on this topic, one area in need of more evidence is on the long-run effects from cumulative exposure to various pollutants. Although it is clear that pollution has shortrun impacts, the potential impacts from exposure over a lifetime may be considerably larger, as hinted at by the results from Almond *et al.* (2009). These impacts may also affect people's investment decisions throughout their life course, suggesting a wide range of potential economic outcomes that may be affected. The empirical issues are more daunting given the challenges in appropriately measuring health outcomes and pollution exposure, and the ability to isolate exogenous variation in pollution, but nonetheless deserve more attention.

The impact of pollution on human capital formation and its deployment in school as well as labor markets also represents a particularly fruitful area for additional exploration. The use of these indirect outcomes can capture a broader range of economic impacts, and they also have the ability to capture subtle, but likely more pervasive health impacts than those captured through standard measures of mortality and hospitalizations. As the improvement in biomedical understanding of the etiology of disease continues, this area of study is likely to rise in frequency and importance.

See also: Health Status in the Developing World, Determinants of. Willingness to Pay for Health

References

Almond, D., Edlund, L. and Marten, P. (2009). Chernobyl's subclinical legacy: Prenatal exposure to radioactive fallout and school outcomes in Sweden. *Quarterly Journal of Economics* **124**(4), 1729–1772.

Bergman, C. (2004). *EPA issues designations on ozone health standards.* Washington, DC: U.S. EPA Press.

Black, S., Devereux, P. and Salvanes, K. (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics* **122**, 409–439.

Chay, K. and Greenstone, M. (2003). The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* **118**(3), 1121–1167.

Cropper, M. and Freeman, III, A. M. (1991). Valuing environmental health effects. In Braden, J. and Kolstad, C. (eds.) *Measuring the demand for environmental commodities*, pp. 165–212. Amsterdam: North-Holland.

Currie, J. and Neidell, M. (2005). Air pollution and infant health: What can we learn from California's recent experience? *Quarterly Journal of Economics* **120**(3), 1003–1030.

Currie, J., Neidell, M. and Schmieder, J. (2009). Air pollution and infant health: Lessons from New Jersey. *Journal of Health Economics* **28**(3), 688–703.

Deschenes, O. and Greenstone, M. (2011). Climate change, mortality and adaptation: Evidence from annual fluctuations in weather in the U.S. *American Economic Journal: Applied Economics* **3**(4), 152–185.

Donald, S. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics* **89**(2), 221–233.

Graff Zivin, J. and Neidell, M. (2009). Days of haze: Environmental information disclosure and intertemporal avoidance behavior. *Journal of Environmental Economics and Management* **58**(2), 119–128.

Graff Zivin, J. and Neidell, M. (2012). The impact of pollution on worker productivity. *American Economic Review* **102**(7), 3652–3673.

Graff Zivin, J. and Neidell, M. (2013). Environment, health, and human capital. *Journal of Economic Literature* **51**(3), 689–730.

Graff Zivin, J., Neidell, M. and Schlenker, W. (2011). Water quality violations and avoidance behavior: Evidence from bottled water consumption. *American Economic Review Papers and Proceedings* **101**(3), 448–453.

Harrington, W. and Portney, P. (1987). Valuing the benefits of health and safety regulation. *Journal of Urban Economics* **22**, 101–112.

Joyce, T., Grossman, M. and Goldman, F. (1989). An assessment of the benefits of air pollution control: The case of infant health. *Journal of Urban Economics* **25**, 32–51.

Lleras-Muney, A. (2010). The needs of the Army: Using compulsory relocation in the military to estimate the effect of environmental pollutants on children's health. *Journal of Human Resources* **35**(3), 549–590.

Mansfield, C., Reed Johnson, F. and Van Houtven, G. (2006). The missing piece: Valuing averting behavior for children's ozone exposures. *Resource and Energy Economics* **28**(3), 215–228.

Moretti, E. and Neidell, M. (2011). Pollution, health, and avoidance behavior: Evidence from the ports of Los Angeles. *Journal of Human Resources* **46**(1), 154–175.

Neidell, M. (2009). Information, avoidance behavior, and health: The effect of ozone on asthma hospitalizations. *Journal of Human Resources* **44**(2), 450–478.

Pope, III, C. A., Schwartz, J. and Ransom, M. (1992). Daily mortality and PM10 pollution in Utah Valley. *Archives of Environmental Health* **47**, 211–217.

Ransom, M. and Pope, III, C. A. (1992). Elementary school absences and PM10 pollution in Utah Valley. *Environmental Research* **58**, 204–219.

Ransom, M. and Pope, III, C. A. (1995). Estimating external health costs of a steel mill. *Contemporary Economic Policy* **13**, 86–97.

Preferred Provider Market

X Martinez-Giralt, Universitat Autònoma de Barcelona and MOVE, Barcelona, Spain

Glossary

Antitrust The legislation and processes in the USA by which a more competitive environment is created through the prohibition of certain practices deemed illegal by antitrust laws, such as price fixing, welfare-reducing mergers, and monopolization.

Bundled discounts A multiproduct seller which has market power over at least one of its products (i.e., the ability to charge an above-cost price for that product) and offers a bundle of several of its products at an above-cost price, attempting to exclude an equally efficient, but less diversified, rival.

Copayment An arrangement whereby an insured person pays a particular percentage of any bills for health services received, the insurer paying the remainder.

Deductible The amount of expenses in an insurance policy that must be paid out-of-pocket before the insurer will pay any expenses.

Downstream competition Capacity of a monopolist to control prices to force competing downstream buyers to sign tying contracts that will lever its monopoly into another market.

Duplicate coverage insurance (double coverage) Situation where an individual contracts two different insurance policies providing coverage for the same events.

First-mover advantage The advantage gained by the initial ('first moving') occupant of a market segment yielding a control of resources that followers may not be able to match.

Idle capacity Production facilities that remain unused because of lack of business. Also, a provider may hold strategic idle capacity to deter entry of a potential competitor by investing in capacity beyond the optimal level.

Indemnity Amount paid by the insurer to the insured by way of compensation for a particular loss. Typically, the insurance policy contemplates a ceiling to this compensation.

Joint ventures Agreement between two (or more) parties to take on a project. The participating parties contribute in money, time, and effort.

Mixed oligopoly Oligopolistic market structure where the objective of at least one firm differs from that of other firms. A particular instance contemplates a public firm aiming at maximizing some notion of social welfare competing with profit maximizing private firms.

Premium competition Insurers' strategic setting of the premium in insurance policies to attract individuals.

Strategic effects Study of the effects of the interrelation among competitors in oligopolistic markets on the decision making and outcome for each participant in the market.

Vertical integration Degree to which a firm owns its upstream suppliers and downstream buyers.

Introduction

In most countries, private health care insurance is provided by managed care organizations (MCOs). They appeared in the late 1990s as an alternative to the traditional fee-for-service health insurance contract. Their main role is to administer and manage the provision of health care services to their clients within a general objective of cost containment in the health care sector. In this sense, an MCO is a middleman contracting with health care providers on the one side and with enrollees on the other. The latter obtain advantageous fees when visiting in-plan providers and the former guarantee a larger base of clients. The most common types of these organizations are preferred provider organizations (PPOs) and health maintenance organizations (HMOs).

An HMO offers health care insurance to individuals as a liaison with providers (hospitals, doctors, etc.) on a prepaid basis. HMOs require members to select a primary care physician, a doctor who acts as a gatekeeper to direct access to specialized medical services whenever the guidelines of the HMO recommend it.

A PPO offers private health insurance to its members (health benefits and medical coverage) from a network of health care providers contracted by the PPO. The main characteristics of a PPO are:

1. health care providers contracted with the PPO are reimbursed on a fee-for-service basis;
2. enrollees in a PPO do not require referral from a primary care physician to access specialized care;
3. enrollees sign a contract defined by a fixed premium, a copayment on the health care services received, and possibly, a deductible;
4. enrollees have freedom to visit out-of-plan providers (with a possible penalty in the form of the payment of a greater share of the provider's fees);
5. drug prescription may be covered as well when enrollees patronize participating pharmacies; and
6. preventive care procedures (check ups, cancer screenings, prenatal care, and other services) may also be available.

To summarize, a PPO is a particular instance of integration between upstream providers and downstream third-party

payers. The aim of this article is to describe how providers compete to become preferred providers.

The PPOs Market Place

Competition in the health care market takes place at different levels. MCOs (and in particular, PPOs) compete to attract enrollees and providers compete for patients and compete to be selected by PPOs. Often, this competition develops in the framework of a regulated market by some public agency aiming at achieving some social welfare goal.

Most of the literature on PPOs deals with the selection of providers, with competition among providers (hospitals and physicians), and with the effects of the design of the insurance contracts on competition. Usually, it is assumed that individuals have already chosen an insurance contract. Some of these individuals may become ill and seek health care services. Sick individuals are referred to as patients. They are the focus of attention of the demand side of the market. Generically, individuals are supposed to make their choices to maximize their level of satisfaction. Focusing the attention on the choice of a PPO, an individual compares on the one hand the premium, co-payment, and deductibles of alternative insurance contracts and on the other hand, the set of in-plan providers. Also, the individual will try to guess how transparent is the information provided by the PPO on medical costs and its negotiation capacity as these are elements characterizing an insurance contract. Also, the individual may consider the plans of the PPOs to enlarge the present set of enrolled providers. All in all, the best plan for an individual reflects the balance between (expected) health care needs, the freedom to choose providers, and his(her) budget constraint. Finally, enrollees should have proper incentives to use the in-plan providers so that the PPOs fulfill their role in the health care system.

In the supply side of the market, one of the reasons for the appearance of MCOs is the need for cost containment in the health care system. As a consequence, managed care has transformed the way hospitals compete for patients and physicians. From competing in quality and provision of services and amenities, managed care introduces the so-called 'selective contracting' of providers. This means that not all available providers in a community are able to contract with the managed care plan. Accordingly, hospitals and physicians compete to be selected as in-plan providers. An issue appears on the size of the PPOs. The negotiation between a provider and a PPO to become in-plan determines the discounts that in turn, are linked to expected utilization. Therefore, the PPO faces a dilemma. Limiting the number of providers in-plan favors achieving the utilization levels, and thus the capacity to offer better deals to enrollees. But a too short list of in-plan providers may discourage individuals to contract with the PPO because it limits the freedom to choose. Empirical evidence seems to point to the prevalence of the obtention of lower prices associated with this selective contracting due to the capacity of the managed care plan to control the number of providers and idle capacity. Also, the size of the managed care plan may be an additional element toward lower prices. The selective contracting mechanism has induced a process of integration between providers and insurers. In this integration, we find upstream health providers deciding first the prices charged to the insurers for a bundle of services, and next insurers deciding the premiums of the (menu of) contracts offered to individuals. The interesting finding is that net revenues, upstream or downstream, result from the combination of a competition effect and a coordination effect. The former reflects the impact of downstream competition on upstream providers; the latter captures the efficiency gains from integration. A PPO, by maintaining some separation between providers and insurers, softens premium competition with respect to the other more integrated structures like HMOs. Accordingly, PPOs emerge as more profitable than HMOs, adding an argument to the popularity of PPOs.

Surprisingly enough, there is very little literature on the process of selecting providers and on competition among providers when different reimbursement rules apply, according to the provider chosen by the patient. Generally, patients have to bear part of the cost of the treatment provided by an in-plan care provider. If an out-of-plan care provider is visited instead, the patient pays the full price and obtains the indemnity from the insurer specified in the insurance contract. It should be clear that the setting of the indemnity associated with the out-of-plan provider is a crucial element in the choice of an insurance contract. Three alternatives can be envisaged, capturing three organizations of the health care systems.

The first one simply does not provide coverage for choices outside the preferred provider set. This is called a pure preferred provider system that captures a pure public system of health provision, such as the Spanish one, where a patient visiting a private provider (instead of a public one) has to bear the full cost of the treatment (unless the patient has some duplicate private insurance). The second alternative, labeled fixed co-payment rule, defines an indemnity equal to what the patient would have obtained had she(he) visited a preferred provider (that is the price of the in-plan provider is used as reference price to determine the indemnity). This alternative captures the idea of indemnity based on a reference price inspired by some features of the French system. Also, it captures some important features of the pharmaceutical sector. Finally, the third alternative, the so-called fixed reimbursement rate rule, considers the same co-payment rate on all providers. It is equivalent to the scenario where all providers have been selected by the insurer. This captures some features of the German system, where, together with the public providers, there is a fringe of private providers regulated through bilateral agreements.

There is also an alternative way to endogenously form the PPO. This is the so-called any willing provider mechanism. Under this approach a third-party payer announces a reimbursement rate and the set of health plan conditions. Any provider finding these acceptable is allowed to join the network.

When providers make simultaneous decisions on prices and qualities, this set-up approaches the primary care sector, whereas when decisions are sequential, first (high-cost, long-run decision on) qualities and then (low-cost, short-run decisions on) prices, the set-up approaches the specialized health care sector. When the market is organized around profit-maximizer providers, the fixed-co-payment rule on the primary health care sector is enough to make providers choose the optimal (welfare-maximizing) price and quality levels. In

contrast, there is no way to attain such an outcome in the specialized health care sector, unless some regulation is introduced. This issue is discussed next.

Alternatively, the mixed public–private provision of health care and the regulation of the market by a public health authority can also be considered. Two scenarios are envisaged. In the first one, an agency regulates both price and quality of the public provider and acts as a Stackelberg leader whereas private providers are followers. It turns out that the first-mover advantage of the public provider coupled with a fixed co-payment rule are sufficient instruments to achieve the first-best allocation. In the second scenario, regulation takes the form of a three-stage game where the regulator sets the level of quality to maximize welfare, then the private providers decide their quality levels, and finally providers compete in prices in a mixed oligopoly fashion. Now, leadership by direct operation of one provider does not ensure achievement of the social optimum, due to the strategic effects resulting from the sequential nature of the decisions. Comparison of these two ways of modeling the role of a public regulator allows to derive some normative conclusions on the implementation of price controls in the health care systems of some European Union member states. All governments have looked at ways to contain health expenditures. Direct and indirect controls over health care providers have been imposed in some countries where co-payments play an important role. In several countries, controls on prices (pharmaceuticals, per-day treatment in hospitals) exist, whereas in others, no such controls exist. Co-payment changes have been frequent in the European countries, mostly limited to the value of the co-payment, whereas maintaining its structure (fixed reimbursement rates). Moreover, co-payments are designed with insurance coverage in mind (typically, they have an upper limit). No role as a market mechanism underlies the choice of the structure and the value of co-payments. Thus, the relative unsuccessful episodes of cost containment through co-payments is not totally surprising. The structure of the co-payment has been kept constant, although the results reported highlight the fact that changing its structure would have a greater impact.

Anticompetitive Scrutiny of PPOs

It has been mentioned in the Section Introduction that a PPO can be seen as the vertical integration between upstream providers and downstream third-party payers. As it is well-known in the economics of regulation, vertical relations may jeopardize market competition. This is so in the private health care market as well.

In the UK, concerns on limits to competition in the provision of private health care, has prompted the Office of Fair Trading (OFT) to its scrutiny. In particular, attention is focused on the level of concentration among providers of private health care, barriers to entry, restrictions on the ability of medical professionals to practice, and consumers' access to providers. The report appeared in December 2011 (the report can be downloaded at http://www.oft.gov.uk/shared_oft/market-studies/OFT1396_Private_healthcare.pdf) led the OFT to undertake a public consultation on its findings which closed in January 2012 (see the associated press release in http://www.oft.gov.uk/news-and-updates/press/2012/26-12). A special report on this study has been published by Health Insurance in January 2011 (issue 158, pp. 14–15, see http://content.yudu.com/A1r4do/HIJan2011/resources/index.htmreferrerUrl=), assessing the viewpoints of private hospitals and doctors, and insurers. They share aims but also face difficult trade-offs. Although insurers agree with hospitals and doctors on the need to guarantee prices and patient flows, they disagree on the patients' capacity to choose among a variety of providers. Finally, hospitals and doctors align with insurers on preventing shortfalls but disagree on the meaning of 'keeping costs at a reasonable level.'

The European Union antitrust authorities are virtually silent on anticompetitive issues around PPOs.

In the USA, the Federal Trade Commission (FTC) acknowledges the changes occurring in the health care market place, so that antitrust enforcement is essential to guarantee the performance of a health care system based on the systems of delivery of health care competing for consumer acceptance. In its report of 2004, the FTC remarks that its activity addresses two basic questions. One refers to the current role of competition in health care and how it can be enhanced to increase consumer welfare. The second one deals with the way antitrust enforcement protects existing and potential competition in health care. Regarding the PPOs, two areas of activity of the FTC are highlighted: bundled discounts and network joint ventures. Bundled discounts refer to the combined sale of two or more products and services at a lower price than the sum of the prices of those goods and services when bought separately. An instance of bundled discount would be proposing discounts to insurers in tertiary services if the insurers made a PPO the sole provider for primary, secondary, and tertiary services. The proper test to prove the existence of bundling requires to show that the price of the bundle is lower than the seller's incremental cost. This is not simple because antitrust laws do not provide clear guidelines.

Physician network joint ventures are under the scrutiny of the FTC because they may reduce and even eliminate competition among the participants in the venture, and they may rise impediments to effective competition among different networks or health plans operating in the same market. Some providers excluded from the network joint ventures have filed complaints of monopolization of the market. However, evidence to support such complaints is difficult to obtain.

A different phenomenon widely studied is hospital mergers. Besides the traditional arguments of enhance efficiency and market power, another motive behind hospital mergers is the improvement of the bargaining position against MCOs. Some empirical evidence suggests that most consolidation of competing hospitals favor price increases in their markets, so that the market power motive seems to offset the efficiency argument.

Silent PPOs

In the early 1990s in the USA, a practice started by which one health plan was selling or renting its provider network (and the network discount rates) to another insurer without the provider's knowledge. This practice has been labeled as 'silent PPO.'

To illustrate, remember first that a physician's reimbursement, when providing out-of-network services, is higher than the in-network fee. Now, consider a physician member of a certain PPO who receives the visit of a patient whose insurance company has no agreement with this physician's practice. Typically, these insurers are organizations without networks of their own. Accordingly, the physician expects to receive the full bill charge from the patient's insurer. The patient's insurer when receiving the bill (and without the knowledge of the patient) assesses whether the physician belongs to a network with a negotiated discount. Then, it enquires about the possibility of the physician's PPO allowing the patient's insurer to use its negotiated discount (this is called the secondary discount market). If so, the physician receives a discounted reimbursement instead of the full payment for the treatment provider, so that even though the patient's insurer does not belong to the PPO, it reimburses the health care services as if the physician would be an in-plan provider. In short, out-of-network services are reimbursed at in-network prices. Figure 1 illustrates the discussion considering a scenario with two PPOs. Physician 1 in PPO1 receives a fee f when treating an in-network patient, and expects a fee $F > f$ when treating an out-of-network patient.

In principle, this need not be an illegal practice. It depends on the terms of the agreement between the insurer and the provider. Such agreements may contain an assignment provision allowing the health plan to offer the contract conditions to anyone willing to pay the agreed fees. Sometimes, there is written consent for the assignment to occur, sometimes there is no express consent or advance knowledge of the provider of such extension of the benefits to other health plans.

Initially, the term 'silent' PPO referred to a PPO where the contract was silent with regard to the commitment of the PPO to direct and encourage its patients to visit the in-plan providers. In the early 2000s however, the term changed to refer to improper and illegal practices. The American Medical Association (AMA), the American Hospital Association (AHA), and the American Association of Preferred Provider Organizations (AAPPO) have taken stance against these practices, and actively tried to encourage legislative actions against them. The problem arises because the providers (physicians and hospitals) realize the situation ex-post, once the treatment has been provided and billed. Accordingly, it falls in the hands of the providers to ensure that only patients enrolled in the PPO receive the discounted fees, whereas outside patients are billed the full price of the treatment received.

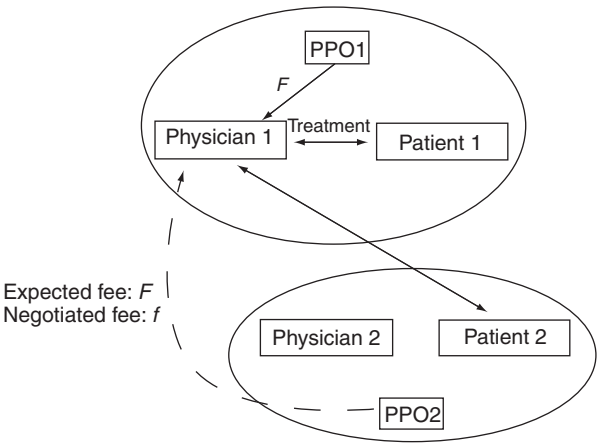


Figure 1 Silent PPOs.

Final Remarks

Preferred provider organizations are the most popular forms of private provision of health care. It balances in the best way the trade-off between insurers and individuals on the one hand, and between insurers and providers on the other. That is, the freedom of individuals to choose providers and the terms of the insurance contract on the one hand, and the base of patients and the discounted fees on the other, with respect to other forms of managed care.

All these agents meet in the market place, where two features are of particular concern. The first one, the structure of PPOs vertically integrating upstream providers and downstream third-party payers, raises the inquiry from the antitrust authorities. The second refers to the degree of transparency of the terms of the contracts between insurers and providers that have promoted legislative initiatives to prevent the better informed party from taking advantage of imperfect information.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance and Health. Health Insurance Systems in Developed Countries, Comparisons of. Health-Insurer Market Power: Theory and Evidence. Managed Care. Private Insurance System Concerns. Supplementary Private Health Insurance in National Health Insurance Systems

Further Reading

Barros, P. P. and Martinez-Giralt, X. (2002). Public and private provision of health care. *Journal of Economics & Management Strategy* **11**(1), 109–133.

Barros, P. P. and Martinez-Giralt, X. (2008). Selecting health care providers: "Any willing provider" vs. negotiation. *European Journal of Political Economy* **24**(2), 402–414.

Boonen, L. H. H. M. and Schut, F. T. (2011). Preferred providers and the credible commitment problem in health insurance: First experiences with the implementation of managed competition in the Dutch health care system. *Health Economics, Policy and Law* **6**, 219–235.

Boonen, L. H. H. M., Schut, F. T., Donkers, B. and Koolman, X. (2009). Which preferred providers are really preferred? Effectiveness of insurers' channeling incentives on pharmacy choice. *International Journal of Health Care Finance and Economics* **9**, 347–366.

Capps, C. S. and Dranove, D. (2004). Hospital consolidation and PPO prices. *Health Affairs* **23**(2), 175–181.

Capps, C. S., Dranove, D., Greenstein, S. and Satterthwaite, M. (2002). Antitrust policy and hospital mergers: Recommendations for a new approach. *The Antitrust Bulletin* **47**(4), 677–714.

Eggleston, K., Norman, G. and Pepall, L. M. (2004). Pricing coordination failures and health care provider integration. Contributions to Economic Analysis & Policy **3**(1), article 20.

Federal Trade Commission (2004). Improving health care: A dose of competition. *A Report by the Federal Trade Commission and the Department of Justice.* Available at: http://www.ftc.gov/reports/healthcare/040723healthcarerpt.pdf (accessed August 2012).

Fuller, D. A. and Scammon, D. L. (2000). Antitrust concerns about evolving vertical relationships in health care. *Journal of Business Research* **48**, 227–232.
Gaynor, M. and Haas-Wilson, D. (1999). Change, consolidation, and competition in health care markets. *Journal of Economic Perspectives* **13**(1), 141–164.
Greany, T. L. (2007). Thirty years of solicitude: Antitrust law and physician cartels. *Houston Journal of Health Law and Policy* **7**(2), 189–226.
Hurley, R. E., Stunk, B. C. and White, J. S. (2004). The puzzling popularity of the PPO. *Health Affairs* **23**, 56–68.
Morrisey, M. A. (2001). Competition in hospital and health insurance markets: A review and research agenda. *Health Services Research* **36**(1), 191–221.
Steadman, K. A. (2006). Silent preferred provider organizations. What is all the noise about? *FORC Journal* **17**, 4 ed., article 3.

Preschool Education Programs

LA Karoly, RAND Corporation, Arlington, VA, USA

Introduction

In recent years, it has become increasingly common for children to be enrolled in preschool education programs for one or more years before the traditional starting age for primary school. According to data from the World Bank, during 2010, 48.3% of preprimary-age children were enrolled in school, a rate that was just 34.1% a decade earlier. Although preprimary enrollment rates in high-income countries far exceed those of low-income ones (82.2% on average vs. 14.9% on average), enrollment rates have been rising since a decade in countries across the development spectrum. Preschool participation rates are not strictly related to the level of economic development, however. Even among high-income countries such as those in the Organization for Economic Cooperation and Development (OECD), there is considerable variation in the share of 3- and 4-year olds enrolled in preschool programs (see Figure 1). Although the preschool enrollment rate during 2009 exceeded 90% in Belgium, Denmark, France, Germany, Iceland, Italy, Norway, Spain, and Sweden, it was not even half that rate in Greece, Ireland, Korea, Switzerland, Turkey, and the US.

The high and rising rates of preschool enrollment across most countries reflect a growing demand among parents for early learning opportunities before the age at which school traditionally begins, as well as enthusiasm on the part of governments across the globe for supporting such programs with public funds. Interest in early childhood investments have been bolstered by research in early childhood development, which has advanced toward a more in-depth understanding of the importance of the early years for lifelong health and development. Emerging evidence from neuroscience, molecular biology, genomics, psychology, and social sciences have converged to provide a new paradigm pointing to the role of both genes and the environment for shaping physical and mental health from the point of conception through to adulthood. Early adversity from nutritional deprivation to insufficient emotional support, or limited cognitive stimulation can trigger the human body's stress management systems in ways that can be protective or harmful, depending on the available supports. These scientific findings serve as a foundation for the theoretical framework put forth by the Nobel Laureate of 2000 in Economics, James Heckman along with his colleagues, which views skill formation as a life-cycle process wherein abilities are both inherited and developed. In this framework, the development of human capital at one stage in life boosts skill attainment at later stages, and early investment improves the productivity of later investments, resulting in a high rate of return to early investment, as skill begets skill.

In the light of these trends and scientific foundations, the goal of this article is to highlight the specific research base that provides support for early education investments. In particular, two strands of research put together strengthen the support on the part of policymakers, practitioners, and parents for preschool education. First, a growing body of evaluation evidence has demonstrated that high-quality early learning programs can boost school readiness and provide long-term benefits in multiple domains. Second, benefit–cost calculations by economists have confirmed that effective programs can more than pay back their costs. Although there are some important caveats and knowledge gaps in each of these research areas, the case for preschool investments rests on a solid research foundation.

Throughout this discussion, it is important to bear in mind that preschool programs defined herein serve children 1 or 2 years before formal primary schooling begins, taking on various forms depending on the country, time period, source of funding, and provider (an even broader array of early childhood intervention models, not being considered here, begin as early as pregnancy and offer services to parents and children in home and center settings in the first 3 years of life, sometimes beyond). Preschool programs may focus on one or more developmental domains including language and cognitive development; behavioral, social, and emotional competencies; and mental and physical health including nutrition. Programs may be delivered in a group setting, such as a child care center or an elementary school; in some countries, home-based providers also offer formal early learning programs. Program intensity can vary, with offerings over 1 year or multiple years that range from part-day programs delivered during the academic year to full-time year-round programming. In addition to programs centered on the child, some programs also engage the child's parents in the learning process through home visits, parenting classes, and other activities. Finally, programs may be fully or partially subsidized by the public sector, with parents or other private sources of support (e.g., employers or philanthropies) filling any gaps.

Evidence from Program Evaluations

For many, the findings from brain research and other developmental science is sufficient justification for providing children – particularly those with disadvantaged backgrounds – with various developmental supports, including formal early learning programs. Yet, preschool programs can take many forms, with considerations for features such as the number of children in a group setting, the ratio of adults to children, the education and training background of the caregivers or teachers, and the choice of an early learning curriculum. With uncertainty over what constitutes an effective program – one that will achieve the goal of supporting children's developmental progress – a body of evaluation research has accumulated to assess the effectiveness of particular program models or specific program features.

The gold standard for evaluation evidence is an experimental design. If outcomes of children who are enrolled in a

 doi:10.1016/B978-0-12-375678-7.00413-2

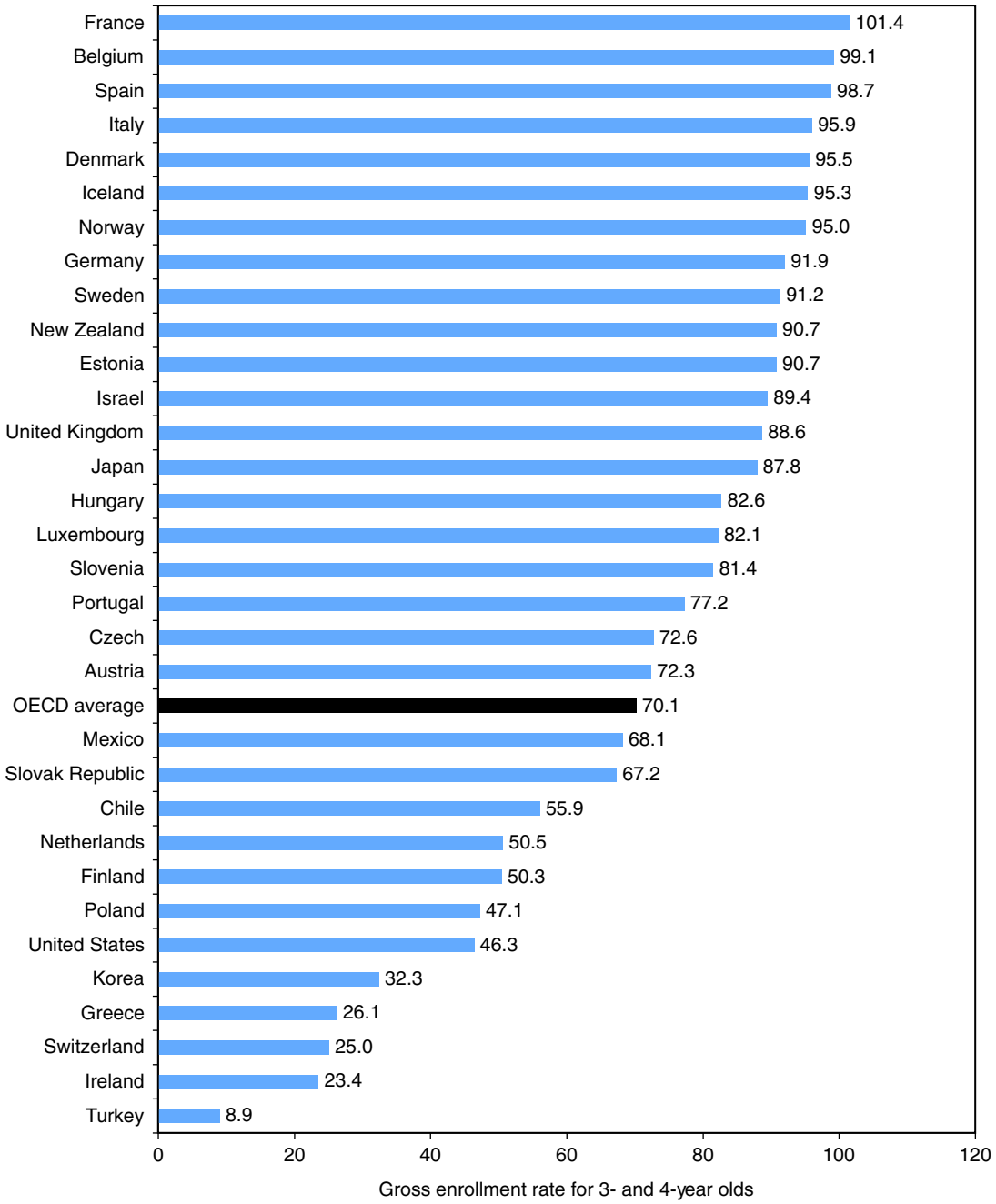


Figure 1 Gross enrollment rates for 3- and 4-year olds in OECD Countries: 2009. Reproduced from OECD (2011). *Health at a Glance 2011: OECD Indicators*. OECD Publishing. http://dx.doi.org/10.1787/health_glance-2011-en. The rate for Canada was not available.

preschool program are compared with those of children not enrolled, the observed differences may arise from the program itself, or from other factors that influence parental choice regarding preschool enrollment as well as child development. This issue of selection bias is avoided with a well-designed experimental study. Random assignment to the treatment group (program participation) versus the control group (no participation) ensures that all other possible determinants of child outcomes are controlled for, so that any observed differences in outcomes between treatment and control children are the result of the intervention.

Of course, experimental studies may be unethical and impractical, or too costly to launch. In such cases, quasi-experimental methods that closely replicate experimental conditions may be feasible. For example, one common method used in the US to evaluate a number of state-funded preschool programs for 4-year olds is a regression discontinuity (RD) design. This approach exploits the strict age-cutoff used to determine when children can enroll in a state preschool program. Effectively, this method uses the accidental birth to compare a child whose birthday occurs just before the enrollment cutoff (and is therefore able to enroll

and experience the preschool treatment) to a child who just misses the birthdate cutoff (and therefore does not experience the program being unable to enroll until next year).

Rigorous Evidence from the USA and Other Countries

In the US, rigorous evaluations of voluntary preschool programs serving 3- and 4-year olds have been conducted for small-scale demonstration programs implemented during the 1960s and targeted to very disadvantaged children, such as the High/Scope Perry Preschool Program and the Early Training Project, and for programs implemented on a large scale and evaluated for more recent cohorts of low-income children such as the Chicago Child–Parent Centers (CPC) program by the Chicago public school system and the federally funded Head Start program. More recently, the RD quasi-experimental method has been used to assess the impact of several state-funded preschool programs targeting children from low-income families as well as Oklahoma's preschool program, which is one of the few programs for 4-year-old children in the US.

Taken together, the evidence from the US evaluation research – synthesized in studies by Barnett, Burchinal, Gormley, Pianta, Shonkoff, and others – has demonstrated that well-designed preschool programs can improve child developmental outcomes on both short- and long-term bases. Based on an assessment of the literature from the US by Karoly and colleagues, Table 1 lists the specific outcomes that have been significantly affected by at least one rigorous (i.e., experimental or well-designed quasi-experimental) evaluation of a preschool program serving children for 1 or 2 years before kindergarten entry. The bulk of the evidence base is in the first tier of Table 1, as all preschool evaluations measure some early childhood outcomes. Most evaluations, for example, find significant favorable effects on one or more developmental measures that are assessed during the program or soon after. Such outcomes in early childhood include measures of general intelligence or intelligence quotient (IQ), assessments of school readiness such as specific prereading and premath skills, as well as gains in socioemotional or behavioral competencies. Meta-analyses across multiple studies being conducted in the US tend to show larger average treatment effect sizes for impacts on cognitive outcomes – in the range of 0.2–0.3, when compared with impacts on socioemotional or behavioral outcomes. Larger impact estimates have been found for specific studies and these indicate that intentional design can boost program impacts beyond those measured for typical programs. The magnitudes found for the most effective programs are large enough to close half or more of the achievement gap between disadvantaged children and their more advantaged peers.

Table 1 Outcomes with favorable impacts from preschool interventions

Lifecourse stage	*Specific outcomes for participating child*
Outcomes in early childhood	IQ behavior, social competence developmental milestones, general health status, immunization abuse and neglect, school readiness
Outcomes in school-age years	Achievement tests, grade repetition, special education use, grades high school completion, college attendance, teen pregnancy
Outcomes in adulthood	Employment earnings, use of social welfare programs, criminal activity, use of tobacco, alcohol, and drugs

A more limited evidence base supports the second and third tiers in Table 1, with confirmation of the benefits from preschool programs continuing into the school-age years and persisting beyond. Studies with follow-up intervals, pertaining to elementary school and later grades – such as the Perry Preschool and Chicago CPC evaluations – demonstrate higher achievement scores, reduced rates of grade repetition and special education use, and higher high school graduation rates. These two studies, with continued follow-up into adulthood (age 40 for Perry Preschool and age 26 for Chicago CPC), also find favorable impacts in other domains like employment and earnings, social welfare program use, criminal activity, and health behaviors. Many of these same impacts are found for other targeted intensive early intervention programs such as the Carolina Abecedarian Project, which provides full-time year-round center-based educational programming till kindergarten entry, starting just few weeks after birth.

In many European countries, where preschool participation rates are almost universal, fewer experimental or quasi-experimental studies have been conducted. Longitudinal studies such as the 1958 British Cohort Study and the 1997 Effective Provision of Preschool Education Project in England provide observational evidence of the favorable effects of preschool participation on child development both in the short and long runs. In developing countries such as Argentina, Jamaica, Mauritius, the Philippines, Turkey, Uganda, Uruguay, and Vietnam, a number of experimental and quasi-experimental evaluations of preschool programs have been implemented, often in combination with other services pertaining to child health and nutrition. Recent syntheses or formal meta-analyses of the evaluations of non-US programs, conducted by Nores and Barnett and by Vegas and Santibañez, among others, document that the favorable impacts found in US studies across multiple developmental domains are replicated throughout the range of low-income to high-income countries. At the same time, there is also considerable variation in the magnitude of the impacts across program models. Such differences may be attributable to program design or the populations served.

Limitations of the Knowledge Base

Although the cumulative evidence from rigorous evaluation studies is very compelling, it is important to recognize the limitations of the research to date. First, much of the evidence of preschool program benefits have stemmed from studies of programs targeting disadvantaged children. This is especially true for the kind of research in the US where both small-scale demonstration programs and large-scale models like Chicago CPC or Head Start serve children in families with limited

resources or other disadvantages. One exception for the US is the evaluation of Oklahoma's universal preschool program, which has demonstrated significant favorable impacts across income and race-ethnic groups, although the evidence suggests that the benefits from preschool participation are greater for the more disadvantaged children. Nevertheless, even children from families with income above poverty are likely to face various stressors that can comprise their healthy development and readiness for school. The Oklahoma results suggest that the benefits from participation in a high-quality preschool program may be broadly shared.

Second, evaluations of specific preschool programs demonstrate proof of the principle that high-quality early learning programs can improve child developmental outcomes on both short- and long-term bases. However, such evidence does not confirm that every program will necessarily have favorable impacts or effects that are equal in magnitude to those found in specific evaluations. In most cases, existing evaluations quantify the impact of specific combinations of preschool inputs as a bundle: group size, staff–child ratio, curriculum, teacher education and training, teacher–child interactions, total hours spent in the program, and so on. The evaluation can neither tease apart the effect of specific inputs nor identify the impacts that would result with some different combination of factors. Essentially, the research to date largely treats each evaluated preschool program as a 'black box' sans the ability to identify which program factors account for the measured impacts. Implementing programs with different combinations of inputs or with different levels of intensity may well produce different impacts, but how much different remains largely unknown. Given these limitations, ongoing research is seeking to understand issues such as whether there are minimum levels of quality required for programs to be effective and how program impacts vary with program dosage (e.g., annual hours in the program).

Third, as participation in some form of early care and education has become more common over time, especially in high-income countries, it has become challenging to measure the impact of preschool program participation against a true 'no program' control group. For example, in the national Head Start experimental evaluation conducted during the early 2000s in the US by Westat and others for the US Department of Health and Human Services, 48% of the 4-year olds in the control group participated in some form of center-based program and 13% attended some other Head Start program than the one they were randomized out of. Thus, the experimental evaluation had measured the effect of participation in Head Start against the status quo where nearly half of 4-year-old children living in poverty were already in some form of early education program. In contrast, when programs like Perry Preschool were evaluated in the 1960s, US children in the control group did not have access to other early learning programs, and were thus in a control condition of parental care only. In other countries with very high preschool enrollment rates, the opportunities for measuring program impacts against a 'no program' control group are very limited. For this reason, ongoing research is centered on assessing the impact on child developmental outcomes from different preschool program designs so as to identify which program models are most effective, rather than trying to assess whether programs have an impact when compared with the alternative of no program participation.

Economic Case for Preschool Investments

In an era of result-based accountability, there has been a growing interest in the US and other countries in demonstrating that the investments in public sector programs generate a favorable economic return to the public sector or at least to society as a whole. This has prompted increased application of benefit–cost analysis (BCA) methods to social policies, including early childhood programs.

The BCA methodology requires (1) a comprehensive measure of program costs relative to the baseline without the program; (2) evidence of the causal impact of the program relative to the same baseline; and (3) the ability to value all the program impacts in a common monetary value, often called 'shadow prices.' The method then compares the present value of the stream of program costs with that of the stream of lifetime program impacts (whether favorable or unfavorable) to determine whether net benefits are greater than zero or alternatively the ratio of benefits to costs is greater than one. This accounting of benefits and costs can be conducted from the perspective of different stakeholders. Most common is to calculate the economic returns for the public sector – accounting for the costs or benefits to taxpayers of a given intervention. The most comprehensive perspective is for the society as a whole, accounting for benefits and costs that accrue to the public sector as well as private benefits and costs that accrue to program participants and non-participants.

The Application of Benefit–Cost Analysis to Preschool Programs

For both preschool programs and other early childhood interventions, one of the challenges in applying BCA is that many of the outcomes affected by these programs are neither easily valued in dollars nor in some other monetary unit. The early childhood outcomes listed in the first tier in **Table 1** including those related to child health are ones where ready shadow prices do not exist for the most part. Consequently, BCA has not been employed for many preschool interventions. When the tool is used, the lack of ready shadow prices means that the economic returns tend to be understated because benefits are undercounted relative to costs.

Given the challenge of placing an economic value on early childhood outcomes, the application of BCA to preschool programs has mostly been limited to those programs with long-term follow-up into the school-age years and beyond, because more of such outcomes can be valued. For example, if a preschool program leads to reductions in the use of special education services, then that represents a savings to the public sector because children will be enrolled in regular education classes instead of the more costly special education programs. Likewise, if a preschool program boosts high school graduation rates, then that can lead to an increase in lifetime earnings for the children when they reach adulthood, and those higher earnings can generate more tax revenue to the

government. Achieving higher educational attainment is also likely to reduce the use of social welfare programs and lower the incidence of crime. Therefore, benefits in these areas may be projected based on any measured educational gains; or these outcomes may be directly observed as they have been in the long-term follow-ups of the Perry Preschool and Chicago CPC evaluations.

As two of the preschool programs with long-term follow-up, the Perry Preschool and Chicago CPC programs have been the focus of a series of BCAs. As shown in Table 2, Perry Preschool has been the subject of at least five different BCAs, three are conducted by the High/Scope evaluation team using updated information at each adult follow-up stage (studies (a), (b), and (c), using follow-up data at ages 19, 27, and 40, respectively) and two are conducted by other research teams using somewhat different methods (studies (d) and (e) based on follow-up data at ages 27 and 40, respectively). The Chicago CPC program has been the subject of two BCAs by the evaluation team using follow-up data at ages 21 and 26 (studies (f) and (g), respectively). Table 2 also shows the benefit–cost ratio for a study by the Washington State Institute of Public Policy, which is based on a meta-analysis of early childhood education programs serving 3- and 4-year-old children in low-income families (study (h)). Thus, rather than a BCA for a specific preschool program, this analysis represents the likely return for high-quality preschool programs on average when implemented at scale.

These results demonstrate several patterns. First, in all cases, the analyses show that the programs generate positive economic benefits with benefit–cost ratios ranging from $2.36–1 to $16.14–1. As Heckman notes, the findings for these programs and other early childhood interventions demonstrate that early childhood investments offer a rare policy option that can promote both economic efficiency, as well as fairness and social justice. Second, as more follow-up data becomes available, the calculated economic returns tend to increase. This is because, the methods of both the Perry Preschool and CPC BCAs to project future benefits are typically too conservative when compared with the actual experiences from later follow-ups. Third, methodological choices matter in the calculated returns. For example, unlike the evaluation teams, the independent estimates for the Perry Preschool program (studies (d) and (e)) made different choices like excluding the value of intangible crime victim costs or using different values for the cost of crime. As the available estimates of crime costs vary widely, these choices can have a considerable impact on the estimated returns. Fourth, the estimates of returns for specific programs will not necessarily generalize those for more generic preschool programs. The lowest benefit–cost ratio in Table 2 is for the more generalized targeted preschool program where the estimated program impacts are assumed to be attenuated because of program scale-up. Thus, this estimate may be closer to what the typical 'real world' program would generate.

Limits on the Generalizability of Existing Economic Analyses

The issue of the generalizability of the findings in Table 2 for either the small-scale Perry Preschool program or the large-scale Chicago CPC program is an important one. Both programs have been designed to serve particularly disadvantaged groups of children, and the estimated program impacts and the associated benefit–cost ratios thereof may not be replicated in other programs or for other populations of children. Programs that are less effective due to lower quality would not be expected to generate the same economic return. Likewise, programs serving more advantaged groups of children would not necessarily have impacts across the same range of outcomes or of the same magnitude. To the extent that impacts are attenuated when quality declines or the population served varies, the economic returns would be lowered accordingly. However, the attenuation of program impacts when programs serve a broader base of children does not necessarily mean that the economic returns will no longer be positive. For example, several studies have estimated the returns to universal preschool programs while assuming that the favorable impacts

Table 2 Reported benefit–cost ratios for preschool programs in the USA

Program/program type	*Source*	*Benefit–cost ratio*
Estimates for specific programs		
(a) Perry Preschool – Age 19 follow-up	Berrueta-Clement *et al.* (1984)	3.56
(b) Perry Preschool – Age 27 follow-up	Barnett (1993, 1996), Schweinhart *et al.* (1993)	8.74[a]
(c) Perry Preschool – Age 40 follow-up	Barnett *et al.* (2005), Nores *et al.* (2005), Belfield *et al.* (2006)	16.14[a]
(d) Perry Preschool – Age 27 follow-up	Karoly *et al.* (1998)	4.11[b]
(e) Perry Preschool – Age 40 follow-up	Heckman *et al.* (2010)	7.1–12.2[ac]
(f) Chicago CPC – Age 21 follow-up	Reynolds *et al.* (2002)	7.14
(g) Chicago CPC – Age 26 follow-up	Reynolds *et al.* (2011)	10.83[a]
Estimates from meta-analysis		
(h) Early childhood education for low-income 3- and 4 year olds	Aos *et al.* (2004)	2.36[a]

[a]Includes value of reduced intangible crime victim costs.
[b]Discount rate is 4%.
[c]Reported range of estimates under alternative assumptions regarding the economic cost of crime.
Source: Adapted from Karoly, L. A. (2012). Toward standardization of benefit–cost analyses of early childhood interventions. *Journal of Benefit–Cost Analysis* **3**(1), Article 4.
Note: The benefit–cost ratios are the ratio of the present discounted value of total benefits to society as a whole (participants and the rest of society) divided by present discounted value of program costs. The discount rate is 3% unless otherwise noted. The value of reducing intangible crime victim costs are excluded unless otherwise noted.

are concentrated among the most disadvantaged children. Even so, the total economic benefits to society as a whole can remain positive, so long as the returns for the disadvantaged groups are sufficiently large.

The economic returns to preschool programs implemented in other countries is also an issue that remains largely unexplored, and the magnitudes of the economic returns demonstrated for programs in the US may not be relevant for other developed or developing countries. Economic returns may be higher or lower depending on how the magnitude of program impacts can vary in other countries and also depending on the variation in the economic values attached to the realized program impacts. For example, the high economic returns to Perry Preschool stem, in part, from the high cost of juvenile and adult crime in the US, costs that are likely to be lower in most other countries in the world. For this reason, rigorous program evaluations in all countries need to be accompanied by careful estimates of program costs and benefits. In addition, in the absence of long-term follow-up data from interventions in both high- and low-income countries, there is a need to develop estimates of the economic value associated with changes in child health and development in the early years.

See also: Economic Evaluation of Public Health Interventions: Methodological Challenges. Education and Health. Education and Health in Developing Economies. Education and Health: Disentangling Causal Relationships from Associations. Nutrition, Health, and Economic Performance. Pay for Prevention. Public Health: Overview

References

Aos, S., Lieb, R., Mayfield, J., Miller, M. and Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth*. Olympia: Washington State Institute for Public Policy.

Barnett, W. S. (1993). Benefit-cost analysis of preschool education: Findings from a 25-year follow-up. *American Journal of Orthopsychiatry* **63**(4), 500–508.

Barnett, W. S. (1996). *Lives in the balance: Age-27 Benefit-cost analysis of the high/scope perry preschool program*. Monographs of the High/Scope Educational Research Foundation, 11. Ypsilanti, MI: High/Scope Press.

Barnett, W. S., Belfield, C. R. and Nores, M. (2005). Lifetime cost-benefit analysis. In Schweinhart, L. J., Montie, J., Xiang, Z., et al. (eds.) *Lifetime effects: The high/scope perry preschool study through age 40*, pp. 130–157. Monographs of the High/Scope Educational Research Foundation, 14. Ypsilanti, MI: High/Scope Press.

Belfield, C. R., Nores, M., Barnett, W. S. and Schweinhart, L. (2006). The high/scope perry preschool program: Cost-benefit analysis using data from the age-40 follow-up. *Journal of Human Resources* **41**(1), 162–190.

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S. and Weikart, D. P. (1984). *Changed lives: The effects of the perry preschool program on youths through age 19*. Monographs of the High/Scope Educational Research Foundation, No. 8. Ypsilanti, MI: High/Scope Press.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A. and Yavitz, A. (2010). The rate of return to the high scope perry preschool program. *Journal of Public Economics* **94**(1–2), 114–128.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., et al. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions*. Santa Monica, CA: The RAND Corporation. MR-898.

Nores, M., Belfield, C. R., Barnett, W. S. and Schweinhart, L. (2005). Updating the economic impacts of the high/scope perry preschool program. *Educational Evaluation and Policy Analysis* **27**(3), 245–261.

Reynolds, A. J., Temple, J. A., Robertson, D. L. and Mann, E. A. (2002). Age 21 cost-benefit analysis of the title I Chicago child-parent centers. *Educational Evaluation and Policy Analysis* **24**(4), 267–303.

Reynolds, A. J., Temple, J. A., White, B. A., Ou, S.-R. and Robertson, D. L. (2011). Age-26 cost-benefit analysis of the child-parent center early education program. *Child Development* **82**(1), 379–404.

Schweinhart, L. J., Barnes, H. V. and Weikart, D. P. (1993). *Significant benefits: The high/scope perry preschool study through age 27*. Monographs of the High/Scope Educational Research Foundation, 10. Ypsilanti, MI: High/Scope Press.

Further Reading

Burchinal, M., Vandergrift, N., Pianta, R. and Mashburn, A. (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly* **25**, 166–176.

Camilli, G., Vargas, S., Ryan, S. and Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record* **112**, 579–620.

Cunha, F., Heckman, J. J., Lochner, L. J. and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. In Hanushek, E. A. and Welch, F. (eds.) *Handbook of the economics of education*. pp. 697–812. Amsterdam: North-Holland.

Gormley, W. T. (2007). Early childhood care and education: Lessons and puzzles. *Journal of Policy Analysis and Management* **26**, 633–671.

Karoly, L. A. and Bigelow, J. H. (2005). *The economics of investing in universal preschool education in California*. Santa Monica, CA: RAND Corporation.

Karoly, L. A., Kilburn, M. R. and Cannon, J. S. (2005). *Early childhood interventions: Proven results, future promise*. Santa Monica, CA: RAND Corporation.

Knudsen, E. I., Heckman, J. J., Cameron, J. L. and Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences* **103**, 10155–10162.

National Scientific Council on the Developing Child and the National Forum on Early Childhood Policy and Programs (2010). *The foundations of lifelong health are built in early childhood*. Cambridge, MA: Center on the Developing Child, Harvard University.

Nores, M. and Barnett, W. S. (2010). Benefits of early childhood interventions across the world: (Under) Investing in the very young. *Economics of Education Review* **29**, 271–282.

Pianta, R. C., Barnett, W. S., Burchinal, M. and Thornburg, K. R. (2009). The effects of preschool education: What we know – how public policy is or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest* **10**, 49–88.

Shonkoff, J. P. and Phillips, D. A. (eds.) (2000). *From neurons to neighborhoods: The science of early child development*. Washington, DC: National Academy Press.

Vegas, E. and Santibañez, L. (2009). *The promise of early childhood development in Latin America and the Caribbean. Issues and policy options*. New York, NY: The World Bank and Palgrave MacMillan.

Relevant Websites

http://developingchild.harvard.edu/
Center on the Developing Child at Harvard University.

http://www.heckmanequation.org
Heckman Equation.

http://nieer.org
National Institute for Early Education Research.

http://www.oecd.org/edu/preschoolandschool/
OECD Directorate for Education.

http://go.worldbank.org/Q0DFS2VJ40
World Bank Early Childhood Development.

Prescription Drug Cost Sharing, Effects of

JA Doshi, University of Pennsylvania, Philadelphia, PA, USA

Introduction

Pharmaceuticals play an increasingly important role in health care. New discoveries of prescription drugs and biologics have transformed the medical management of several diseases. Pharmaceutical expenditures have also dramatically grown over time in part due to increases in the quantity demanded and the high prices of these new medications. An ongoing challenge for third-party payers in countries with public or private insurance coverage for prescription drugs, has been how best to balance the competing goals of affording access and financial risk protection versus avoiding overconsumption of prescription drugs and controlling costs. To contain pharmaceutical spending, various cost-containment policies have been implemented by payers across countries. One such policy, cost sharing between patients and insurers, is common in many developed countries, particularly the US. Cost sharing, as referred to in this article, is the portion of the drug's cost to be paid for by the patient at point-of-service, although the insurer (or payer) covers the remainder of the cost. Other terms used for cost sharing in some countries include user fees, user charges, consumer charges, or prescription drug charges.

The goals of this article are to present an economic framework for the effects of prescription cost sharing, provide an overview of the different forms of prescription cost sharing, and summarize the empirical evidence related to their effects on prescription and medical service use, expenditures, and outcomes.

Economic Framework

Insurance involves a tradeoff between the gains from risk protection and the losses from the incentives to use more medical care caused by the presence of insurance. Patients who have no insurance coverage for prescription drugs, as is the case in many developing countries, face full price (i.e., 100% cost sharing) for all drugs, and are fully at risk for any and all drug expenses. However, patients with full insurance coverage for prescription drugs, pay no costs (i.e., 0% cost sharing) and this may raise potential for moral hazard i.e., additional insurance-induced utilization of drugs. To mitigate the potential problem of moral hazard, insurers in most countries with private or public drug coverage require some cost sharing for prescription drugs, i.e., patients pay some portion of the costs associated with the drugs used as opposed to receiving full coverage. As pharmaceutical expenditures have continued to rise over time, insurers across many countries have increased the levels of prescription cost sharing and/or experimented with alternative forms of cost sharing to control pharmaceutical spending. In many countries, however, the increase in the drug cost sharing amounts has more likely been related to the payer's budgetary burden or deficit problems rather than the recognition of or change in moral hazard of the covered population.

Prescription drug cost sharing may reduce payer drug expenditures through several mechanisms. First and most directly, an increase in patient drug cost sharing reduces payer drug spending by shifting responsibility of a portion of the drug costs from the payer to the patient. Hence, even if drug demand was perfectly inelastic cost sharing would still result in reductions in payer drug expenditures.

The second mechanism through which prescription cost sharing may reduce drug expenditures is through reductions in the price of prescription drugs, either directly by causing pharmaceutical manufacturers to lower drug prices or offer rebates/discounts (as is the case with reference pricing or tiered copayments discussed in the next section) or indirectly via encouraging patients to shift to lower priced (perfect or imperfect) drug substitutes. Examples of the latter include switches from brand name to generic drugs, nonpreferred to preferred brand drugs, and drugs above the reference price to those at the reference price.

Third, assuming that demand for drugs is not inelastic additional reductions in payer expenditures may be obtained by reductions in the overall quantity of drugs demanded with an increase in patient cost sharing. This decrease in quantity, however, can result from either a reduction in the probability of drug use among potential new users (i.e., noninitiation of newly prescribed medications) and/or a reduction in the number of prescriptions refilled among current users (i.e., nonadherence to or discontinuation of ongoing medications). The question then arises with regard to what types of drugs do the patients forego or delay taking and what other services they might substitute for the foregone drugs.

There are several ways that prescription drug cost sharing may affect nondrug health care expenditures. Firstly, certain medical services are direct complements to drugs (as, e.g., physician visits or laboratory tests) and their quantity demanded may also decrease with an increase in prescription cost sharing. For instance, physician office visits are an example of a complement given that the disease diagnosis is made and a prescription is written during a physician visit. Furthermore, certain newly prescribed medications or those prescribed for chronic illnesses require ongoing laboratory monitoring and follow-up visits with the physicians. Hence, if an increase in prescription drug cost sharing reduces the quantity of drugs demanded then it may also reduce the number of physician visits and lab tests required and associated health care expenditures.

However, for many conditions, particularly chronic illnesses, medical treatments may be substitutes for drug treatment in the short term (as, e.g., surgery rather than drug treatment for gastroesophageal reflux disease) or longer term (as, e.g., inpatient and outpatient care associated with coronary artery bypass surgery after a heart attack or kidney dialysis in diabetes patients which could have been prevented by use

 doi:10.1016/B978-0-12-375678-7.01211-6

of long term drug treatment). This substitutability between drugs and other medical services is often the basis of one of the arguments raised against policies intended to increase prescription cost sharing.

Economic theory suggests that when patient cost sharing rises, rational patients will reduce the consumption of drugs whose costs exceed the marginal benefits and continue to consume drugs whose marginal benefit exceeds the costs to the patient. The primary assumption underlying this is that patients are fully informed of the marginal benefits of the drugs to make such evaluations and rational choices. However, critics argue that patients (and providers) often have imperfect information about the marginal benefits from drugs and other treatments, and hence may forego, delay, or decrease adherence to beneficial drugs, which in turn may adversely affect health outcomes and lead to increases in use of other medical services. If the increases in the spending due to the resultant medical service use offset the savings in prescription drug expenditures, then the prescription cost sharing policy will fail in its intended goal of cost containment and improving efficiency.

These considerations have implications for both the design and evaluation of prescription cost sharing by private and public insurers who provide coverage for both drugs and medical services. To the extent that medical services are complements or substitutes for drugs then the drug demand elasticity alone may not determine optimal cost sharing levels; both the own price elasticity of demand and cross-price elasticities of demand need to be factored in (Goldman and Philipson, 2007). For example, even though the own price elasticity of drug demand is high, the optimal cost sharing for the drug may be lower (or even zero) if hospitalizations and emergency room visits are substitutes for the drug, as lower drug cost sharing itself will result in a lower use of these other more expensive services. Similarly, evaluations of the effects of prescription cost sharing policies should not take a siloed view of the impact of cost sharing on only pharmaceutical use and spending but should also examine the effects on patient health outcomes, other medical service use and spending, and more importantly the cost offsets and net effects on total spending.

Types of Cost Sharing

Copayments, Coinsurance, and Tiered Formularies

The two most common types of prescription cost sharing used by payers across countries are copayments and coinsurance. Copayment is a fixed monetary amount per prescription to be paid by the patient (e.g., $10), and coinsurance is a fixed percentage of the total cost of the prescription to be paid by the patient (e.g., 20%). Copayments provide better financial risk protection than coinsurance, because they better cover higher-priced drugs. Under coinsurance patient out-of-pocket payments vary directly with the price of the drugs; hence, it provides stronger incentives for the patient to choose lower priced drugs than copayments.

Payers in several ex-US countries in Europe and Australasia typically require modest drug cost sharing with several protections such as an annual out-of-pocket maximum or stop-loss, sometimes related to income and exemptions for vulnerable groups (e.g., low-income, elderly, disabled, children, pregnant women). Unlike US payers, some of these countries also impose the same level of cost sharing (copayment or coinsurance) for all drugs, regardless of the cost or type of medication This is partly explained by the fact that payers in these countries rely less on patient cost sharing and more heavily on other forms of supply-side cost containment (i.e., reimbursement and price controls). By contrast, most payers in the US use patient cost sharing as a primary tool to control costs and hence have moved away from requiring such flat cost sharing levels. Instead, they have adopted tiered formularies which require differential levels (or tiers) of cost sharing for different types of drugs based on factors such as the cost of the drug to the insurer, relative to the cost of close alternatives. Most private plans in the US have adopted three-tiered formularies that require the lowest cost sharing for Tier 1 drugs (typically generics), a second and higher level of cost sharing for Tier 2 drugs (typically preferred brand-name drugs), and a third and highest level of cost sharing for Tier 3 drugs (typically nonpreferred brand name drugs). Private plans are now increasingly imposing a fourth tier requiring even higher cost sharing (often coinsurance rates) for very expensive specialty biologic medications (see Section Specialty Tier Cost Sharing for more detail). Tiered or differential cost sharing creates financial incentives for patients to use less expensive drug substitutes such as generics and preferred brands. Insurers are also often able to negotiate discounts and/or rebates with brand-name drug manufacturers in exchange for placement on the preferred brand tier, as opposed to the nonpreferred tier.

Reference Pricing Surcharges

Reference pricing leads to another form of cost sharing wherein the incentives are set up for patients to use less expensive drug substitutes. The insurer sets the maximum ('reference') price it will cover (based on prices of low-cost benchmark drugs) for a group of therapeutically similar drugs that are deemed to be close substitutes for one another in the treatment of specific diseases. Patients who want to use a higher-priced drug must pay out-of-pocket for the entire difference between the retail price of that drug and the reference price covered by the insurer. Reference pricing is commonly used in European countries (e.g., Germany and Netherlands), Canada (e.g., British Columbia), and New Zealand. However, its use is rare in the US.

Deductibles

A deductible is the expense that must be paid 100% out of pocket by patients before an insurer will start paying for any drug expenses. They are generally used in combination with other forms of cost sharing such as copayments or coinsurance. Except for certain private plans in the US Medicare Part D program (which offers voluntary prescription insurance to the elderly and disabled), most other payers in the US typically do not subject medications to deductibles. Drug deductibles are also rarely used in European countries,

except for Denmark and Sweden. Some Canadian provinces such as Manitoba and British Columbia also use drug deductibles, the levels of which vary based on patient income.

Benefit Caps

Benefit caps include cost sharing features which require 100% out-of-pocket payments for all prescriptions filled either after the patient's total drug spending reaches a predefined amount (i.e., annual benefit caps or coverage gap) or after a patient fills a predefined number of prescriptions in a month (i.e., prescription limits). These somewhat extreme forms of cost sharing are only seen in the US For example, the standard benefit design under Medicare Part D implemented in 2006 included a coverage gap (i.e., 'doughnut hole') wherein beneficiaries were required to pay 100% of drug costs after total drug spending exceeded a predefined amount ($2250 in 2006) until they reached the catastrophic cap amount ($5100 in 2006) and were then required to pay only 5% of all drug costs. Before the availability of Medicare Part D, drug benefits available through private Medicare managed care plans typically included such annual benefit caps (but without catastrophic coverage). Similarly, some US state Medicaid programs that cover low-income Americans impose a limit on the number of prescriptions covered per patient in each month. Such policies may reduce prescription drug spending by shifting the entire financial responsibility for drugs to the patient after the benefit cap is reached as well as reducing overall drug use if patients try to avoid exceeding the benefit cap. Heavy medication users such as the elderly and the chronically ill are most likely to face the brunt of these policies.

Specialty Tier Cost Sharing

Specialty drugs include high-cost self-injectables, infusions, and certain inhalations and oral agents; they are typically biologically derived and/or require cold-chain distribution. As compared to the 'traditional' chemically synthesized medications (i.e., small-molecules), specialty biologic medications (i.e., large-molecules) are manufactured in more complex, highly involved processes and often represent significant advances in treatments for complex, chronic conditions such as cancer, rheumatoid arthritis, and multiple sclerosis. At the same time they are priced 10 to 20 times higher than traditional drugs which means patients needing such medications are at financial risk for high out-of-pocket expenditures in the absence of insurance coverage. Initially, US insurers covered specialty biologics in a manner similar to other traditional pharmaceuticals, particularly given that a very small proportion of members were users of these medications. However, as the specialty biologic market has expanded significantly over the past two decades and spending growth for these medications has far outpaced the growth for traditional drugs insurers have begun to alter the coverage structure for specialty biologics.

In the US, insurers have created additional, higher cost sharing tiers, often called specialty tiers, on which these drugs are placed. The intent of these tiers is to isolate specialty drugs from lower cost small-molecule entities and to assign specific cost sharing and utilization management tools to these types of medications. Although even the standard three tiers (generics, preferred brands, and nonpreferred brands) require a copayment, the specialty tier often carries a percent coinsurance as opposed to a copayment. This means that beneficiaries are responsible for a far greater proportion of the costs of these expensive medications than they would if the insurer had imposed a copayment. The creation and utilization of specialty tiers was originally most noticeable in the US Medicare Part D program, in which drug plans were allowed to place drugs with monthly costs exceeding a certain threshold (e.g., $600 in 2012) on a specialty tier and require as high as a 33% coinsurance. Today, virtually all Part D plans that use tiered cost sharing structures have created a specialty tier. Although specialty tiers are less commonly used by US employers or private insurers, their adoption has been increasing rapidly in recent years especially as new specialty medications for more common conditions such as osteoporosis and rheumatoid arthritis have begun to enter the market.

Empirical Evidence on the Effects of Drug Cost Sharing

This section summarizes the empirical evidence on the effects of drug cost sharing for traditional drugs and specialty biologics.

Effects of Cost Sharing for Traditional Drugs

A substantial number of studies examining the effects of cost sharing on traditional small-molecule medications have been published in the literature. Numerous reviews with varying scopes and objectives have attempted to qualitatively summarize these studies which are highly heterogeneous in terms of the study design, study population, study outcomes and disease group/drug classes examined, type of data source, and statistical methods; hence, often resulting in mixed findings (Gibson *et al.*, 2005; Goldman *et al.*, 2007; Gemmill *et al.*, 2008; Austvoll-Dahlgren *et al.*, 2008; Eaddy *et al.*, 2012). The vast majority of these studies have been conducted in the US followed by Canada. A few studies have examined the impact of cost sharing increases in Europe (UK, Sweden, Netherlands, Italy, Germany, and Denmark), Australasia (Australia and New Zealand), and Asia and the Middle East (Taiwan, Nepal, and Israel).

Except for the RAND Health Insurance Experiment (HIE) which was a randomized trial in the US (Newhouse, 1994), all the remaining studies are observational in nature. Unfortunately, the coinsurance rate for drugs was made identical to and varied at the same time as the coinsurance rate for other medical services in the RAND HIE; hence, it does not provide an unbiased estimate of the demand elasticity for drugs alone. It is important to note that the methodology used in many of the remaining observational studies has also been reported to be of low to moderate quality (Austvoll-Dahlgren *et al.*, 2008; Soumerai *et al.*, 1993). In particular, several studies are cross-sectional in design whereas others have evaluated outcomes

before and after the cost sharing change without including a control group to account for biases arising due to contemporaneous trends. The latter was often the case in studies from ex-US countries wherein the government increased cost sharing for the entire population (Goldman *et al.*, 2007; Austvoll-Dahlgren *et al.*, 2008).

Prescription drug use and expenditures

Majority of the evidence indicates that higher levels of drug cost sharing are associated with lower total prescription expenditures (Gibson *et al.*, 2005; Goldman *et al.*, 2007; Gemmill *et al.*, 2008). There is also consistent evidence to indicate that the payer's prescription expenditures decrease and not surprisingly, patient out-of-pocket expenditures increase. As noted earlier, the magnitude of the reductions in drug expenditures varies based on factors such as the type and magnitude of change in cost sharing, therapeutic drug class being studied, and the type of study population subject to cost sharing. Some studies report price elasticity estimates reflecting the percentage change in drug spending that would be associated with a 1% increase in cost sharing. If studies that lacked a control group or had very small changes in cost sharing are excluded, price elasticity estimates have been reported to range from -0.2 to -0.6 (that is, 10% increases in cost sharing via copayments or coinsurance are associated with a 2–6% reduction in prescription drug expenditures) (Goldman *et al.*, 2007).

A few studies from the US, Canada, and Taiwan have examined the effect of coinsurance rates rather than fixed dollar copayments. Overall, the effect of coinsurance is at the low end of the price elasticity range of -0.2 to -0.6 (Goldman *et al.*, 2007). Although surprising, this is primarily explained by the fact that in most of the study settings the coinsurance was accompanied by an out-of-pocket maximum per prescription or per year which likely subdued the effect of the coinsurance (Goldman *et al.*, 2007). However, a recent study of US adults with employer sponsored coverage examined coinsurance versus copayments of equal dollar amounts and reported that the use of (adherence to) diabetes medications was lower under coinsurance, perhaps due to the additional uncertainty generated for patients in their out-of-pocket spending depending on changes in the price of drugs (Dor and Encinosa 2010).

As opposed to the relatively modest effects of low (and often flat) levels of cost sharing reported in some of the earliest studies published in the literature, a majority of the recent studies of tiered cost sharing suggest larger effects (Goldman *et al.*, 2007). For example, results from a study of privately insured US adults suggest that doubling drug copayments was associated with 60% reductions in drug spending in the first year (i.e., 'short run' elasticity of -0.6) and a further 20% in the second year (i.e., 'long run' elasticity of -0.8) (Gaynor *et al.*, 2007). A majority of the evidence from studies examining reference pricing also suggests that it lowers prescription drug spending, at least in the first one or two years after implementation (Gemmill *et al.*, 2008).

Very few studies have examined the effect of deductibles. In particular, evidence on deductibles in private plans within the Netherlands suggests price elasticity estimates to be only -0.06 to -0.08 (Gemmill *et al.*, 2008). However, a sizeable number of studies have examined the impact of benefit caps such as annual caps on drug spending, monthly prescription limits, and the coverage gap in Medicare Part D (Goldman *et al.*, 2007; Gemmill *et al.*, 2008; Polinski *et al.*, 2011). As one would expect, prescription drug spending is significantly lower with these extreme forms of cost sharing. For instance, one of the studies of an annual cap of $1000 on total drug spending reported 28% lower prescription drug spending in US elderly patients in a Medicare managed care plan with the cap compared to those not subject to the cap.(Hsu *et al.*, 2006) Similar levels of reductions in prescription drug spending have been reported during the coverage gap in Medicare Part D (Polinski *et al.*, 2011).

There is also consistent evidence to indicate that the payer's prescription expenditures decrease with most forms of cost sharing; not surprisingly, increases in patient out-of-pocket expenditures have also been well documented in the literature (Gibson *et al.*, 2005; Gemmill *et al.*, 2008; Polinski *et al.*, 2011). Although clearly some of the reductions in payer spending on drugs arise from cost-shifting to patients, part of them also arise due to decreases in the overall quantity demanded and/or the price of the prescription drug expenditures (Gibson *et al.*, 2005; Goldman *et al.*, 2007; Gemmill *et al.*, 2008; Eaddy *et al.*, 2012).

The existing evidence largely suggests that the decrease in quantity of drugs demanded occurs through both a reduction in the probability of drug use or initiation (Gibson *et al.*, 2005; Gemmill *et al.*, 2008; Solomon *et al.*, 2009) and a reduction in the number of prescriptions refilled among current users via nonadherence to or discontinuation of ongoing medications; however, the results vary by type of cost sharing change, study population, and drug class studied (Gibson *et al.*, 2005; Eaddy *et al.*, 2012). For instance, a US study of elderly patients with employer sponsored retiree coverage reported that higher cost sharing significantly delayed the initiation of drug therapy within 1 year and 5 years in patients newly diagnosed with hypertension, hypercholesterolemia, and diabetes (Solomon *et al.*, 2009). In a recent review of US and Canadian studies examining the relationship between patient cost sharing and medication adherence, 56 (85%) of the 66 included studies demonstrated a statistically significant negative relationship (Eaddy *et al.*, 2012). The remaining studies demonstrated either limited or nonsignificant findings. Extrapolating across the studies, the authors reported that for each dollar increase in patient copayments, adherence would be expected to decrease by 0.4% (that is, a $10 increase would be associated with a 3.8% drop in overall adherence); however, they note that the actual decline may be larger or smaller depending on the study population and type of cost sharing change given the wide range in the study results. The evidence tends to be more mixed for tiered formularies wherein some studies report decreased adherence and increased discontinuation and others find no changes depending on the drug class and population studied (Gibson *et al.*, 2005). However, majority of the evidence indicates that reference pricing is generally not linked to worsened adherence with drugs at the therapeutic class level (patients may switch from a drug above the reference price to one at the reference price but rarely completely discontinue taking any drug within the therapeutic class) (Goldman *et al.*, 2007). The evidence on

benefit caps (annual caps on drug spending, monthly prescription limits, or the coverage gap in Medicare Part D), however, consistently suggests a reduction in the probability of use of medications and higher rates of discontinuation across common therapeutic classes (Goldman *et al.*, 2007).

Several studies indicate that the reductions in the medication use with higher cost sharing occur for both drugs considered less 'essential' (e.g., antihistamines) as well as more 'essential' (e.g., medications for hypertension, diabetes, and asthma) for stabilizing or improving health (Gibson *et al.*, 2005; Goldman *et al.*, 2007). Although economic theory predicts that reductions in use should be larger for 'less' essential medications relative to 'more' essential medications, the empirical evidence is quite mixed and inconsistent with some studies reporting evidence of such differences and others reporting no such differences (Gibson *et al.*, 2005; Goldman *et al.*, 2007; Gemmill *et al.*, 2008).

Empirical evidence suggests that drug cost sharing in the form of reference pricing and tiered formularies may also reduce drug expenditures either directly by causing pharmaceutical manufacturers to lower drug prices or offer discounts and indirectly via encouraging patients to shift to lower priced drug substitutes (Gibson *et al.*, 2005; Gemmill *et al.*, 2008). For instance, studies from Canada, New Zealand, the Netherlands, and Germany suggest that reference pricing was associated with manufacturers lowering prices of drugs (typically to the reference price) in several therapeutic categories (Gemmill *et al.*, 2008). There is also consistent evidence that patients typically switch from drugs priced above the reference price to drugs priced at the reference price (Goldman *et al.*, 2007; Gemmill *et al.*, 2008). Similarly, adding a third tier for nonpreferred brand drugs has been typically associated with reductions in the use of these medications and an increase in the use of preferred brand-name drugs; however, the extent of substitution varies by drug class (Gibson *et al.*, 2005). Perhaps surprisingly, there is little evidence of generic substitution in the face of differential copayments for generics and brand-name drugs (Gibson *et al.*, 2005; Gemmill *et al.*, 2008). These results may be due to the fact that the generic brand copayment differentials in most studies have been quite small and a majority of the evaluations were conducted during a time frame when fewer generic options were available for brand-name drugs compared to what exists today given patent expirations for numerous block-buster drugs. For the US, this may also reflect the fact that in the widespread presence of state generic substitution laws (wherein a pharmacist can substitute a generic even if the script is written for a brand, unless the physician requires the brand), differential copayments for generics and brand-name drugs may yield minimal additional impact

Health care outcomes, use, and expenditures

In this section, the empirical evidence on the complementary and substitution effects of prescription cost sharing is reviewed. If prescription cost sharing decreases prescription drug use, then to the extent that physician visits serve as complements to the use of prescription drugs (given that they are required for having a prescription written) their use may also decrease. However, reductions in medication use, particularly for drugs considered 'essential' for stabilizing and improving health may have consequences in terms of worsened disease symptoms and clinical outcomes and consequent substitution with other more resource-intensive health care services such as emergency room visits, outpatient visits, hospitalizations, and nursing home admissions.

Although there is a large literature on the effects of cost sharing on prescription use and spending, far fewer studies have examined its effects on health outcomes and nondrug health services utilization and expenditures. The lack of studies on health outcomes is largely due to the unavailability of clinical data (e.g., laboratory test values), vital status, and/or patient self-reported information (e.g., worsening of symptoms) to link with the administrative data on prescription and medical service use. The most compelling results on clinical outcomes come from a US study examining elderly patients in a Medicare managed care plan with a cap of $1000 on annual drug costs compared to those in an uncapped plan (Hsu *et al.*, 2006). Not only were relative rates of mortality higher among those in the capped plan, but physiologic outcomes such as blood pressure, cholesterol, and hemoglobin A1c levels were relatively worse among patients being treated for hypertension, hyperlipidemia, and diabetes, respectively. Evidence on associations between mortality and cost sharing also comes from studies in Canada (Quebec) and Italy (Gemmill *et al.*, 2008).

Most of the evidence on nonmedical service use comes from studies conducted in the US and Canada. Furthermore, the findings are mixed and depend largely on the population being studied and the type of cost sharing changes being evaluated. Evidence from a few studies in the US and Germany suggest that physician visits serve as complements to the use of prescription drugs and their use is lower with increases in drug cost sharing (Gemmill *et al.*, 2008). Contrary evidence on physician visits serving as substitutes comes primarily from studies in Canada (British Columbia) wherein there was no cost sharing for physician visits and hence, perhaps some patients substituted free physician care for prescription drugs in the face of increases in drug cost sharing (Gemmill *et al.*, 2008).

Evidence of substitution effects with other medical services comes primarily from US based studies focusing specifically on patients with chronic diseases such as congestive heart failure, hyperlipidemia, hypertension, and schizophrenia; such studies have linked higher drug cost sharing with greater likelihood of outpatient, inpatient, and/or emergency care (Goldman *et al.*, 2007; Eaddy *et al.*, 2012; Gemmill *et al.*, 2008). Also, and not surprisingly, the evidence of such adverse effects is much stronger and clear-cut for more extreme forms of cost sharing that involve benefit caps in the form of annual benefit caps and monthly prescription limits especially which have been applied in some settings for the elderly and/or poor patients in the US (Hsu *et al.*, 2006; Soumerai *et al.*, 1993). Increase in serious adverse events (acute care hospitalizations, long-term care admissions, mortality) and emergency room visits have also been reported among the poor and the elderly in Canada (Quebec) after cost sharing requirements changed from minimal or no copayments to a 25% coinsurance with income-based annual out-of-pocket maximums (Tamblyn *et al.*, 2001).

The studies that have not found substitution or complementary effects between drug cost sharing and nonmedical

service use are generally those that evaluated cost sharing changes designed to encourage patients to switch to lower cost drug substitutes (i.e., tiered formularies and reference pricing) rather than deter any drug use (Gemmill *et al.*, 2008); those that examine a broader population (not limited to specific chronic diseases) (Goldman *et al.*, 2007); and/or those in which the cost sharing did not have an impact on medication adherence either due to the nature or magnitude of cost sharing changes or the price inelasticity given the severity of illness (e.g., post heart attack) in the population being studied (Eaddy *et al.*, 2012; Gibson *et al.*, 2005).

It is unclear from this sparse literature with mixed findings on whether the net effect of an increase in drug cost sharing results in cost savings or whether it results in increased long-term or short-term total spending due to substitution with other resource-intensive medical services. Most studies examining nonmedical service use have failed to extend their evaluation to examining cost offsets. As expected, formal or informal cost offset calculations from studies of benefit caps in the form of annual caps on drug spending and monthly prescription limits indicate that the drug cost savings are offset by the increases in nonmedical service use such as hospitalizations, emergency room visits, or nursing home admissions in elderly and poor patients (Soumerai *et al.*, 1993; Hsu *et al.*, 2006). A more recent study of drug cost sharing in privately insured adults with employer sponsored coverage suggests that 35% of the cost savings due to reductions in drug spending were offset by increases in other medical spending; however, increases were observed only in outpatient spending (no effects on inpatient spending) (Gaynor *et al.*, 2007).

Effects of Cost Sharing for Specialty Drugs

Specialty drugs represent a relatively new innovation in the pharmaceutical market and hence few studies to date have been conducted that examine the effects of higher cost sharing on the use of such medications and related outcomes and spending. Furthermore, almost all the evidence on this topic comes from data on privately insured patients in the US from a time period when few patients were subject to aggressive cost sharing strategies for specialty drugs that are increasingly becoming more commonplace (e.g., in the US Medicare Part D program).

Specialty drug use and expenditures

The earliest study addressing this topic examined specialty drug use and spending among privately insured patients with one of four conditions (cancer, kidney disease, multiple sclerosis, and rheumatoid arthritis) which are treated with specialty drugs. The authors estimated a fairly inelastic price elasticity of total drug spending, ranging from -0.01 for cancer drugs to -0.21 for rheumatoid arthritis drugs (Goldman *et al.*, 2006).

A recent study estimated the elasticity of demand associated with five specific specialty drugs used to treat cancer in privately insured patients. The authors found that demand elasticity for initiating the five specialty cancer drugs ranged from -0.19 to -0.26 and that for continuing to fill prescriptions among those who initiate was even lower ranging from -0.04 and -0.11 (Goldman *et al.*, 2009). Evidence from a study focusing specifically on use of specialty biologics in privately insured patients with rheumatoid arthritis also suggests that the demand for such RA drugs is relatively inelastic with elasticity estimates of -0.93 and -0.038 for biologic initiation and continuation, respectively. In other words, the authors found a 9.3% reduction in the probability of initiating therapy and a 3.8% reduction in the probability of continuation among those who initiate if average out-of-pocket costs for these drugs were to double (Karaca-Mandic *et al.*, 2010). Another study examining the impact of patient out-of- expense on prescription abandonment (defined as the patient never actually taking possession of the medication despite evidence of a written prescription generated by a prescriber) suggested somewhat more price sensitivity for multiple sclerosis drugs (Gleason *et al.*, 2009). A majority (83%) of the patients with multiple sclerosis had an out-of-pocket expense of $100 or less and these patients had an abandonment rate of 5.7%. In the remaining patient groups whose out-of-pocket expense was greater than $200, the abandonment rate was significantly higher, with more than 1 in 4 patients abandoning their specialty drug prescriptions (Gleason *et al.*, 2009).

Finally, a recent study has reported that long-term users of antiinflammatory, immunosuppressant, cancer, and multiple sclerosis medications whose specialty medication copayments had been increased only experienced a minor decline in adherence, suggesting that responses to cost sharing changes for specialty biologics are not as large as those for traditional small-molecule products (Kim *et al.*, 2011).

Health care outcomes, use, and expenditures

No study has examined the impact of higher cost sharing for specialty drugs on health care outcomes and nondrug medical services use and expenditures. Further research on this topic is needed.

Value-Based Insurance Design

As highlighted in the previous sections of this article, several studies have documented that increased prescription cost sharing not only reduces use of low-value medications but also use of and adherence to highly effective medications needed to appropriately manage chronic conditions. This may be because patients underestimate the marginal benefits of the drugs. This issue has raised the concept of 'value-based' cost sharing, now commonly referred to as value based insurance design (VBID) in the US (Chernew *et al.*, 2007; Pauly and Blavin, 2008).

The premise of VBID is that patient copayment or coinsurance levels are set relative to the value offered by the medication (benefit and costs), and not its cost alone (Chernew *et al.*, 2007). This implies lower cost sharing when clinical benefits of the drug exceed the costs (i.e., medication offers high value) and high cost sharing when benefits do not justify the cost (i.e., medication offers low value), regardless of the actual cost of the drug. However, not all classes of drugs are amenable to value-based cost sharing and due to practical hurdles applications of VBID in the US have been limited

to lowering cost sharing for broad classes of drugs such as antihypertensives, cholesterol-lowering agents, anti diabetic agents, and/or asthma medications (Chernew *et al.*, 2007; Choudhry *et al.*, 2010). Although such VBID approaches are being increasingly adopted by US employers and insurers, only a small number of studies, mostly observational and many without control groups, have evaluated the direct impact of reducing copayments on prescription and medical service use, spending, and outcomes (Fairman and Curtiss, 2011; Choudhry *et al.*, 2010).

One of the first quasi-experimental studies in this area evaluated a VBID program implemented by a large US employer as part of a disease management (DM) program for diabetes Chernew *et al.* (2008). A control employer that used the same DM program but did not change cost sharing levels was utilized to conduct a pre-post analysis to understand the effect of the VBID program. For five classes of medications, copayment levels were reduced from $5 to $0 for generics, from $25 to $12.50 for preferred-brand drugs, and from $45 to $22.50 for nonpreferred-brand drugs. Among patients using angiotensin converting enzyme (ACE) inhibitors or angiotensin II receptor blockers (ARBs), beta-blockers, diabetes drugs, or statins, there was a statistically significant increase in adherence of 2.6–4.0 percentage points. These improvements in adherence translated to price elasticities of -0.11 to -0.20. Among patients taking inhaled corticosteroids, adherence increased by 1.9%, but the effect was not significant (Chernew *et al.*, 2008).

The Post-Myocardial Infarction Free Rx Event and Economic Evaluation (MI FREEE) trial was the first randomized controlled study to examine the impact of VBID; furthermore it not only examined medication adherence to medications prescribed to patients following a myocardial infarction but also clinical outcomes and total health care costs (Choudhry *et al.*, 2011). Patients who were discharged from a hospital after suffering a myocardical infarction were randomized at the employer level to either no cost sharing (i.e., VBID group) or usual cost sharing (i.e., control group) for statins, beta-blockers, ACE inhibitors, and ARBs. Adherence to medication was significantly higher by 4–6 percentage points in the no cost sharing group relative to the control group. However, there were no significant differences between the two groups in the primary end point of the rate of first major vascular event or revascularization. Nevertheless, the secondary outcome of total major vascular events was significantly lower among the no cost sharing group compared to the control group (21.5 vs. 23.3 per 100 person-years). Furthermore, total health care spending was similar between the two groups. The MI FREEE trial concluded that the elimination of cost sharing improved adherence to medication, decreased rates of total major vascular events, and decreased patient spending without increasing total health care costs.

It should be noted that the MI FREEE trial was conducted in patients who had suffered a heart attack in whom improved medication adherence due to reductions in cost sharing is more likely to be cost neutral due to reductions in future cardiovascular events. However, few studies have examined the impact of VBID on medical and total spending when used in the general population of patients with chronic conditions such as hypertension, hyperlipidemia, diabetes, and asthma. The limited number of studies suggest that VBID is either cost-neutral or at times cost saving; however, the reliability of the evidence is highly questionable given that these studies either fail to report results from a payer's perspective, are based on economic modeling using weak assumptions, or do not permit isolation of the impact of the VBID program from concurrent disease management programs (Fairman and Curtiss, 2011). Despite the lack of strong evidence of effectiveness or cost-effectiveness of copayment reductions for medications, employers and payers are increasingly adopting VBID approaches. Hence, there is a clear need for further rigorous evaluations of the clinical and economic outcomes of the impact of lowering medication copayments.

Conclusion

Substantial evidence suggests that patient cost sharing for prescription drugs can reduce third-party payer drug expenditures. These reductions in drug spending have been shown to occur via several mechanisms including a direct shift in costs to patients as well as reductions in the quantity and price of drugs used. However, the mechanisms and magnitude of effects vary considerably by type of change in cost sharing, type of therapeutic drug class, and the type of study population subject to cost sharing. Reductions in medication use have been reported for drugs essential for maintaining or improving health as well as other drugs. Evidence suggests that these reductions in use occur via both decreases in the probability of drug initiation and increases in nonadherence to or discontinuation of ongoing medications for chronic conditions. Although limited data exist on the effects of prescription cost sharing on health outcomes and spending, evidence in certain conditions suggests that higher cost sharing is associated with increased used of nondrug medical services such as hospitalizations and emergency care visits.

Specialty biologics represent a marked shift in the pharmaceutical landscape, with more complex manufacturing methods and significantly higher prices as compared to traditional drugs. The limited evidence on the effects of higher cost sharing for these medications suggest relatively inelastic demand for these products. However, this evidence base is exclusively limited to the experience of privately insured US patients from a time period when very few patients were subject to aggressive cost sharing strategies for specialty drugs such as those seen in the US Medicare Part D program.

In conclusion, as private and public payers continue to experiment with prescription benefit designs in the coming years, research examining the effects of various drug cost sharing policies on prescription and medical use, spending, and outcomes will be essential and of high value.

See also: Pharmaceutical Pricing and Reimbursement Regulation in Europe. Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Value-Based Insurance Design

References

Austvoll-Dahlgren, A., Aaserud, M., Vist, G., et al. (2008). Pharmaceutical policies: Effects of cap and co-payment on rational drug use. Cochrane Database System Reviews, Issue 1 Art. No.: CD007017. DOI: 10.1002/14651858.CD007017

Chernew, M. E., Rosen, A. B. and Fendrick, A. M. (2007). Value-based insurance design. *Health Affairs* **26**, 195–203.

Chernew, M. E., Shah, M. R., Wegh, A., et al. (2008). Impact of decreasing copayments on medication adherence within a disease management environment. *Health Affairs* **27**, 103–112.

Choudhry, N. K., Avorn, J., Glynn, R. J., et al. (2011). Full coverage for preventive medications after myocardial infarction. *New England Journal of Medicine* **365**, 2088–2097.

Choudhry, N. K., Rosenthal, M. B. and Milstein, A. (2010). Assessing the evidence for value-based insurance design. *Health Affairs* **29**, 1988–1994.

Dor, A. and Encinosa, W. (2010). How does cost-sharing affect drug purchases? Insurance regimes in the private market for prescription drugs. *Journal of Economics & Management Strategy* **19**, 545–574.

Eaddy, M. T., Cook, C. L., O'Day, K., Burch, S. P. and Cantrell, C. R. (2012). How patient cost-sharing trends affect adherence and outcomes:A literature review. *Pharmacy and Therapeutics* **37**, 45–55.

Fairman, K. A. and Curtiss, F. R. (2011). What do we really know about VBID? Quality of the evidence and ethical considerations for health plan sponsors. *Journal of Managed Care Pharmacy* **17**, 156–174.

Gaynor, M., Li, J. and Vogt, W. (2007). Substitution, spending offsets, and prescription drug benefit designs. *Forum for Health Economics and Policy* **10**, 1–31.

Gemmill, M. C., Thomson, S. and Mossialos, E. (2008). What impact do prescription drug charges have on efficiency and equity? Evidence from high-income countries. *International Journal for Equity in Health* **7**, 12.

Gibson, T. B., Ozminkowski, R. J. and Goetzel, R. Z. (2005). The effects of prescription drug cost sharing: A review of the evidence. *American Journal of Managed Care* **11**, 730–740.

Gleason, P. P., Starner, C. I., Gunderson, B. W., Schafer, J. A. and Sarran, H. S. (2009). Association of prescription abandonment with cost share for high-cost specialty pharmacy medications. *Journal of Managed Care Pharmacy* **15**, 648–658.

Goldman, D. P., Jena, A. B., Lakdawalla, D. N., et al. (2009). The value of specialty oncology drugs. *Health Service Research* **45**, 115–132.

Goldman, D. P., Joyce, G. F., Lawless, G., Crown, W. H. and Willey, V. (2006). Benefit design and specialty drug use. *Health Affairs* **25**, 1319–1331.

Goldman, D. P., Joyce, G. F. and Zheng, Y. (2007). Prescription drug cost sharing: Associations with medication and medical utilization and spending and health. *Journal of the American Medical Association* **298**, 61–69.

Goldman, D. P. and Philipson, T. J. (2007). Integrated insurance design in the presence of multiple medical technologies. *American Economic Review* **97**, 427–432.

Hsu, J., Price, Mary, Huang, Jie, et al. (2006). Unintended consequences of caps on medicare drug benefits. *New England Journal of Medicine* **354**, 2349–2359.

Karaca-Mandic, P., Joyce, G. F., Goldman, D. P. and Laouri, M. (2010). Cost sharing, family health care burden, and the use of specialty drugs for rheumatoid arthritis. *Health Services Research* **45**, 1227–1250.

Kim, Y. A., Rascati, K. L., Prasla, K., et al. (2011). Retrospective evaluation of the impact of copayment increases for specialty medications on adherence and persistence in an integrated health maintenance organization system. *Clinical Therapeutics* **33**, 598–607.

Newhouse, J. P. and the Insurance Experiment Group (1994). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press.

Pauly, M. V. and Blavin, F. E. (2008). Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics* **27**, 1407–1417.

Polinski, J. M., Shrank, W. H., Huskamp, H. A., et al. (2011). Changes in drug utilization during a gap in insurance coverage: An examination of the medicare Part D coverage gap. *PLoS Medicine* **8**, e1001075.

Solomon, M. D., Goldman, D. P., Joyce, G. F. and Escarce, J. J. (2009). Cost sharing and the initiation of drug therapy for the chronically ill. *Archives of Internal Medicine* **169**, 740–748.

Soumerai, S. B., Ross-Degnan, D, Fortess, E. E. and Abelson, J. (1993). A critical analysis of studies of state drug reimbursement policies: Research in need of discipline. *Milbank Quarterly* **71**, 217–252.

Tamblyn, R., Laprise, R., Hanley, J. A., et al. (2001). Adverse effects associated with prescription drug cost-sharing among poor and elderly persons. *The Journal of the American Medical Association* **285**, 421–429.

Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment

AD Sinaiko, Harvard School of Public Health, Boston, MA, USA

Glossary

Cross-price elasticity The effect of a change in the price of one good or service on the consumption or utilization of another.

Marginal cost The additional cost incurred if the output rate is increased by a small amount.

Marginal value The maximum amount someone is willing to pay for a small increase in consumption or utilization.

Moral hazard Moral hazard can occur when the insurer has imperfect information on the likely behavior of insured individuals. There are two main types. Ex ante moral hazard refers to the effect that being insured has on safety behaviour, generally increasing the probability of the event insured against occurring. Ex post moral refers to the possibility that insured individuals will behave in such a way after an insured event has occurred that will increase the claim cost to insurers, partly because the user-price of care is lower through insurance and demand may therefore rise. It is also often related to insurance fraud.

Price elasticity of demand It is a measure of the change in quantity demanded of a good with a change in its price. It may also be called own-price elasticity of demand.

Introduction

In health insurance, cost-sharing refers to payments that a patient makes directly (i.e., out of pocket) for medical services. Cost-sharing includes a deductible, which is the amount of money a patient pays for services before their health insurance coverage begins, copayments, which are flat payments made for particular products or services (e.g., $15 for a doctor visit or $10 for a prescription), and coinsurance, which is when patients are required to pay a fixed percentage of the cost of their care (e.g., 20% of the cost of a hospital stay).

The extent that cost-sharing affects patient demand for care has long been an important topic in health economics, and research aims to answer questions about how to trade-off moral hazard and risk protection and determine optimal levels of cost-sharing. Risk protection describes protection from financial loss in the case of serious illness, and is often the reason people purchase health insurance. With more complete insurance (i.e., lower cost-sharing) consumers have better protection against losses in wealth from medical care utilization. In contrast, moral hazard in health insurance occurs when patients choose to consume care that they value less than its marginal cost because they pay little or no cost-sharing for the care. Thus, in contrast to risk protection, more complete insurance induces greater use of economically inefficient care.

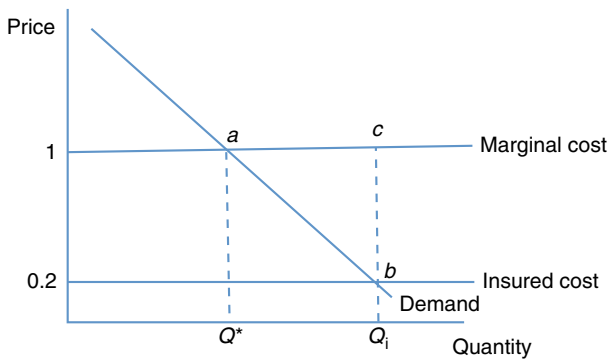


Figure 1 Moral hazard.

The framework that economists have traditionally used to analyze cost-sharing and to depict the inefficient use of resources due to moral hazard is illustrated in Figure 1. A demand curve (D) shows the valuation placed by the consumer on units of medical care, and the marginal cost curve represents opportunity cost, i.e., the valuation placed by consumers on those resources if used to produce other goods or services. In a competitive economy, price equals marginal cost. Each additional unit of medical care that is consumed beyond where the marginal valuation equals marginal cost (point a in Figure 1) carries with it inefficiency, because the medical care is valued less than the cost of the resources used to produce it. This deadweight loss is equal to the vertical distance between the demand curve and the marginal cost curve. It can be seen from this framework that the subsidy provided by health insurance, which reduces the consumer's cost for medical care from marginal cost to the insured cost (here assuming a 20% coinsurance rate), induces additional resources into the production of medical care services beyond the efficient level (the insured consumer demands the amount Q_i, which is greater than Q^*). In Figure 1, the deadweight loss from this moral hazard is depicted by the triangle abc. Deadweight loss from moral hazard varies with price elasticity of demand: Insurance providing coverage for services that are more price elastic will induce more inefficiency than insurance coverage for services that are less price elastic.

The argument made in this welfare economics framework requires several strong assumptions, many of which likely do not hold in healthcare settings. One that is certainly violated

 doi:10.1016/B978-0-12-375678-7.00801-4

is that patients have well-defined preferences for medical care and that when they visit a doctor those preferences are honored in the manner they wish. Instead, the asymmetrical information between a patient and his physician makes the agency relationship central to delivery of healthcare (e.g., patients visit their physician precisely to obtain a diagnosis and be advised on treatment) and there is considerable evidence that the way that providers are paid affects how patients are treated. Under these circumstances, an observed demand curve does not accurately reflect patient's preferences. There is also growing evidence that in complex, high-stakes settings (an apt description of many healthcare settings) consumers do not make rationally consistent choices, which violates a tenant of utility maximization that is required for this framework to hold. Hurley (2000) offers a thorough discussion and critique of the traditional framework, and reviews an alternative framework, which taking a distributional approach at the societal level, suggests the focus be on utilization with an aim toward allocating resources so as to achieve equality of health across individuals. The behavioral response to cost-sharing and estimates of elasticity of demand are important to achieving optimal outcomes under this framework as well.

A Brief Review of the RAND Health Insurance Experiment

The gold standard for evidence on the price elasticity of demand for medical care is from the RAND Health Insurance Experiment (HIE), a large-scale project that implemented a randomized study design to assess the impact of cost-sharing on demand for healthcare in the 1970s and early 1980s. The HIE randomly assigned nonelderly families to commercial insurance plans with varied coinsurance rates and deductible levels, and followed them from 3 to 5 years to study their medical care use and health. Briefly, researchers found that consumers facing higher cost-sharing consumed less care, and they estimate an elasticity of demand for medical care of 0.2 for almost all services studied, except for mental health services and dental services for which demand is estimated to be more responsive. For an average person the difference in medical care utilization did not have an adverse impact on health; individuals who were sick and had low income did experience some adverse health effects as a result of consuming less care in the higher cost-sharing plans. Aron-Dine *et al.* (2013) have recently reexamined this experiment and note that nonlinearities in health insurance contracts can have important effects on how consumers make decisions about their spending at different levels of budgets.

Need for More Recent Evidence

A few factors have led to continued interest in research on the price elasticity of demand for medical care following the HIE. The HIE studied the cost-sharing provisions included in insurance policies of the 1970s. At that time it was rare for a health plan to have a network of physicians or to provide coverage for prescription drugs. At present, insurers create networks of physicians and drug formularies in order to increase their negotiating power to receive more favorable prices. Physicians, physician groups, or pharmaceutical companies have an incentive to grant a lower price or a discount to be in the network or on the formulary in order to obtain a greater volume of business due to the lower patient payment for the drug or visit. For patients, physician networks and formularies has meant that cost-sharing is differential by provider or by drug, so that patients pay lower cost-sharing to see a physician who is in their plan's network or to buy a prescription drug that is a generic or a preferred brand (i.e., on the 'formulary'). Differential cost-sharing raises the empirical question of estimation of cross-price elasticities, which is the impact of a change in the price of one medical service on use of another medical service. If changing the price of one medical service increases (decreases) the use of another medical service, this would either add (or subtract) from the moral hazard that occurs as a result of demand for the service whose price changed.

The way health plans pay providers has also changed since the 1970s, with implications for estimates of elasticity of demand. Capitation, where plans pay providers a fixed, per-member-per-month payment, has been employed by Health Maintenance Organizations (HMOs) and Preferred Provider Organizations (PPOs) over the past 20 years and is an important component of the new Accountable Care Organizations. Under this payment mechanism providers are incented to deliver less care, despite any demand by patients, because they keep the balance of their capitated payments. These incentives did not exist in the health plans that were studied under the HIE, where providers were reimbursed their full charges on a fee-for-service basis.

For the most part, the HIE also did not study use of differential cost-sharing by type of service, though one plan in the HIE varied the cost-sharing a patient would pay for inpatient versus outpatient services (i.e., cost-sharing was applied to outpatient services and inpatient services were free to the patient). The price of outpatient care was not found to affect the demand for inpatient care. However, consumers today face numerous differential prices for medical services, including across types of visits (office visits vs. emergency room visits), types of drugs (generic vs. brand), and even for some, according to the efficiency and quality of their provider through tiered provider networks. In the current environment, optimal cost-sharing depends on cross-price elasticities along with own-price elasticities.

Finally, the HIE focused on a generally healthy, commercially insured population and did not study whether demand response varied differentially for subgroups of consumers (e.g., low income/uninsured, the elderly, and racial minorities).

Thus, due in part to the evolution in benefit design and in part to the desire to test the generalizability of the HIE estimates, several studies of demand response to aspects of medical care have been conducted since the HIE. All are observational studies, the best generally consisting of natural experiments such as the introduction of new insurance policies when patients are subject to plausibly exogenous changes in cost-sharing for medical services.

Estimates of Own-Price Elasticity of Demand since the HIE

Prescription Drugs

Much recent work has studied demand response for drugs in formularies, where the typical structure is either two-tiered, where the lowest cost-sharing tier contains generic drugs and the higher tier branded drugs, or more commonly, three-tiered, where the lowest cost-sharing tier contains generic drugs, the middle tier 'preferred' branded drugs, and the highest cost-sharing tier 'nonpreferred' branded drugs. Several papers review this literature and find that overall, increasing pharmaceutical cost-sharing results in lower utilization of prescribed drugs, i.e., own-price elasticities are negative, and in steering patients away from nonpreferred branded drugs. Gibson *et al.* (2005) review 30 studies from 1974 to early 2005 and report price elasticity of demand ranging from -0.1 to -0.4. In their review of 20 years of this literature (1985–2006), Goldman *et al.* (2007) report that for each increase in cost-sharing of 10%, prescription drug spending decreased by 2–6% (i.e., price elasticity of demand range from -0.2 to -0.6). Huskamp *et al.* (2003) and Goldman *et al.* (2007) show that these findings are consistent across classes of drugs, including medications known to be efficacious for certain chronic diseases, such as statins, angiotensin-converting-enzyme (ACE) inhibitors, antihypertensives, and antidiabetics.

Research on pharmaceutical cost-sharing among non-commercially insured populations is still emerging. The most recent work on the effect of cost-sharing for pharmaceuticals has studied demand response among the elderly following the introduction of Medicare Part D, the Medicare prescription drug benefit. A large literature has found evidence that the introduction of Part D plans has led to lower out-of-pocket spending and higher utilization of prescription drugs, suggesting negative elasticity of demand for pharmaceuticals among the elderly. Joyce *et al.* (2009) found this effect on demand particularly among lower-income beneficiaries. Duggan and Scott Morton (2010) estimate the effect of Part D on changes in utilization and report that their estimates suggest an own-price elasticity of -0.38 for a Medicare recipient with average prescription drug spending.

Future work should continue to study the extent to which demand response varies for population subgroups, especially vulnerable groups such as low-income populations and racial/ethnic minorities. Some recent work by Chernew *et al.* (2008) compares the experience of patients living in low-income areas (based on median household income in patient's zip code of residence) with that of patients in high-income areas, and finds that elasticity of demand for pharmaceuticals is higher for patients in low-income areas. For the most part data limitations (specifically the lack of an indicator for race or socioeconomic status in administrative claims data) have made this research difficult.

Effect of New Health Plan Designs

The rise of managed care since the mid-1990s has meant provider payment in the form of capitation that has increased pressure on providers to curb patient utilization of care, as discussed above. Natural experiments involving the expansion of managed care provide an opportunity to study the elasticity of demand for medical care in the presence of these supply-side incentives. Chandra *et al.* (2010) study changes in copayments for physician office visits and prescription drugs that accompanied the transition of the elderly (retirees) to PPO and HMO plans in the early 2000s by California Public Employees' Retirement System, and estimate that arc elasticities of demand in a managed care environment are less than 0.1, which is about half of that observed in fee-for-service plans studied in the HIE. Goldman *et al.* (2006) find that in the context of implementation of parity for mental health services, supply-side management techniques kept utilization levels largely unaffected by increases in the generosity of mental health coverage, a particularly interesting finding because elasticity of demand for mental healthcare was found to be twice that for other medical care in the fee-for-service environment of the HIE.

Another form of insurance design that has experienced increasing popularity since the early 2000s is high-deductible health plans, where consumers face a large deductible (e.g., $1000 for individuals and $2000 for families) before their health plan begins to cover their costs of care. Preventive services are often exempt from the deductible, and these plans may be paired with a Health Savings Account (HSA). In a large study of the HSA plans from over 700 employers (but from 1 health insurer), LoSasso *et al.* (2010) found that the deductible reduced medical spending by approximately 5%, and the services that were affected tended to be smaller patient-driven services.

Cross-Price Elasticities

Several studies have investigated the impact of differential cost-sharing across medical services, which is now prevalent in health insurance, on demand. In general, this literature reports trends in spending and utilization suggesting that services are substitutes (which implies that cross-price elasticity is positive) or complements (i.e., negative cross-price elasticity), but does not calculate and estimate elasticity of demand.

Pharmaceuticals and Other Medical Services

The vast majority of the studies on this topic consider the cross-price elasticity between pharmaceuticals and other medical services. Evidence is suggestive that pharmaceuticals and other medical services are substitutes. In their review, Goldman *et al.* (2007) summarize the evidence that increased cost-sharing is associated with adverse medical events and outcomes for patients with chronic diseases including congestive heart failure, lipid disorders, diabetes, and schizophrenia. For example, Soumerai *et al.* (1991, 1994) showed that limiting the number of prescriptions in a month in a Medicaid program to three saved money on drugs but led to offsetting increases elsewhere; in the case of schizophrenic patients, the increase in other spending was 17 times the savings on drugs.

Evidence suggesting similar positive cross-price elasticities has also been observed in the Medicare population following

the introduction of Medicare Part D. McWilliams *et al.* (2011) studied spending on nonpharmaceutical medical services in a nationally representative sample of Medicare beneficiaries, and found that the introduction of Medicare Part D resulted in a significantly differential reduction in spending for beneficiaries with limited prior drug coverage (i.e., beneficiaries who had more significant decreases in out-of-pocket spending on pharmaceuticals and presumably increased their utilization and adherence) than beneficiaries with generous prior drug coverage (i.e., those who experienced less change in out-of-pocket spending). They observe the reduced spending predominantly from lower use of inpatient services and skilled nursing facility services. Likewise, in their study of California retirees, Chandra *et al.* (2010) find that increased cost-sharing for pharmaceuticals led to increased utilization of inpatient hospital services. Applying this finding in the opposite direction, Rosen *et al.* (2005) have simulated that it would save money and improve outcomes if ACE inhibitors were made available for free for elderly diabetics in Medicare. When compared with the default cost-sharing in the Medicare drug benefit, the induced increase in the use and cost of ACE inhibitors if they were free would be more than offset by other averted medical costs.

Provider Network

Early efforts (in the 1980s and mid-1990s) by health plans to organize physicians into networks were heavily driven by the provider price or the amount of the discount a provider granted the plan from their price, and simply allowed for physicians to be 'in-network' or 'out-of-network.' These network designs required patients to pay significantly more of the cost of their care if they chose to see an out-of-network physician, and to my knowledge, there is no literature estimating cross-price elasticities for in-network versus out-of-network physicians.

More recently, insurers have begun to offer 'tiered provider networks,' where health insurers sort providers into tiers based on cost and quality performance, and patients pay lower cost-sharing to see a provider in a higher-performing tier. In a tiered network, although the cost-sharing differential across hospitals can be significant (i.e., differential deductible or coinsurance) the cost-sharing difference across providers is usually more modest (differential office visit copay amounts on the order of $10 or $20).

Few studies have assessed consumer response to differences in cost-sharing across tiered providers (the cross-price elasticity of demand) despite the fact that health plans have been experimenting with tiered networks since the early 2000s. Two recent studies analyzed patient-level claims data and find some evidence that consumers switch to preferred providers when the price differential between preferred and nonpreferred tiers is large (i.e., on the order of hundreds of dollars). Scanlon *et al.* (2008) found that some workers were more likely to select a preferred hospital (preferred status based on whether the hospital met set patient safety standards) for medical visits, workers in a second union and all patients admitted for a surgical diagnosis were no more likely to choose the preferred hospitals. Rosenthal *et al.* (2009) analyzed patient loyalty following the narrowing of a PPO physician network that excluded 3% (48 out of 1800) of physicians from the network. The authors find that 81% of patients of affected physicians did not continue to see those physicians following their exclusion from the network, and an additional 7% of patients saw their excluded physician only one more time.

Implications for Insurance Design

An important finding from this literature is that the simple rule that insurers should offer less generous coverage for medical services with higher demand elasticity, is not necessarily ideal. First, the large body of evidence suggesting strong cross-price elasticities across medical services (i.e., that drug consumption lowers the use of other medical services) would call for more generous coverage for prescription drugs, despite the fact that their demand response is relatively elastic. In other words, some patients, especially those with certain chronic diseases, can be induced by less cost-sharing to take actions today that will reduce their future use of medical services and/or improve their future health. It thus follows that optimal cost-sharing could vary across persons and across specific medical goods and services. This logic underlies value-based insurance design (VBID), a mechanism proposed by Chernew *et al.* (2007) where copayments for high-benefit services, such as medications for treatment of chronic disease, would be kept low or set to zero. Although early experience with VBID has shown promise, Baicker and Goldman (2011) offer a thoughtful discussion of the challenges of how to design the benefit design beyond pharmaceutical coverage for individuals with chronic illness, and how to address the potential for risk segmentation of patients into plans with this design.

Final Comments

The empirical research conducted since the 1970s reports estimates of own-price elasticities that are largely in line with those from the RAND HIE. Demand response for pharmaceuticals is also found to be slightly more responsive than that for other services. There is some preliminary evidence suggesting that the most vulnerable populations are more responsive to out-of-pocket medical spending; however, this question should be tested further.

In the context of managed care, cost-sharing is not the only tool to curb inefficient use of medical services. Along with capitation, insurers can apply tools such as gatekeeping and utilization review to affect consumption of medical services. Supply-side mechanisms may affect the estimates of demand response and limit the generalizability of these findings because the effect of the supply-side policies may not be known. However, that physicians have been found to treat their patients in a consistent pattern, such that patterns of care induced by managed care supply-side incentives spillover to a physician's entire patient panel, might limit concerns about generalizability.

There is currently still much more to learn. Efforts to estimate elasticity of demand for medical services for uninsured

populations is now possible due to currently ongoing natural randomized experiment among a Medicaid eligible population in Oregon (Finkelstein *et al.*, 2012); this evidence, not yet available, will greatly enrich the literature. Several opportunities for future research in this area remain. One potential vein is to study how to incorporate findings from behavioral economics, which indicate that consumer decision-making is subject to heuristics and biases, to improve the effect of lowered cost-sharing on consumer demand and the implementation of VBID. Improving our understanding of how cost-sharing can lead to overall cost savings or encourage demand for high quality, effective medical care is an exciting future research agenda.

See also: Demand Cross Elasticities and 'Offset Effects'. Managed Care. Moral Hazard. Prescription Drug Cost Sharing, Effects of. Value-Based Insurance Design

References

Aron-Dine, A., Einav, L. and Finkelstein, A. (2013). The RAND health insurance experiment, three decades later. *Journal of Economic Perspectives* **27**(1), 197–222.

Baicker, K. and Goldman, D. (2011). Patient cost sharing and health care spending growth. *Journal of Economic Perspectives* **25**(2), 47–68.

Chandra, A., Gruber, J. and McKnight, R. (2010). Patient cost-sharing and hospitalization offsets in the elderly. *American Economic Review* **100**(1), 193–213.

Chernew, M. E., Gibson, T. B., Yu-Isenberg, K., et al. (2008). Effects of increased patient cost-sharing on socioeconomic disparities in health care. *Journal of General Internal Medicine* **23**(8), 1131–1136.

Chernew, M. E., Rosen, A. B. and Fendrick, A. M. (2007). Value-based insurance design. *Health Affairs* **26**(2), w195–w203.

Duggan, M. and Scott Morton, F. (2010). The effect of Medicare Part D on pharmaceutical prices and utilization. *American Economic Review* **100**(1), 590–607.

Finkelstein, A., Taubman, S., Wright, B., et al. (2012). The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics* **127**(3), 1057–1106.

Gibson, T. B., Ozminkowski, R. J. and Goetzel, R. Z. (2005). The effects of prescription drug cost sharing: A review of the evidence. *American Journal of Managed Care* **11**(11), 730–740.

Goldman, D. P., Joyce, G. F. and Zheng, Y. (2007). Prescription drug cost sharing: Associations with medication and medical utilization and spending and health. *Journal of the American Medical Association* **298**(1), 61–69.

Goldman, H. H., Frank, R. G., Burnam, M. A., et al. (2006). Behavioral health insurance parity for federal employees. *New England Journal of Medicine* **354**(13), 1378–1386.

Hurley, J. (2000). An overview of the normative economics of the health sector. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1A. Amsterdam: Elsevier.

Huskamp, H. A., Deverka, P. A., Epstein, A. M., et al. (2003). The effect of incentive-based formularies on prescription-drug utilization and spending. *New England Journal of Medicine* **349**(23), 2224–2232.

Joyce, G. F., Goldman, D. P., Vogt, W. B., Sun, E. and Jena, A. B. (2009). Medicare Part D after two years. *American Journal of Managed Care* **15**(8), 536–544.

LoSasso, A. T., Shah, M. and Frogner, B. K. (2010). Health savings accounts and health care spending. *Health Services Research* **45**(4), 1041–1060.

McWilliams, J. M., Zaslavsky, A. M. and Huskamp, H. A. (2011). Implementation of medicare Part D and nondrug medical spending for elderly adults with limited prior drug coverage. *Journal of the American Medical Association* **306**(4), 402–409.

Rosen, A. B., Hamel, M. B., Weinstein, M. C., et al. (2005). Cost-effectiveness of full medicare coverage of angiotensin-converting enzyme inhibitors for beneficiaries with diabetes. *Annals of Internal Medicine* **143**(2), 89–99.

Rosenthal, M. B., Li, Z. and Milstein, A. (2009). Do patients continue to see physicians who are removed from a PPO network? *American Journal of Managed Care* **15**(10), 713–719.

Scanlon, D. P., Lindrooth, R. C. and Christianson, J. B. (2008). Steering patients to safer hospitals? The effect of a tiered hospital network on hospital admissions. *Health Services Research* **43**(5p2), 1849–1868.

Soumerai, S. B., McLaughlin, T. J., Ross-Degnan, D., Casteris, C. S. and Bollini, P. (1994). Effects of limiting medicaid drug-reimbursement benefits on the use of psychotropic agents and acute mental health services by patients with schizophrenia. *New England Journal of Medicine* **331**(10), 650.

Soumerai, S. B., Ross-Degnan, D., Avorn, J., McLaughlin, T. J. and Choodnovskiy, I. (1991). Effects of medicaid drug-payment limits on admission to hospitals and nursing homes. *The New England Journal of Medicine* **325**(15), 1072.

Further Reading

Ketchum, J. D. and Simon, K. I. (2008). Medicare Part D's effects on elderly patients' drug costs and utilization. *American Journal of Managed Care* **14**(supplement 11), SP14–SP22.

McGuire, T. G. (2012). Demand for health insurance. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *Handbook of health economics*, vol. 2. North Holland: Elsevier.

Newhouse, J.-P. and The insurance experiment group (1993). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge: Harvard University Press.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**(3), 531–537.

Remler, D. and Greene, J. (2009). Cost-sharing: A blunt instrument. *Annual Review of Public Health* **30**, 293–311.

Yin, W., Basu, A., Zhang, J. X., et al. (2008). The effect of medicare Part D prescription benefit on drug utilization and expenditures. *Annals of Internal Medicine* **148**(3), 169–177.

Zeckhauser, R. J. (1970). Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**(1), 10–26.

Pricing and Reimbursement of Biopharmaceuticals and Medical Devices in the USA

PM Danzon, University of Pennsylvania, Philadelphia, PA, USA

Introduction

The US is the largest market for pharmaceuticals in the world. It is also one of the few countries in the industrialized world that does not regulate pharmaceutical prices. This largely reflects the predominance in the US of competing private health insurance plans. Although the public insurance programs, mainly Medicare and Medicaid, now account for more than 40% of total health expenditures in the US, their design has historically been influenced by the design of the private insurance sector. In particular, Medicare Part D, which covers outpatient drugs for seniors, was designed to be run by competing private sector prescription drug plans (PDPs) and the enabling legislation explicitly bars the government from negotiating drug prices.

In the US, as in every other country, insurance that provides financial protection to consumers thereby also tends to make consumers insensitive to prices. For patented biopharmaceuticals, this enables producers to charge higher prices than they would in the absence of insurance. In countries with either national health insurance or regulated social insurance, government payers respond to this effect of insurance by constraining prices either directly, through price controls, or indirectly by making reimbursement contingent on cost-effectiveness. In the US no single private payer has sufficient market power to control pharmaceutical prices. Rather, the negotiation of prices with producers and the design of cost-sharing and other access controls are dimensions of competition between health plans. Fundamentally, health plans that are more restrictive on price to suppliers can offer consumers lower cost plans but with restricted access to services.

In practice, most private health plans and Medicare use similar approaches to negotiating prices and/or setting reimbursement rules for pharmaceuticals. This article describes the predominant reimbursement rules used by these US payers and the effects of these rules on manufacturer pricing, highlighting the important differences based on where a drug is dispensed and between onpatent brands and generics.

Overview of the Drug Distribution System and Price Levels

In the US pharmaceutical market there are multiple prices, corresponding broadly to different levels of the distribution chain and whether the price is a list price or a transactions price (that reflects discounts or rebates off the list price). The main price levels and types are outlined here in summary, and the following sections describe the system and its effects in more detail.

Pharmaceutical manufacturers typically sell drugs to wholesalers at a list price, the wholesale acquisition cost (WAC), sometimes with modest discounts (1–2%) for prompt payment. Wholesalers distribute drugs to retail pharmacies (including mail-order pharmacies) and hospital pharmacies, adding a competitively determined mark-up to cover their distribution costs. Traditionally, an estimate of the average price at which wholesalers sold to pharmacies was published by pricing agencies as a list price called average wholesale price (AWP). This list price was widely used by payers as a basis for their reimbursement to pharmacies. However, AWP became an increasingly unreliable (usually inflated) measure of the actual average transaction price at which wholesalers sell to pharmacies. Although AWP and other list prices have remained the basis for payer reimbursement to pharmacies, there is usually a significant discount. For example, a payer may set pharmacy reimbursement at AWP-18%, where the discount off AWP is negotiated between the payer and the pharmacy chain. This reimbursement price is intended to allow the pharmacy to cover the cost of drug acquisition plus a competitive dispensing fee. The payer reimburses the pharmacy at this price, net of any patient cost-sharing that the pharmacy must collect from the patient, depending on the payer's plan design.

Private payers also negotiate discounts from manufacturers of patented drugs directly. These negotiated discounts are usually paid by electronic transfer from the manufacturer to the payer, thus bypassing the wholesaler/retailer system. This preserves confidentiality of the payer-specific discount amounts and prevents price arbitrage, that is, the manufacturer transfers the discount directly to the payer for whom it is intended. The ability of payers to extract these discounts from manufacturers depends on the payer's ability to influence drug use through its formulary design. The average price received by manufacturers, taking into account these discounts given to private payers, is called the average manufacturer price (AMP). Medicaid and some other public payers by statute get mandatory rebates off AMP.

This system of a manufacturer list price, combined with negotiated discounts to private payers and mandatory rebates to government payers, is similar for generics, except that the negotiated discounts given by generic manufacturers are targeted at dispensing pharmacies, rather than at health plans/payers, because the pharmacies are the ultimate decision-makers for multisource drugs (off-patent products with originator and generic suppliers). This is described below.

Discussion of pharmaceutical pricing in the US tends to focus on prices charged by manufacturers, including both the list prices and the discounts/rebates, because these exmanufacturer prices form the basis for prices paid by final payers, with the addition of wholesaler and pharmacy margins that are competitively determined. The wholesale segment in the US is highly concentrated with three firms accounting for more than 80% of the national market, reflecting large economies of scale. It is nevertheless highly competitive, partly due

to increasing concentration and strong competition at the retail pharmacy (including mail order) level, where the top six chains now account for more than 60% of dispensing sales. Pharmacy regulation in the US requires that retail pharmacies employ a licensed pharmacist, but there are no requirements that pharmacies be owned by pharmacists and no restrictions on chain pharmacies, in contrast to restrictive pharmacy ownership regulations in many other industrialized countries. Over the last two decades, the large chain pharmacies, such as Walgreens, RiteAid, and CVS, have grown by increasing their number of outlets and the range of products and services they offer, besides drugs. Conversely, some large supermarkets and department stores like Walmart operate pharmacies within their stores. These large, chain retailers take advantage of economies of scale and scope, and play a major role in driving competition in the wholesale and the generics sector, as described below. Within each geographic market multiple chains compete and competitive pressure on pharmacy margins is enhanced by the bargaining power of large health plans and pharmacy benefit managers (PBMs) such as Express Scripts and Caremark, including competition from their mail order pharmacies that compete with bricks and mortar pharmacies. Similarly, for inpatient drugs, hospitals purchase through large group purchasing organizations that negotiate with wholesalers and put competitive pressure on distribution margins.

Thus in the US, the exmanufacturer list price and discounts, the wholesale mark-ups, retail mark-ups, and final prices to payers/consumers are freely determined, constrained by market competition, in contrast to most other countries where prices and mark-ups at each of these levels are set by regulation. In this article 'price' refers to the exmanufacturer price, before discounts, unless otherwise noted. The term 'cost-sharing' is used to refer to the component of the final price paid by the consumer.

Why Biopharmaceuticals Markets are Different

Although US biopharmaceutical markets are structurally competitive, as described above, several important factors differentiate biopharmaceutical markets and pricing from those for most goods. For most goods that are sold in reasonably competitive markets to reasonably informed consumers, standard economic theory implies that competition will align prices with value to consumers and marginal cost to producers, yielding outcomes that are broadly consistent with economic efficiency. Achieving efficient pricing for pharmaceuticals is complicated by several factors. First, R&D is roughly 17% of sales for the US-based originator pharmaceutical industry, compared to 4% for other US industries. Marginal cost-pricing, which is the expected outcome in competitive markets, would achieve first best static efficiency but would fail to cover total costs and would violate the requirement for dynamic efficiency that producers capture the full social surplus produced by innovation. To address the need for R&D incentives, patents (and other exclusivities) bar generic competitors for a limited term. Patents intentionally enable originator firms to price onpatent products above marginal cost and thus potentially recoup R&D expenses. Although patent-induced pricing above marginal cost may lead to only 'second best efficient' utilization of patented products, patents are the generally accepted way to pay for R&D, as reflected in the World Trade Organization's Trade-Related Intellectual Property (TRIPS) provisions. Thus, patents and the resulting temporary market power are not intrinsically a cause for concern over pharmaceutical prices.

Second and more problematic is the effect of comprehensive insurance coverage on pricing. Insurance protects consumers from financial risk and, through cross-subsidies, makes health services more affordable to low-income consumers. However, because such insurance makes patient demand highly price-inelastic, insurance creates the potential and incentives for manufacturers to charge prices that exceed the level that would result from patents alone. Public and private insurers may use various strategies to constrain this 'producer moral hazard.' In most industrialized countries payers either control prices directly, through price or reimbursement regulation, or require evidence that the drug is cost-effective, which indirectly limits the manufacturer's price based on the drug's incremental effectiveness. Third, patients, payers, and even physicians often lack good information about effectiveness of medical goods and services, which may undermine price-sensitivity. For biopharmaceuticals and devices, this uncertainty is mitigated by regulation of safety, efficacy, manufacturing quality, and promotion. Thus the effect of insurance is the main cause for concern over pharmaceutical prices.

Private and public payers in the US use neither direct price regulation nor indirect price control through incremental cost-effectiveness thresholds as a requirement for reimbursement. Rather, US payers influence the prices charged by manufacturers primarily through use of tiered formularies that offer preferred formulary position and therefore larger market share to drugs that are favorably priced (or give larger discounts), relative to therapeutically similar drugs. Medicaid and other smaller public programs receive mandatory rebates off the manufacturer's price. These approaches leave list prices unconstrained but do achieve significant discounts on onpatent drugs in crowded drug classes with several close therapeutic substitutes. However, these approaches provide little constraint on prices of drugs that are more unique, including most specialty drugs and biologics. By contrast, reimbursement and substitution rules for generics result in highly price competitive generic markets and very low generic prices in the US. The following sections describe in detail these reimbursement rules and their effects on pricing in the US.

Reimbursement Rules for Onpatent Brands

In the US, payer rules and approaches to pharmaceutical reimbursement differ, depending on where the drug is dispensed – retail pharmacy, physician office, or hospital inpatient (Table 1). Reimbursement differences largely reflect the historical evolution of insurance coverage. Retail pharmacy (54% of prescription sales) and mail order pharmacy (17% of sales) dispense self-administered drugs and are reimbursed on a fee-for-service basis by a private insurer's pharmacy benefit or by Medicare's Part D benefit for seniors. Drugs dispensed as part of an inpatient hospital admission (approximately 10% of sales) are covered by the patient's inpatient benefit

Table 1 US: Reimbursement rules depend on product type and distribution channel

Channel	*Retail pharmacy*	*Physician office*	*Hospital inpatient*
Medicine type	Orals, creams, and self-injectibles	Biologics, infusions, and vaccines	All types
Benefit	Pharmacy/Medicare D	Medical/Medicare B	Hospital/Medicare A
Reimbursement	Tiered formularies with access controls	Buy-and-bill with ASP + 6%	Hospital is paid DRG per admission

(Medicare Part A for seniors). Inpatient drugs are reimbursed, along with all other inpatient costs, in a single bundled payment for the hospital admission. Drugs that are dispensed in physicians' clinics (approximately 12% of sales), including infusions and vaccines, are covered by the patient's medical benefit (Medicare Part B) which pays for physicians' services. Because many new, expensive biologics are physician-dispensed, including many oncologics, the reimbursement rules for this category are critical to pricing of biologics in the US.

Pharmacy-Dispensed Drugs

Primary care drugs

Private health plans use PBMs to manage drugs that are dispensed through retail pharmacies. PBMs developed in the 1990s as stand-alone, independent contractors that managed drug benefits on behalf of self-insured employers and other health insurers. Since then, some large insurers have developed their own in-house PBMs that compete with the stand-alone PBMs. When Medicare Part D was created in 2003 to provide outpatient drug coverage for seniors, administration of the Part D benefit was assigned to competing, private entities called prescription drug plans (PDPs), which are similar to PBMs, with important differences noted below. Many private health insurers and PBMs also serve as PDPs.

Private PBMs and Medicare PDPs use similar strategies to manage drug costs. Specifically, a pharmacy and therapeutics (P&T) committee, which includes physicians and pharmacists, evaluates alternative drugs and designs the formulary, that is, the list of drugs that are covered, with associated patient cost-sharing levels and any other controls. Most plans use a formulary with three or more tiers with corresponding copayments. The first tier is for generics and has a US$0–10 copayment per prescription (or a month's supply of a chronic medication); the second tier includes preferred onpatent brands with a modest (US$25–45) copayment; and the third or nonpreferred brand tier has significantly higher copayment, currently approximately US$45–90 per month. In addition, a fourth tier is increasingly used for expensive specialty drugs and usually has a 25–30% coinsurance of the drug's price. Additional tiers with high coinsurance rates may also apply to 'lifestyle' drugs. These tiers and associated differential copayments are designed to incentivize patients and their physicians to accept generics, if available, or choose 'preferred' brands among onpatent brands.

In addition to these differential copayments, plans increasingly also use direct controls to achieve appropriate utilization. Most common are a step edit (a computerized block that automatically rejects reimbursement of a drug unless the patient meets certain conditions, such as prior failure on a generic alternative) and prior authorization (which requires the physician to obtain prior approval from the health plan before a drug is reimbursed).

Formulary design with tiered cost-sharing, step edits, and prior authorizations enables PBMs/PDPs to shift drug utilization toward preferred drugs. This ability to 'shift share' within a therapeutic class gives plans leverage to negotiate price discounts from manufacturers in return for preferred formulary positioning. For example, a plan that is willing to severely limit the number of preferred brands and impose a large copay differential for nonpreferred brands creates leverage in price negotiations, because a drug manufacturer may be willing to give a large discount to be the only brand on the preferred tier, whereas they may give little or no discount if they share the preferred tier with all competitor products in the class.

Thus, this tiered formulary approach enables PBMs to gain significant leverage over manufacturer prices provided that there are several clinically similar drugs in a class and the PBM is able/willing to limit patient choice of drugs. More restrictive PBM plans that limit patient choice and impose high cost-sharing on nonpreferred drugs can get lower prices and offer lower premiums. Essentially, this process structures patient cost-sharing and utilization controls to increase the cross-price demand elasticity facing manufacturers. It gives payers and patients a trade-off between drug choice and cost of coverage. It has worked reasonably well for large, primary care therapeutic classes, such as statins or antiulcerants, where the availability of several, therapeutically similar drugs has enabled PBMs to drive deep discounts, particularly once generics become available in a crowded class. Because these discounts are confidential, comprehensive data are not available and conclusions here are based on anecdotal and the limited, publicly available data.

Discounting has been challenged by retail pharmacists in antitrust litigation alleging collusive pricing and price discrimination by drug manufacturers (Scherer, 1997). Dispensing pharmacies do not receive the discounts on onpatent drugs comparable to those given to PBMs because pharmacies cannot - and arguably should not – independently influence a physician's/patient's choice between therapeutic substitutes. This litigation conspicuously excluded off-patent drugs and generics, because for these drugs the discounts go to the pharmacies as decisionmakers in choosing between generically equivalent versions of a prescribed compound (see below). Under the settlement of this litigation, manufacturer discounts were to be made available on the same terms to all purchasers; however, because PBMs/PDPs and payers design the formularies that drive therapeutic substitution, they remain the main recipients of discounts on onpatent drugs, whereas pharmacies (including mail-order pharmacies) are the main recipients of discounts on generics.

Consistent with this theory, that the largest discounts go to payers that have greatest control over market share, the

conventional wisdom is that Kaiser gets among the deepest discounts, because Kaiser, as a staff-model health maintenance organization whose physicians work only for Kaiser, can enforce formulary adherence and steer utilization toward preferred drugs. By contrast, most private payers have limited ability to enforce their formularies and influence the prescribing practices of independent physicians because each payer's patients account for a small fraction of each physician's practice.

Specialty drugs

The tiered formulary approach works reasonably well for large classes with several drugs that are close clinical substitutes. However, it works less well for specialty drugs and other classes with few close substitutes. 'Specialty drugs' refer to relatively high-priced drugs used to treat complex diseases for which most prescribing is done by specialist physicians, such as cancer, rheumatoid arthritis, multiple sclerosis, and all rare diseases. Even specialty drugs for the same indication often differ in efficacy and tolerability for individual patients, such that doctors and patients are unwilling to accept payer control over clinical choices. Over the past decade, biopharmaceutical innovation has increasingly shifted toward such specialty drugs, including many biologics that each has distinct risks and benefits. For such specialty drugs, PBMs' only tools to control spending are high patient cost-sharing and/or prior authorizations and step edits, to assure that the patient tries any cheaper alternatives first. These mechanisms at most control utilization, but have little direct effect on price. The limited control of payers over the price of specialty drugs is one factor making these drugs a more attractive target for pharmaceutical R&D compared to primary care therapeutic classes with close substitutes and potential genericization.

Medicare PDPs have taken the lead in placing pharmacy-dispensed specialty drugs (which Medicare defines as drugs that cost US$600 or more a month) on a fourth 'specialty' tier with a 25–33% coinsurance, and PBMs are increasingly following this approach. These high coinsurance percentages applied to very expensive drugs potentially imply patient cost-sharing of hundreds of dollars per month. Simple insurance theory suggests that such high patient cost-sharing may imply inappropriately high financial risk for patients and make patients highly price sensitive (and noncompliant), which might constrain manufacturer prices. However, in practice, the majority of patients are protected from such high cost-sharing by other features of their coverage or by manufacturer assistance programs. Specifically, low income seniors are protected from most cost-sharing by Medicare Part D's low income subsidy, and all seniors are protected by the catastrophic stop-loss on Part D cost-sharing, which in 2013 is US$4750 per year, after which the patient pays at most 5% of the drug price (0 for Medicaid-eligibles). Moreover, manufacturers are required to give Medicare patients a 50% discount while they are in the coverage gap ('doughnut hole') where they must pay the full drug cost. These discounts are ignored in calculating beneficiary's out-of-pocket expenses, so effectively they reach the stop-loss after lower cost-sharing. Some private patients may also face high cost-sharing and some currently have no catastrophic stop-loss. For such patients, manufacturers increasingly provide patient assistance programs (PAPs) for low income patients and cost-sharing coupons for other patients (such coupons are illegal for Medicare patients).

Thus although patients nominally face high cost-sharing for specialty drugs, in practice actual marginal cost-sharing is often minimal due to the combination of supplementary insurance through Medicaid and other private coverage, stop-loss limits, copay coupons, and patient assistance programs. In that case, cost-sharing is ineffective at constraining manufacturer prices for specialty drugs. There is little robust evidence on effects of this recent high cost-sharing for costly, specialty drugs, and obtaining reliable estimates is difficult if those who truly do face the 25–30% coinsurance simply forego the treatment. However, it seems likely that for an increasing fraction of new drugs, patient costsharing, which is the main approach to constraining prices in the US, cannot simultaneously constrain manufacturer pricing and enable appropriate patient use.

The fact that Medicare PDPs typically have significantly higher copayments for nonpreferred brand drugs than private PBMs, and more PDPs use specialty tiers with a 25–30% coinsurance for specialty drugs, suggests that PDPs' increasing use of these cost-sharing strategies may partly reflect the greater financial and adverse selection risk born by PDPs, due to three factors. First, to incentivize PDPs to control costs, by law the PDP is at risk for 15% of a patient's cost beyond the Medicare catastrophic threshold (US$6955 in 2013). By contrast, PBMs are not directly at risk for the drug spending of their enrollees, rather, they are reimbursed a fee per script and retain a fraction of the discounts they negotiate. Second, PDPs face greater adverse selection risk because most Medicare beneficiaries can choose between several stand-alone PDPs. If one PDP in an area were to offer more generous coverage of specialty drugs, it might attract a disproportionate share of the patients who need these and other drugs. By contrast, each private employer offers their employees only one PBM, hence that PBM does not face adverse selection within the employee pool. Third, by law Medicare PDPs are exempt from tier exemption requests for drugs on a specialty tier, hence use of a specialty tier may reduce the administrative cost burden of exemption requests for PDPs, which would likely be significant if the PDP were to place some specialty drugs on a preferred tier while putting others on a nonpreferred tier with very high cost-sharing.

Medicaid

Unlike the Medicare Part D drug benefit, which is operated by private sector entities that use similar tiered formularies and negotiated discount strategies to private PBMs, the federal-state Medicaid program uses mandatory rebates. Because Medicaid beneficiaries are low income families with children, seniors and the disabled, even modest patient cost-sharing may lead to noncompliance. Rather than use tiered cost-sharing, since 1990 Medicaid has required manufacturers to give a mandatory rebate equal to the greater of 15.1% off the AMP (which is the manufacturer's average price charged to the private sector, including discounts) or the 'best price' (largest rebate) given to any private payer. For generics, the mandatory Medicaid rebate was a flat 11%, unrelated to discounts to other payers. When Medicare Part D was established in 2003, drug coverage for 'dual eligible' seniors (who are eligible for both

Medicaid and Medicare) was exempted from these Medicaid rebates, and rebates to Medicare PDPs were exempted from the definition of 'best price.' Under the Affordable Care Act of 2010 (ACA), the minimum Medicaid rebate on brand drugs was increased to 23.1% (13% for generics) and Medicaid Managed Care Organizations are required to pay this rebate on Medicaid-eligible enrollees.

By requiring that manufacturers of brand drugs give to Medicaid the largest discount they give to any private purchaser, Medicaid's 'best price' rule effectively raised the cost of giving discounts that exceed the mandatory minimum Medicaid rebate (15.1% before 2010, now 23.1%) to private payers. Manufacturers rationally give discounts to customers who use formularies to create elastic demand. But paying the government a rebate for Medicaid usage has no effect on drug utilization by Medicaid patients, as the rebate is unrelated to preferred formulary status or incentives of patients or prescribers. Therefore, from the perspective of manufacturers, tying a mandatory rebate to Medicaid to a discount given to private payers reduces the overall elasticity of response to private rebates beyond the mandatory minimum, which is now the weighted average of the (presumably elastic) response of the private enrollees and the totally inelastic response of Medicaid enrollees. Thus, the Medicaid best price requirement reduced manufacturer willingness to give discounts to private payers in excess of the mandatory minimum Medicaid rebate, particularly for drugs with relatively high usage by Medicaid patients.

Empirical studies have confirmed that private sector rebates declined in response to this Medicaid best price. When Congress established Medicare Part D, discounts given to Medicare PDPs were explicitly excluded from the Medicaid best price calculations, in order to encourage manufacturers to give deep discounts to PDPs. The 2010 increase in the mandatory minimum Medicaid rebate to 23.1% means that the 'best price tax' now only applies to private sector discounts larger than 23.1%, hence increased discounting up to this 23.1% threshold is expected, ceteris paribus.

Mandatory Medicaid rebates may also create an incentive for manufacturers to raise the price from which the rebate is calculated. Anticipating this effect, Medicaid requires an additional rebate equal to the cumulative excess increase in a drug's price over the consumer price index (CPI), since the drug's launch. This 'excess-CPI rebate' has not been sufficient to eliminate increases in manufacturer prices for onpatent drugs faster than the CPI in recent years. Thus, the Medicaid mandatory rebate provisions have probably contributed to both higher list prices and smaller discounts for private sector payers. Consistent with this, Duggan and Scott Morton (2006) found that drugs with a higher Medicaid share experienced larger increases in prices (including discounts) to private payers.

Physician-Dispensed Drugs

Drugs that require infusion or injection, including many cancer drugs and other biologics, are dispensed in physician clinics. For Medicare patients, physician-dispensed drugs are covered under Medicare Part B (which covers physician services) rather than Medicare Part D (which covers pharmacy-dispensed outpatient drugs). Before 2005, Medicare reimbursed dispensing physicians at 95% of AWP, an unregulated list price, and most private payers followed suite. This reimbursement rule created incentives for manufacturers to compete for market share by offering discounts off AWP to physician practices, in order to increase the margin between their acquisition cost and the reimbursement. Evidence has confirmed that financial incentives influenced physician prescribing choices (Epstein and Johnon, 2012). Lawsuits have also alleged that some manufacturers raised AWP in order to increase the physicians' margin. The margin accrued to dispensing physicians because Medicare and other payers did not attempt to reduce their reimbursement price to capture the discounts. This contrasts to payer response to a similar incentive system for generics in the US (see below) or in Japan. In Japan, manufacturers similarly offer discounts below the reimbursement price to physicians who dispense drugs, as an inducement to use their drugs. However, the Japanese payers (partially) capture these competitive discounts by reducing their reimbursement price paid to dispensing physicians, based on a biennial audit of actual acquisition prices.

Following substantial litigation over the alleged manipulation of AWP and large margins given to dispensing physicians, in 2005 Medicare changed its Part B reimbursement, intending to align reimbursement more closely to actual acquisition prices. Under the new rules, Medicare Part B reimburses dispensing physicians at the manufacturer's Average Selling Price (ASP) plus a six percent margin. Manufacturers are required to report each drug's ASP quarterly, which is defined as the volume-weighted average manufacturer selling price, including all discounts, lagged two quarters. In the short run, this shift to ASP + 6% reimbursement reduced the prices that Medicare Part B pays for drugs. But in the longer run, the ASP + 6% formula eliminates incentives for manufacturers to compete on price, because any discounts offered to physicians in quarter T reduce the ASP and hence reduce the reimbursement price and the 6% margin for all physicians in period T + 2. Moreover, the ASP + 6% reimbursement rule creates perverse incentives for manufacturers to compete by charging high rather than low prices, because a higher price offers a larger margin to the dispensing physician. The main impact of the perverse ASP + 6% incentives is for higher launch prices. Raising prices postlaunch by more than one or 2% a quarter risks squeezing physicians' margins because their reimbursement only rises after a two quarter lag. Because many private payers follow Medicare reimbursement, this Part B reimbursement rule and its perverse incentives have probably contributed to higher prices for oncologics and other biologics in the US.

Despite the perverse price-increasing incentives created by Medicare's ASP + 6% reimbursement rule, two factors may provide some constraint. First, as prices for these drugs rise and increasingly exceed $40 000 or even $100 000 per treatment course, physicians that 'buy-and-bill' face significant cash flow cost and even risk, if reimbursement is uncertain. Thus, uptake of some very costly drugs has initially been slower than expected, at least until payer reimbursement is assured. Second, the 20% Medicare Part B patient cost-sharing should in theory act as some constraint on manufacturer

prices. However, in practice most seniors are protected from this cost-sharing by supplementary insurance through either Medicaid (for low income seniors) or Medigap (employer-sponsored or privately purchased). Similarly, although private insurance plans usually have cost-sharing, most private patients have a stop-loss limit on annual out-of-pocket costs and such limits become mandatory under the PPACA. Further, for uninsured or privately-insured patients who face significant out-of-pocket costs, most manufacturers offer copay coupons or patient assistance programs (copay coupons are illegal for Medicare patients). As a last resort, although physicians cannot waive copayments without risk of violating antikickback statutes, they can refer patients to a hospital outpatient department, which may waive copayments. Thus similar to the situation for pharmacy-dispensed specialty drugs, most patients are protected from the nominally high cost-sharing on physician-dispensed drugs. If so, manufacturers face highly inelastic demand and little if any constraint on pricing.

Hospital Inpatient Drugs

Drugs that are dispensed as part of an inpatient episode are generally not reimbursed separately but are included in the bundled diagnosis-related group (DRG) payment for the hospital admission. Medicare updates its DRG payment rates over time, based on national average costs, by DRG, as reported in hospital cost reports. Private payers negotiate various forms of bundled payment for inpatient hospital care, with private rates generally above Medicare rates but also no separate reimbursement for inpatient drugs. Thus in the short run the cost of new inpatient drugs (or price increases for existing drugs) are borne by hospitals, with pass-through to payers with a lag, if/when the drug becomes standard of care and reflected in average cost for the DRG. In exceptional circumstances, a very high-priced new drug may be reimbursed separately from the DRG temporarily, until its cost is included in an increased DRG payment.

This system of bundled payment for inpatient admissions puts hospitals at risk for inpatient drug costs in the short run. Hospitals therefore have incentives to be price sensitive in designing their formularies and negotiate price discounts with manufacturers in return for preferred formulary placement. Larger hospital systems that negotiate on their own behalf and can enforce formularies have greater bargaining power and get larger discounts than smaller hospitals and those that bargain indirectly through group purchasing organizations (GPOs). However, as with PBMs/PDPs, hospitals have little or no leverage to negotiate discounts for drugs that have few or no close substitutes, which includes many specialty drugs.

Generics

In 2011, 80% of all prescriptions were dispensed as generics in the US, and for compounds with a generic available the generic share of scripts was 94%. By contrast, generics account for only 27% of dollar value of sales (IMS, Health Informatics Institute, April 2012). This high generic share by volume reflects both the large percentage of drugs for which patents have expired (or been successfully challenged) and the rapid conversion to generics once they become available. The much lower generic share of sales by value reflects the low generic prices, relative to originator prices. Compared to most other high income and some middle income countries, generic penetration is more rapid and generic prices are lower absolutely in the US (Danzon and Furukawa, 2006).

The high-generic volume share and low generic prices in the US relative to most other industrialized countries reflects several institutional features that interact to produce a highly price-competitive generic market in the US. First, the statutory rules governing generic entry are designed to reduce costs of entry and encourage patent challenges. Under the 1984 Hatch Waxman Act, generic versions of chemical drugs can file an Abbreviated New Drug Application (ANDA) with the Food and Drug Administration (FDA). By demonstrating bioequivalence to the originator drug, generics can be approved as substitutable for the originator, while simply referencing the originator's clinical trials to establish safety and efficacy, rather than doing new clinical trials. This dramatically reduced the cost and time required for generic approval. The Hatch-Waxman Act also incentivized generics to challenge patents, by offering a 180-day market exclusivity to the first completed ANDA that successfully challenges the originator's patents (a Paragraph IV filing). This FDA requirement, that generics demonstrate bioequivalence to the originator, is the basis for confidence on the part of physicians, consumers, and payers that generic substitution by pharmacists is safe. Although bioequivalence and substitutability apply to the great majority of small molecule drugs, certain compounds are considered too high risk or too difficult to characterize to permit safe substitutability.

Second, payers incentivize patients to accept generics by structuring formularies with low copayments (US$0–10) for generics, whereas the patient may have to pay the full price or a nonpreferred tier copayment for the originator. Third, the rapid uptake of generics reflects the rules and financial incentives for pharmacy substitution. The great majority of states have adopted the default rule that pharmacists may substitute any FDA-substitutable generic, even if the physician writes the script for the originator brand, unless the physician explicitly requires that the brand be dispensed. Thus the pharmacy decides which version of the compound to dispense, if substitutable generics are available and the physician does not require the brand. Payers incentivize pharmacies to prefer low-priced generics by reimbursing the same 'maximum allowable cost' (MAC) regardless of whether they dispense a generic or the brand. The MAC is similar to a reference price used in many other countries. Payers generally set the MAC at a relatively low generic price, though methodologies for setting and updating MACs vary. Because the pharmacy captures the margin between the MAC reimbursement and the acquisition cost of the drug, generic manufacturers compete by offering discounts on the acquisition cost to increase this margin. Over time, the payers revise the MACs downward to capture some of this discounting on generic prices, which leads to further price cutting by generics. Thus once multiple generics enter for a given drug, aggressive price competition and rapid generic erosion of brand sales occur, because pharmacy substitution is incentivized by MAC reimbursement and patient acceptance of generics is incentivized by low cost-sharing.

In a pharmacy-driven generics market such as the US, where generics are required by regulation to be bioequivalent and pharmacies are authorized and incentivized to substitute, generics have little incentive or possibility to use branding to create a perceived quality differential. Generics are therefore unbranded and compete on price and service to their highly-price conscious pharmacy customers. By contrast, in countries where generics are not required by regulation to be bioequivalent, which includes most middle and lower income countries, actual and perceived quality differences can play a big role in choice between supposedly similar products. In such regulatory regimes, there are often multiple 'similar' or 'copy' products that claim to have the same active ingredient as the originator brand, but there is no assurance that they in fact have exactly the same active ingredient and an equivalent therapeutic effect, quite aside from issues of substandard or counterfeit products. Given such intrinsic quality uncertainty, generic producers have strong incentives to sell branded generics, where brand becomes a proxy for quality. With quality uncertainty, physicians prescribe by brand – either the originator brand or a specific branded generic – and pharmacy substitution is not legally authorized, although it may not happen in practice. Branded generics are promoted and detailed to physicians, just like originator brands, which adds significant marketing costs. Generics also promote their brand to pharmacies in countries where pharmacies dispense without a prescription. In such branded generic markets, competition between generics focuses on brand as a proxy for quality, not price. On the contrary, branded generic prices are relatively high, compared to the originator price in that country and compared to US generic prices, in part because a low price might be interpreted as a proxy for low quality. In physician-driven, branded generic markets, originator brands usually retain a significant market share even after patent expiry, in contrast to the virtually complete originator brand erosion in the US.

Generic prices have traditionally been lower in the US than in major European markets, including Germany, France, and Italy, in contrast to onpatent brand prices which are higher in the US. But since the late 1990s most western European countries have changed their regulatory and reimbursement rules to permit and incentivize pharmacy substitution and generic price competition, which has increased savings to payers from generics. In Germany, since 2007 payers are authorized to contract directly with generic companies, using competitive tenders to drive and capture savings from price competition. So far, branded generics markets remain the norm in Latin America, Africa, and most Asian countries, including China. In such countries, generic quality is uncertain and some branded generic prices are relatively high. Some multinational originator companies are entering these relatively high-margin, branded generic markets through licensing arrangements with branded generic producers. This strategy draws on their brand selling expertise and brand image. However, because this physician-driven branded generic model delivers only modest savings to consumers and payers, compared to the pharmacy-driven unbranded, price-competitive generics model of the US, it seems likely that the US unbranded, price-competitive generic model will eventually become the norm in most countries.

The relatively high concentration on the purchaser side of the US generic market probably contributes to low generics prices. As discussed earlier, the US retail pharmacy market has become concentrated into large chains and large national wholesalers purchase on behalf of independent pharmacies. These concentrated buyers purchase a sufficient absolute volume and market share to have significant leverage in price negotiations with generic suppliers. Because these large purchasers are the decisionmakers for (intramolecule) generic substitution, generic suppliers target their discounts to them in the first instance, whereas originator companies target discounts on onpatent drugs to payers or PBMs who are the decisionmakers with respect to formulary design and (intermolecule) therapeutic substitution. The system relies on competition at the retail pharmacy and PBM levels to pass on these discounts on generics and onpatent brands to ultimate payers and consumers.

Competitive pressure on generic prices also depends on number of generic competitors. Price competition is weaker when there are few generic competitors, which occurs in at least two instances. First, the Hatch-Waxman Act intentionally grants a 180-day exclusivity to the first ANDA generic to successfully challenge all relevant patents. During those 180 days, the originator typically maintains or raises its price, but may also launch an authorized (licensed) generic to capture some of the more price sensitive market. During this period of at most two generics, their pricing is typically approximately 60–80% of the originator price. By contrast, once the exclusivity period expires, if multiple competing generics enter, generic prices fall rapidly to 10–20% of the preexpiry originator prices. Thus, the much higher price and margin during the exclusivity period creates an incentive for generic firms to incur the significant litigation costs and risks in challenging patents. Second, more complex formulations and specialty drugs with small markets typically attract fewer generic competitors than oral formulations with large markets. With fewer competitors, generic prices and margins can remain relatively high.

Brand Pricing after Patent Expiry

The evidence indicates that originator firms typically raise prices before and after patent expiry, rather than reduce price to compete with generics. One theory that accounts for such pricing is that the originator pursues a segmentation strategy (Frank and Salkever, 1992). In this model, before patent expiry, the originator selects the profit-maximizing price based on the weighted average of price elasticities of all customers in the market. After patent expiry, the more price elastic customers switch to generics, and the originator targets only the most brand-loyal, price-inelastic customers, which results in a higher, profit-maximizing price. This model was appropriate in the early 1990s, when many consumers in the US paid out-of-pocket for drugs and pharmacy substitution was less the norm. It is also useful in understanding price competition in self-pay, branded generic markets in emerging markets. However, in the current US context where most consumers have tiered insurance coverage and most states have pharmacy substitution, the price-inelastic market segment is very small. Moreover, during the 180-day exclusivity period, some originator firms have given

some payers sufficiently high discounts to get the originator drug placed on the generics tier, such that consumers have no reason to prefer generics. This strategy rarely continues beyond the 180-day exclusivity period, presumably because the discount required to compete on price with generics becomes too large.

A second rationale for originators to raise price before and after patent expiry is to encourage patients and payers to switch to a new formulation or a follow-on version of the drug that still enjoys patent protection or data exclusivity and therefore is not subject to generic competition. For example, before the patent-expiry on the standard twice-a-day tablet formulation, the firm may launch a once-a-day, timed release version of the drug that receives some market exclusivity for doing new clinical trials. By raising the per-day price of the old tablet such that it exceeds the price of the new delayed release formulation, the firm encourages payers and patients to prefer the new formulation. If the script is written for the exclusivity-protected formulation, pharmacies cannot substitute a generic because substitution is permitted only within the identical formulation. Such launch of a new formulation or product, together with price increase on the older formulation, is the most effective defense against generic erosion on a patent-expired drug in the US.

Biosimilars

In contrast to the price-competitiveness of generics for small molecule (chemical) drugs, biosimilar versions of biologics are unlikely to compete aggressively on price in the US, for several reasons. First, the higher clinical trial and manufacturing costs of biosimilars are expected to result in fewer competitors. Second, the greater complexity of biologics molecules means that the FDA is unlikely to declare them to be substitutable with the originator or with each other, except possibly for the simplest biologics. If pharmacies cannot substitute, physician-prescribers will be the decisionmakers. Thus biosimilars in the US will likely be branded products, detailed to physicians and marketed on brand rather than price, more like branded originator drugs than unbranded, price-competitive chemical generics. It is possible that payers may use tiered formularies, tiered cost-sharing, step edits, and prior authorizations to attempt to drive utilization toward preferred biosimilars, in which case they may extract significant discounts. However, this would be a departure from their current passive role with regard to use of biologics and other specialty drugs.

For biosimilars that require infusion in a physician office, the nonsubstitutable biosimilar would receive a different reimbursement code from the originator and have a separate ASP. Under current Part B reimbursement, this could discourage price competition by biosimilars, because reducing the price would reduce the physician's margin. The PPACA therefore provides that the 6% margin would be calculated on the originator's ASP, regardless of whether the originator or the biosimilar is dispensed. This eliminates any financial incentive for physicians to prefer the originator versus the biosimilar, but still provides at best weak incentives for biosimilars to compete on price.

As availability of clinically similar biologics increases, including both biosimilars and 'bio-betters,' payers may attempt to change reimbursement rules for specialty drugs in order to stimulate some price competition and value-based purchasing. Some US payers are requesting comparative effectiveness data and evidence of outcomes as a condition of favorable formulary placement. Some payers are also starting to adopt bundled payments for episodes of care that includes drugs. If providers are at risk for the cost of drugs as part of a care episode, they have strong incentives for cost-conscious choices with respect to volume and type of drugs. In addition to DRGs for inpatient care, Medicare has adopted bundled payment for dialysis, and at least one private payer uses bundled payment for certain episodes of cancer care. Therapeutic reference pricing, as used in Germany for certain classes of drugs, has also been mentioned in the US, but so far seems unlikely.

Discussion and Conclusions

The US reimbursement system for pharmaceuticals, which permits manufacturers to set prices freely and relies on patient cost-sharing and health plan bargaining to drive discounts off these prices, has been reasonably effective at constraining prices for drugs in crowded classes with clinically close substitutes. Once generic entry occurs, pharmacy substitution and reimbursement incentives assure low generic prices and rapid generic erosion. However, for specialty drugs – which typically have few close therapeutic substitutes and include many biologics – current insurance reimbursement and cost-sharing arrangements create incentives for manufacturers to set high prices with few constraints, especially for physician-dispensed drugs. Although patients nominally face very significant cost-sharing, such cost-sharing is ineffective at constraining manufacturer prices due to supplementary insurance, stop-loss limits, manufacturer coupons, and patient assistance programs. The ACA stop-loss limits will further reduce the elasticity of demand facing manufacturers for high-priced drugs. Stop-loss limits provide appropriate financial protection for patients but imply that other payer constraints on prices may be appropriate.

After double-digit growth rates in the late 1990s, the rate of growth of drug expenditures has moderated since the early 2000s in the US. This is due largely to savings from patent expirations and consequent generic erosion on many high-volume drugs, combined with a modest flow of new drugs that have generated insufficient new sales to replace sales lost to patent expiries. The resulting savings to payers and consumers have created budget headroom for higher prices on newly-launched drugs and substantial postlaunch price increases. However, this will change as the wave of patent expiries tapers off around 2015 and if the recent uptick in number of new drug approvals continues.

Thus, the combination of an increasing share of new drugs (biologics, orphan drugs, and other specialty drugs) for which traditional PBM/PDP-tiered formulary mechanisms work poorly, with more (and appropriate) stop-loss limits on patient cost-sharing under the ACA, make some change in reimbursement mechanisms increasingly likely. Although none of the ACA provisions relate directly to pharmaceutical

reimbursement, the inducements for bundled payments and outcomes-based reimbursement for hospital and physician providers may eventually spill over to pharmaceuticals. If payers require evidence of comparative and cost-effectiveness, as an input for making coverage and reimbursement decisions, this would incentivize manufacturers to set prices commensurate with incremental health benefit delivered. This form of flexible and indirect price constraint, that aligns prices with incremental health benefit, provides more appropriate incentives for R&D and for efficient use of drugs than either the status quo or alternative price control mechanisms that have been proposed.

See also: Biosimilars. Markets with Physician Dispensing. Patents and Regulatory Exclusivity in the USA. Pharmaceutical Marketing and Promotion. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Regulation of Safety, Efficacy, and Quality

References

Danzon, P. M. and Furukawa, M. F. (2006). Prices and availability of biopharmaceuticals: An international comparison. *Health Affairs* **25**(5), 1353–1362.

Duggan, M. G. and Scott Morton, F. (2006). The distortionary effects of government procurement: Evidence for Medicaid prescription drug purchasing. *Quarterly Journal of Economics* **121**(1), 1–30.

Epstein, A. J. and Johnon, S. J. (2012). Physician response to financial incentives when choosing drugs to treat breast cancer. *International Journal of Health Care Finance and Economics* **12**(4), 285–302.

Frank, R. G. and Salkever, D. S. (1992). Pricing, patent loss and the market for pharmaceuticals. *Southern Economic Journal* **59**, 165–179.

Scherer, F. M. (1997). How US antitrust can go astray: The brand name prescription drug litigation. *International Journal of the Economics of Business* **4**, 239–256.

Further Reading

Danzon, P. M. and Chao, L. W. (2000). Cross-national price differences for pharmaceuticals: How large, and why? *Journal of Health Economics* **19**, 159–195.

Danzon, P. M., Towse, A. K. and Mulcahy, A. (2011). Setting cost-effectiveness thresholds as a means to achieve appropriate drug prices in rich and poor countries. *Health Affairs* **30**(8), 1529–1538.

Duggan, M. G. and Scott Morton, F. (2012). The medium-term impact of Medicare Part D on pharmaceutical prices. *American Economic Review* **101**, 387–392.

Frank, R. G. and Salkever, D. S. (1997). Generic entry and the pricing of pharmaceuticals. *Journal of Economics & Management Strategy* **6**, 75–90.

Garber, A., Jones, C. I. and Romer, P. M. (2006). Insurance and Incentives for Medical Innovation. *Forum for Health Economics and Policy* **9**(2), Biomedical Research and the Economy, Article 4.

Hoffman, J. M., Li, E., Doloresco, F., et al. (2012). Projecting future drug expenditures. *American Journal of Health System Pharmacy* **69**(5), 405–421.

Jacobson, M., Earle, C., Price, M. and Newhouse, J. (2010). How Medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment. *Health Affairs* **29**(7), 1391–1399.

Lakdawalla, D. N. and Yin, W. (2010). Insurers' negotiating leverage and the external effects of medicare part D. *NBER Working Paper 16251*. Cambridge, MA: National Bureau of Economic Research.

Saha, A., Grabowski, H., Birnbaum, H., Greenberg, P. and Bizan, O. (2006). Generic competition in the US pharmaceutical industry. *International Journal of the Economics of Business* **13**, 15–38.

Scott Morton, F. (1997). The strategic response by pharamceutical firms to the Medicaid most-favored customer rules. *Rand Journal of Economics* **28**, 269–290.

US Congressional Budget Office (2005). *Prices for brand-name drugs under selected federal programs*. Washington, DC: Congressional Budget Office.

US General Accounting Office (1993). *Medicaid: Changes in drug prices paid by HMOs and hospitals since enactment of rebate provisions*. Washington, DC: US General Accounting.

Pricing and User Fees

P Dupas, Stanford University, Stanford, CA, USA

Glossary

Allocative efficiency This exists when a product or service is allocated in such a way as to maximize the health benefit of the population.

Placebo–price effect When a higher price for a product or service affects beliefs about its quality and this belief translates into higher actual experience of the product's efficacy.

Price-elasticity The responsiveness (elasticity), of the quantity demanded of a good or service to a change in its price.

Sunk cost When costs previously incurred cannot be retrieved or saved by ceasing production, they are said to be 'sunk'. They should be irrelevant to a rational calculating decision maker.

Introduction

Governments throughout the world intervene in the health sector. One motivation is that under Article 25 of the Universal Declaration of Human Rights access to adequate healthcare is a fundamental human right. A second motivation is that the health sector is subject to many market failures, due to, for example, consumption externalities, imperfect information, and imperfect credit. Consumption externalities exist when private consumption of health services yields positive or negative social returns. Immunization is a prime example: when individuals get immunized, the disease transmission rate falls, yielding increased protection for the whole population. In the case of imperfect information, people do not have all the information they need to make healthcare decisions. There may, for example, be information asymmetries: the patient typically has less information about her private returns to consuming a given health procedure or drug than the physician trying to sell that service or drug. In the case of imperfect credit, people might not be able to finance lumpy health investments that yield positive returns, such as preventive care. In the presence of these market failures, private consumption of health products and services is socially suboptimal. Thus, in the presence of positive externalities, immunization rates are too low; in the presence of information asymmetries that favor the provider who has incentives to oversell, the utilization rate for services or drugs is too high; and with imperfect credit investments in preventive products is too low. Governments intervene in the health sector to remedy these market failures and achieve the social optimum. They do so in four main ways: public provision of health services, subsidies for private provision, regulation of private provision, and public provision of information. In the first three cases, an important dimension of a government's intervention concerns the pricing of health products and services.

This article is concerned with the question of how to decide what price should be set or charged for what health service or product. How much should patients pay for health services at government clinics? How large should the subsidy be for preventive health products that exhibit positive externalities? What should be the maximum price that private medical practitioners are allowed to charge for primary care services? Although many of the points discussed here are directly relevant to the issue of optimal price controls, price regulation, which is covered in the previous article, will not be discussed. Instead, focus is on the issue of user fees (or user charges) for public health services. Public health services account for more than two-thirds of medical services provided in sub-Saharan Africa and between one-third and two-thirds in Southeast Asia. For our purposes, public health services will be broadly defined. They will include the delivery of subsidized health products, even if receipt of these products does not require a health professional. In other words, as defined here, public health services encompass both inpatient and outpatient medical care and the implementation of subsidies for vaccines, bed nets, antimalarial pills, and other privately produced health products.

User fees have implications for cost-effectiveness, allocative efficiency, equity, progressivity of public healthcare spending, and quality of service. Each of these is a desirable policy end in itself, and so each is an important factor in the optimal pricing decision. However, they are not always compatible with each other. Furthermore, they all have to be financed from a single, and typically constrained, budget. Thus governments have to tradeoff over them. Although the relative importance accorded to each factor will depend on the government's objectives, most are likely to place nonzero weights on most factors. In addition, these five factors have to be considered both for the product or service in question and for other products and services funded from the same budget. Each factor, then, has to be carefully analyzed when setting a pricing (user fee) system for public health services. This article reviews the theory and empirical evidence on the effects of user fees on each factor.

Cost-Effectiveness

A first reason to charge a user fee for a given health service or product may be to reduce the public cost per unit provided. This per-unit cost reduction would make it possible to increase the quantity provided within a given budget – that is, to

 doi:10.1016/B978-0-12-375678-7.00123-1

become more cost-effective. Two conditions are necessary for a user fee to reduce the unit cost.

User Fees and Administrative Costs

User fees can only improve cost-effectiveness if the administrative costs of collecting and managing the fee revenue are lower than the fee itself. This is an obvious enough condition, but one that is not always easy to satisfy. Record keeping at the point of service in many developing countries is done manually. This makes the aggregation of data needed for effective management time-consuming and difficult. In particular, ensuring that fees are properly collected and remitted in full can require either costly monitoring, or costly incentives for health workers, or both.

User Fees and Fixed Costs

When fixed costs are important, the impact of user fees on cost-effectiveness will depend on the price-elasticity of demand. Fixed costs for healthcare are often large: both the facility costs and staff costs must be paid whether patients use the facility or not. This means that a change in demand can have a substantial impact on the average cost. Imagine, for example, that 15 patients use a prenatal clinic daily when the price is zero, but only 9 when a user fee is charged. The fixed costs, for example, the salary of the prenatal nurse, would be the same for 9 patients as for 15. As such, the per patient cost of delivering prenatal care could actually be higher when fees are introduced and utilization rates decrease. A recent example of the effects of utilization rates on cost-effectives in the presence of important fixed effects comes from a randomized trial in Udaipur, India. When parents were offered in-kind incentives to use free immunization services, demand increased so much that costs per child immunized were halved compared with when they received free service alone. In other words, with higher demand fixed costs were spread over many more beneficiaries, so that the negative price was more cost-effective than a zero price.

The question of the price-elasticity of demand for health services gained prominence in the mid-1980s. After the Alma-Ata 'Health for All' Declaration was signed in 1978, which made access to basic healthcare a fundamental right, many countries in Africa implemented free primary healthcare. In the mid-1980s, however, it became apparent that free delivery was not financially sustainable, and in 1987 the World Health Organization, United Nations Children's Fund and a group of African health ministers launched the Bamako Initiative calling for self-financing mechanisms at the local level, including user fees, particularly, for drugs. The evidence available at the time was mixed and therefore controversial. Earlier studies, using cross-section variation in prices, had estimated that demand for healthcare was relatively price-inelastic. The price range over which the elasticity could be estimated was relatively narrow, making it difficult to gauge how truly price-sensitive people were. Later studies, using the introduction (or suspension) of user fees, found large drops (increases) in utilization in response to the policy change. For example, the suspension of user fees in Madagascar following a political crisis has been credited for a very large uptick in utilization rates. All in all, the empirical evidence so far suggests that policymakers should give close attention to potential reductions in utilization rates when considering user fees as an instrument for reducing the unit costs of public health goods.

Allocative Efficiency

Perhaps a second reason to charge user fees is to improve allocative efficiency – that is, ensure that the product or service is provided to those that actually need it the most. In particular, user fees may help prevent overutilization, that is, utilization by those for whom the conferred benefits, both private and social, are lower than costs of providing them the good. If user fees do prevent such waste, then the resulting reduction in demand would be desirable even when universal access to services is one of the objectives. There are three mechanisms through which user fees can act on allocative efficiency: screening effects, psychological effects, and moral hazard deterrence effects.

User Fees Can Improve Allocative Efficiency

First, by screening out those who do not value the product or service enough to pay for it. This allocative role of market prices is a standard tenet of price theory. Households, just like governments, are budget-constrained. They, too, are unlikely to invest in a health good – be it prenatal care or a water purification product – if the expected benefits are lower than the costs, both the monetary costs and the time costs. Persons with a simple cold, for example, is less likely to be interested seeking care if it will cost money or it will take 2 h of their day. Instead if they have severe malaria, they will probably want to make the investment in medical care, even at a high cost. In the end, whether or not they seek care and what kind of care they seek will depend on the prices they face. In other words, the prices determine the allocation of public health goods. As it is theoretically based on the user's valuation of the good, this priced-determined allocation is efficient.

The allocative efficiency of prices breaks down under three market imperfections – externalities, imperfect credit, and imperfect information. First, when there are positive externalities to private consumption, private consumption should be subsidized up to the social value to ensure that those for whom the private value is lower than the market price (but the sum of the social and private value is higher than the price) still invest in the good.

Second, when credit markets are not perfect, people cannot borrow to invest in goods that yield positive returns. Limited access to credit means that people may not have the ability to pay for their full valuation of the good; ability to pay and willingness to pay can then become disjoint. User fees could in such cases bar access to some people for whom the health returns to the health good would be high, but who are too poor to pay for it.

Third, when information is imperfect people may not know exactly the private value of the good. For example, they may not have the information or the ability to process

available information; they would then not be able to assess how much they would benefit from, and so should want to pay for, a given medical procedure or product. For some goods, they may only be able to acquire the information by first trying the good. In these cases of imperfect information, user fees could screen out those that do not know they need the product, preventing them from ever learning that the good could benefit them. However, in the presence of imperfect information, how much a service or product is sold for may be interpreted as a signal of its quality. If so, setting user fees too low could discourage usage by setting too low expectations about the quality being provided. In a randomized study in urban Zambia, researchers find that offering a new, unknown water purification product at too low a price dampened demand for the product compared with a higher-priced, well-known product. This signaling effect of prices can be mitigated by information provision; however, in the Zambia study, accompanying the subsidy with a marketing message that informed customers that the new product was as effective as the well-known product led to higher demand at subsidized prices. Likewise, information that the user cost is subsidized might be sufficient to ensure that low prices are not taken as a signal of low quality (i.e., if people infer something about the value of a service from the extent to which it is subsidized).

Evidence from recent randomized experiments suggests that the extent to which user fees can improve allocative efficiency depends on the context as well as the good. In another randomized study in urban Zambia, researchers randomized the fee charged for a chlorine-based water purification product. They found that higher fees screen out households that would not benefit health wise, because, for instance, they would use the product for house cleaning rather that water purification. In rural Kenya, researchers randomized the fees charged for artemisinin-combination therapy (ACT), the latest class of antimalarial drugs, and found that higher fees increase the likelihood that the ACTs are bought by those with a verified case of malaria. However, when they randomized the fee charged for antimalarial bed nets, also in rural Kenya, they found the opposite result: higher fees significantly reduced demand by screening out those who value the product and would use it efficiently if they got it for free but cannot afford to pay for it. A third study in rural Kenya found that the reduction in demand associated with high fees can prevent households from learning the true private value of antimalarial bed nets, which dampens future willingness to pay.

Given these dynamic learning effects, and the pervasiveness of credit constraints, a potential alternative to setting user fees is to use nonmonetary costs as an allocative mechanism. Studying the adoption of a water chlorination product similar to those in the Zambia studies discussed earlier, a study in Kenya shows that compared with simply handing out the product for free at clinics, distributing free vouchers redeemable at a local store can improve allocative efficiency. This is because the transaction cost of going to the store to redeem the coupon, though small, seems to be enough of a deterrent to dissuade people from picking up a chlorination product they will not use, while not discouraging those who will actually use the product.

User Fees Could Improve Allocative Efficiency

Second, by way of the psychological effects of prices, including the sunk cost fallacy and placebo–price effects. The effectiveness of some health goods will depend on the behavior – compliance – of the user. For example, the effects of an iron supplementation regimen on anemia depend on the behavior of the recipients. If they do not comply with the regimen, say taking a pill once a week instead of once a day, the treatment will not work as well. The same goes for a bed net; if it is not hung up or used, it will not protect anyone from malaria. Some services also have the same property. If a pregnant woman does not listen to the nurse during her prenatal care visit, she might not learn enough to benefit from the visit. For such goods, user fees might help induce the complementary behavior required for full effectiveness. They could do so through two psychological channels – the sunk cost fallacy and placebo–price effects

First, the sunk costs fallacy. When this fallacy is operative, the higher the price paid for a good, the higher the likelihood that it is used to its full potential. This is because the buyers want to avoid feeling that they wasted money. People, it seems, do not recognize when they should consider costs incurred in the past as sunk costs. Studies in the US have found evidence of sunk cost fallacy effects for entertainment products. It is possible that such effects could apply for health products, with, for example, people who pay more likely to comply with an expensive course of treatment or more likely to use a bed net. Two of the randomized pricing studies discussed in Section User Fees Can Improve Allocative Efficiency were specifically designed to test for sunk costs fallacy effects for health products, one in urban Zambia and one in rural Kenya. However, neither found evidence for such effects.

Second, placebo–price effects. In this case, paying a higher price increases the psychological investment of the user, boosting effectiveness. Thus, it was found that people who are charged full price for a drink supposed to boost mental acuity perform better on mental tasks than those who are told they had received a price discount. Whether such placebo–price effects are at play for public health goods remain to be directly tested. The evidence from Zambia and Kenya mentioned in Section User Fees Can Improve Allocative Efficiency indirectly suggests that such placebo–price effects are not large enough to boost usage of water chlorination products or bed nets, but they could increase the effectiveness of medication for mental health, for example. There is no evidence to date on this issue.

User Fees Can Improve Allocative Efficiency

Third, by deterring *ex ante* moral hazard. If health goods are costly, people are more motivated to stay healthy. Thus, when treatment for injury is expensive and out of pocket, people would be more motivated to avoid injuries than when treatment is free. More to the point, people would have a higher incentive to invest in preventive goods if curative care is costly. Note, however, that this argument can be used to motivate larger fees for curative services, but not for preventive care. For preventive care, the argument is exactly the opposite; user fees would reduce preventive investments, leading to higher demand for curative care in the future, potentially increasing

total healthcare costs. Evidence on the importance of *ex ante* moral hazard in the context of a developing country is rare, probably because it is considered unlikely or minimal. Looking at the impact of introducing health insurance for informal workers in Nicaragua, a study found no evidence of moral hazard behavior.

Equity

Improving equity in access to healthcare (if not equity in health) is one of the objectives of most governments. The impact of user fees on equity in access depends on the price-elasticity of the demand for health services and products, and how it varies by socioeconomic status.

As discussed in Sections User Fees and Fixed Costs and User Fees Can Improve Allocative Efficiency, there is a large literature on the price-elasticity of demand for health goods. The price-elasticity of the demand for preventive care has recently received a lot of attention, and the evidence from randomized field experiments suggests quite a large price sensitivity in a number of settings. The evidence on the price-elasticity of the demand for curative care is somewhat mixed and for the most part imperfectly estimated, but overall it suggests that user fees tend to compromise access, especially among the poor, who tend to be more price-sensitive than others.

An obvious way to amend a user fee system to foster equity is to price discriminate, that is, charge the poor less than the rich for a given health service or product, for example, through the distribution of vouchers. This is not always easy to do in practice. Most of the poor in developing countries are subsistence farmers or employed (often self-employed) in an informal business. This means that they are not part of the tax base; there is thus no record of their earnings, which makes it difficult to identify who should be eligible for the lower fee/voucher.

Progressivity of Public Health Spending – the Redistributive Implications of User Fees

A related issue is that of redistribution. Redistribution is often a health policy objective. In that case, public health services are an integral part of poverty alleviation efforts. An important consideration when setting user fees is their impact on who benefits from public health spending, or benefit incidence.

Higher user fees for health goods can make public spending regressive as they disproportionately affect the poor. If user fees are set below the average cost but remain substantial enough that they reduce the demand for public health services proportionately more for the poor than for the rich, then they would make public health spending regressive: benefits would accrue disproportionately to the rich. Even if user fees do not reduce health service utilization among the poor, they could have negative redistributive implications through negative cross-price-elasticities. The more the poor have to pay for their healthcare, the less money they have left to invest in, say, education. If user fees for health reduce enrollment in public schools among the poor, that might undermine the goal of primary education for all, another objective common to most governments of developing countries.

Although, to the best of our knowledge, there is no evidence on cross-price elasticities for publicly provided services, there is, as discussed in Section User Fees Can Improve Allocative Efficiency, a large literature on the price-elasticity of demand for health goods, and this literature suggests that the poor are much more price-sensitive than others, which would imply that user fees are likely regressive. A study looking at public spending on curative healthcare in seven sub-Saharan countries found that public health spending is disproportionately benefiting the less poor, consistent with the price-elasticity literature. The richest 20% receive much more than 20% of public health subsidies, whereas the poorest 20% receive less than 20%. This is because a large fraction of public health subsidies go to services that the poor do not use, such as hospital care, which the poor do not access because they typically live far from any hospital. A recent review of the evidence compiled in the 2004 World Development Report shows that this phenomenon is not limited to Africa.

A potential solution to ensure that public health spending is targeted at the poor is, here again, to use the strategy of price discrimination. Although price discrimination for a given product can be difficult when identifying the poor is itself difficult, an alternative is to charge high fees for products and services that only the rich demand/use, and low or no fees for the products and services used primarily by the poor. That would mean, for example, charging high fees for hospital care and low fees for care at primary facilities; or if there is geographic segregation, charging higher fees in richer areas and lower fees in poorer areas.

Quality of Service

The last factor to consider in the pricing decision is that of the quality of the healthcare received by the population. This means considering both the quality of those health services subsidized or provided by the government as well as the quality of the alternatives that people would have to resort to if user fees deter them from accessing public services.

User Fees and the Quality of the Services for Which a Fee Is Charged

User fees can have a direct positive impact on the quality of the services for which they are charged if the revenue they generate is retained by the local facility charging them, and used locally. This can come about through two main mechanisms. First, the user fees can finance quality improvements such as maintenance or renewal of the equipment or the facility or in-service training for health workers. Second, the revenue from the user fees can be used to incentivize health workers: if health workers can pocket the user fees, they have a higher incentive to be present and serve than if their payoff function is flat. However, pay-for-service can lead to overprovision of services, that is, moral hazard on the part of the provider. There is, to the best of our knowledge, no rigorous evidence to date on these issues in the context of a developing country. It is therefore not known how potential quality improvements attained through user fees would compare with

direct investments in quality by the government, such as incentive pay systems paid out of general revenue. There is, however, some evidence from the private sector suggesting that the margins that can be made by providers on health products are so low that the incentives effect is almost nonexistent. In a randomized trial in Zambia, it was found that nonfinancial rewards (e.g., social recognition) for agents selling condoms are more effective than allowing the agents to keep a margin on their sales. Please also see the article by Miller and Babiaraz in this section of the Encyclopedia for a more detailed discussion of pay for performance considerations in developing countries.

Even if the revenue from user fees is not used to directly finance quality improvements, user fees could impact quality indirectly. One can think of two such potential indirect effects. First, the total revenue raised in user fees by a given health facility could be interpreted by the central authority as a signal of the quality of the services this facility provides. Indeed, it has been shown that demand is responsive to quality levels. The government could then allocate quality-enhancing projects based on this measure of quality, or use it as a way to monitor the local providers. Second, user fees might provide incentives for users to monitor their local providers and to demand better care: if they have to pay for the service, they have an incentive to demand high quality to ensure they get their money's worth. This argument was put forth quite forcefully by the 2004 World Bank Development Report titled 'Making Services Work for Poor People.' It is not clear, however, that users can easily judge the quality of the services they receive. One study shows that, despite extremely low quality of the healthcare they are getting, and their poor resulting health status, people in Udaipur (India) are quite satisfied with their own health and the services they receive. As such, community monitoring of local health providers might require information provision, such as through report cards, even in the presence of user fees. Such report cards were found very effective in Uganda.

User Fees and Health Outcomes

Even if user fees can enhance the quality of the services for which fees are charged, the quality change might not translate into better health outcomes for the population if user fees reduce utilization of those services and divert people to private alternatives of low quality, such as private practitioners with dubious qualifications, or self-treatment. It is therefore critical to know the price and quality of the alternatives available to people, as well as the likely impacts of a change in public sector fees, in order to fully assess the ultimate impact of user fees on the quality of healthcare that is received.

The effects of users on quality may be dynamic. An example is that of pricing for antimalarial drugs. Artemisinin-based therapies now constitute the only treatment effective against *Plasmodium falciparum* in Africa, where parasite resistance to earlier generations of antimalarials is widespread. Monotherapies are cheaper to produce than combination therapies (which combine an artemisinin derivative with a partner drug), and therefore favored by consumers. However, the use of monotherapies is suboptimal from a social standpoint as it contributes to faster resistance development to artemisinin. This means that high fees for combination therapies today may lead to a lower drug quality in the future, if they deter demand and instead lead patients to purchase monotherapies from the private sector. Here again, considering cross-price elasticities is thus critical when determining optimal pricing strategies.

Conclusion

Governments intervene in the healthcare sector primarily to improve health outcomes. However, their ability to intervene is limited by a budget constraint. This means that optimal pricing for public health services has to strike a delicate balance: it has to minimize the likelihood that needy persons do not access the health products or services that could benefit them, while also minimizing the likelihood that these products and services are used by those for whom the returns are low. The critical parameters to take into consideration when setting a price or user fee are thus price-elasticities: the price-elasticity of the demand for the health product or service under consideration, and how it varies with income and health status; and also the cross-price elasticities of other human capital investments that the government might care about.

Overall, the empirical evidence suggests that the price-elasticity of the demand for health products and services is relatively important in developing countries, but often not because of frivolous demand at low prices – rather, because of underutilization at high prices. This suggests that in many cases the introduction of user fees might need to be paired with exemptions for the poor in order to achieve the objectives of improving aggregate health outcomes and equity of access. The question then becomes one of cost-effectiveness: if running a scheme of user fees with exemptions is costly to administer, it might be much simpler and no more costly to have a blanket no-fee policy.

Identifying what price is optimal for a given service, drug or product, given the local context and given the objective function, is not necessarily simple, as the discussion above highlighted. Even once it has been identified, implementing the chosen price schedule is not necessarily that simple either. Providers at public health facilities might demand under-the-counter payments from clients for drugs, services, and other products on top of the set user fee. The importance of this type of petty corruption – which could undermine even the most carefully designed and progressive user fee system – are a part of a separate but related and extremely important theme on which research has been and is currently being performed.

See also: Efficiency and Equity in Health: Philosophical Considerations. Health Care Demand, Empirical Determinants of. Health Services in Low- and Middle-Income Countries: Financing, Payment, and Provision. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance

Experiment. Rationing of Demand. Resource Allocation Funding Formulae, Efficiency of. Willingness to Pay for Health

Further Reading

Ashraf, N., Berry, J. and Shapiro, J. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review* **100**, 2383–2413.

Ashraf, N., Kelsey, B. J. and Kamenica, E. (2013). Information and subsidies: Complements or substitutes? *Journal of Economic Behavior and Organization* **88**, 133–139.

Cohen, J. and Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Quarterly Journal of Economics* **125**(1), 1–45.

Cohen, J., Dupas, P. and Schaner, S. (2012). Price subsidies, diagnostic tests, and targeting of malaria treatment: Evidence from a randomized controlled trial. *NBER Working Paper 17943*. Cambridge, MA: National Bureau of Economic Research.

Dupas, P. (2009). What matters (and what does not) in households' decision to invest in malaria prevention? *American Economic Review* **99**, 224–230.

Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics* **3**, 425–449.

Dupas, P. (2012). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *NBER Working Paper 16298*. Cambridge, MA: National Bureau of Economic Research.

Gertler, P. and Hammer, J. (1997). Strategies for pricing publicly provided health services. *Policy Research Working Paper 1762*. Washington, DC: Wold Bank.

Jack, W. (1999). *Principles of health economics for developing countries*. Washington, DC: World Bank.

Jameel Abdul Latif Jameel Poverty Action Lab (2011). The price is wrong. *J-PAL Bulletin*. Available at: http://www.povertyactionlab.org/publication/the-price-is-wrong (accessed 17.06.13).

Mwabu, G. (1997). User charges for health care: A review of the underlying theory and assumptions. *Working Paper 127*. New York: The United Nations University, World Institute for Development Economics Research.

World Bank (2003). *World development report 2004: Making services work for poor people*. Washington, DC: World Bank.

Primary Care, Gatekeeping, and Incentives

I Jelovac, University of Lyon, Lyon, France, and CNRS, Ecully, France

Glossary

Capitation Payment arrangement for health care service providers that pays a physician or group of physicians a set amount for each enrolled person assigned to them, per period of time, whether or not that person seeks care.

Fundholding (in the National Health Service in Britain) System enabling general practitioners to receive a fixed budget from which to pay for primary care, drugs, and non-urgent hospital treatment for patients.

Gatekeeping Feature of many health care systems according to which a patient can access specialized care only after a PCP has issued a referral.

Managed-care Program intended to reduce unnecessary health care costs through a variety of mechanisms.

Patient (self-) selection Selection of a health care organization by patients. A consequence of patient (self-) selection is that, when both gatekeeping and free access systems coexist, gatekeeping is expected to attract individuals who are healthier on average than the free access does.

Primary Care Provider (self-) selection Selection of a remuneration system by PCPs. A consequence of PCP (self-) selection is that, when several remuneration methods coexist, each one is expected to attract PCPs according to their personal characteristics.

Referral Transfer of care for a patient from one clinician to another.

Third-party payer Organization other than the patient or the health care provider involved in the financing of personal health services.

Introduction

In its World Health Report 2008, the World Health Organization (WHO) advocates in favor of a central role for primary care in health care systems. The WHO defines the specific features that should characterize primary care to ensure improved health and social outcomes: person-centeredness, continuity, comprehensiveness, and integration. Person-centeredness is about adapting medical advice to individual life circumstances. Continuity would allow the best use and sharing of information between individuals and primary care providers (PCPs). The concepts of comprehensiveness and integration stress the multiple roles of a PCP: health promotion and prevention, diagnosis and treatment or referral, and chronic or long-term care. As for the organization of health systems, the WHO promotes to switch the entry point to the health system from hospitals and specialists to PCPs.

Evidence has been reported about favorable medical outcomes in systems with an emphasis on primary health care. For instance, continuity of care contributes to lower all-cause mortality. Person-centeredness is responsible for an improved quality of life and increased treatment compliance. Comprehensiveness contributes to better health outcomes and to fewer patients admitted for preventable complications of chronic conditions. The effect of the supply of primary care is not so clear-cut. A superficial observation negatively relates the number of PCPs per capita to the mortality rates. However, more PCPs are expected to work in areas with a worse case-mix of patients. Accounting for this simultaneity effect, the negative relationship between the number of PCPs and the mortality rates disappears.

On top of the evaluation of the benefits of primary care, the contribution of health economics to the trend toward strengthening primary care has been taking the following forms: to analyze possible organizations to bring primary care upfront; and to think of ways to make individuals and health care providers adhere to the aim of strengthening primary care. A lot of attention has been devoted to the gatekeeping role of PCPs and to the incentives of PCPs and patients to adequately use primary care as an entry point to the health care system. The selection of patients and PCPs into different primary care organizations is also an interesting issue even though it has been less debated so far. The notions of primary care, gatekeeping, incentives, and selection are therefore the core of this article. The aim is to understand each of these aspects independently from each other as well as their interactions with each other.

The remaining of the article is structured as follows. The following section discusses gatekeeping versus direct access to specialists. The next section reports on patients' incentives to use primary care as an entry point to the health system. It also analyzes the PCPs' incentives to fulfill their roles. The next section tackles the issue of selection that appears when several organizations for providing primary care or accessing specialized care coexist. The penultimate section extends the discussion to the supply of specialized care. The final section concludes the article.

Gatekeeping

The most obvious way to bring primary care upfront is to forbid patients' direct access to specialists. The PCP is thereby empowered with a gatekeeping role. Patients can access specialized care only after the PCP has issued a referral. The WHO has stressed the importance of the gatekeeping system as an organizational model to structure health care. Gatekeeping is typical of the health care systems in Denmark, Finland,

 doi:10.1016/B978-0-12-375678-7.01311-0

Ireland, Italy, the Netherlands, Norway, Portugal, Spain, and the UK; whereas Austria, Belgium, France, Germany, Greece, Iceland, Luxembourg, Sweden, and Switzerland allow free access to most medical specialists.

Empirical comparisons between gatekeeping systems and systems with free access to specialists repeatedly report the following three effects. Gatekeeping decreases patients' satisfaction, even though it earns a better acceptance in countries where specialists are in short supply as in the UK. Also, gatekeeping is significantly associated with a lower utilization of health services and lower expenditures.

To appreciate the influence of gatekeeping on the utilization of medical services and on the resulting expenditure, it is important to understand the possible relationships between gatekeeping, medical utilization, and medical expenses. Gatekeeping is primarily meant to limit the use of expensive specialist services to the necessary cases only and to avoid them for patients needing primary care only. Therefore, a decrease in utilization and expenses can reflect an efficient use of medical services only if it decreases unnecessary visits to specialists. Empirical evidence on unnecessary care under free access to specialists is therefore needed to support this relationship; otherwise it is admitted to think that gatekeeping can cause a decrease in necessary specialized care too.

Another aspect of the relationship between gatekeeping versus free access, utilization, and expenses is selection. Gatekeeping in the public system coexists with free access in the private sector in countries such as Spain and the UK whereas they coexist in the private sector in Switzerland and in the USA. When both gatekeeping and free access systems coexist, the authors expect gatekeeping to attract members who are healthier on average than the free access system does. This selection process would automatically result in lower medical utilization and expenses for the gatekeeping system, independently of a possible gain in efficiency. Limited evidence is available about the existing efficiency effect, once the selection bias is accounted for.

The effects of gatekeeping versus free access are also dependent on the financial incentives they are associated with. For example, gatekeeping is often associated with PCPs' financial incentives to limit referrals to specialists, whereas system with free access provides generally little incentives of this kind. Therefore, the lower medical utilization and costs observed in gatekeeping systems might be due to the financial incentives rather than to the gatekeeping barrier itself. The empirical literature on gatekeeping versus direct access to specialized care so far has not disentangled the effect of both patients' and PCPs' financial incentives from the effects of constrained access to specialists.

Some theoretical arguments help comparing gatekeeping with direct access considering the optimal provision of incentives to PCPs. The next section discusses how to provide adequate incentives to PCPs. At this stage and for the sake of comparison between gatekeeping and free access, it can be mentioned that incentives to PCPs are meant to minimize two types of possible errors: the use of specialized services when unnecessary (type-I error) and the lack of specialized treatment when necessary (type-II error). Without adequate incentives to PCPs, gatekeeping is expected to generate more type-II errors than type-I errors. Conversely, free access may result in more type-I than type-II errors. With adequate (though costly) incentives to PCPs, one can minimize both types of errors in a gatekeeping system because decisions are in the hands of the PCP. However, type-I errors would remain in a free access system because those are independent of the PCPs' decisions. Therefore, when optimal incentives are provided, gatekeeping performs better, in theory. Conversely, free access might perform better when the patients' pressure to refer anyway is high or when the quality of the patients' self-health information is either highly accurate (in which case the patients' self-referral is very efficient) or weakly accurate (in which case the PCPs' financial incentives are very costly).

Incentives

Incentive mechanisms are increasingly popular in the health care sector to deal with the inefficiencies caused by asymmetric information between physicians, patients, and third-party payers. Incentive mechanisms exist for both patients and physicians to encourage the adequate utilization of health care services. The incentives for the efficient utilization of primary care versus specialists is discussed here, starting with the incentives directed to patients and following on with those directed to PCPs.

In a free access system, patients can be motivated to access PCPs first by making them bear an additional out-of-pocket payment when they choose to directly visit a specialist. This system is in use in France since 2005 as well as in some health maintenance organizations in the USA. It is a soft version of the gatekeeping system. However, financial incentives directed to patients bring the issue of equity in the access to health services, which was absent from the pure gatekeeping system. The financial incentives directed to patients may also limit the scope for quality (non-price) competition between specialists. Indeed, if the patients' co-payment is proportional to a specialist's fee, the best reply of the specialist to an increase in patients' co-payment is to decrease his fee to sustain demand. The specialist earns thereby less revenue per patient, which results in lower incentives to invest in costly quality. However, for these financial incentives to prove efficient in terms of utilization of primary versus specialized care, empirical evidence is needed about the actual behavior of patients with and without out-of-pocket payments. The French experience with out-of-pocket payments has proven disappointing because direct access to a specialist was very limited even when no such payments were due.

As for the PCPs, incentives are generally provided through their remuneration scheme and they are expected to influence their referring behavior. The aim is to limit the discretionary and unnecessary referrals to specialists. Many authors have written about the incentive properties of the most traditional payment schemes. If one ignores the issue of referrals, the classical analysis of traditional payment systems yields the following conclusions. Fee-for-service (FFS) payments may encourage physicians to provide too many medical services to maximize their revenue. Capitation may lead physicians to limit either the amount or the quantity of the medical services they provide. FFS can thus be responsible for excessive health care costs and utilization, whereas capitation can be responsible for

low quality/amount of care. Salaried doctors have an incentive to minimize their effort during the consultation because they receive the same income irrespective of this effort.

However, some of the aforementioned incentives can be reversed in the case of PCPs who can refer patients to specialized care. PCPs paid by capitation can save on personal costs by simply referring their patients to expensive specialized care. PCPs' altruism reinforces this effect because referrals allow for own cost minimization without prejudice to the patient.

Salary may bring the same incentives as capitation regarding referrals. To minimize their effort during consultation, salaried doctors have an incentive to refer more often than needed. This is even more so if one considers that PCPs derive utility from the well-being of their patients, because referrals do not harm the patients.

PCPs paid on an FFS basis earn more revenue when treating the patients on their own rather than referring them to a specialist. In that case, FFS would lead to lower total costs and quality compared to capitation, if it is supposed that the costs and the quality of specialized care are higher than those of PCPs' care and it can be abstracted from professional duty considerations.

In theory, fundholding shares the same incentives as FFS regarding referrals. Fundholding enables PCPs to receive a fixed budget from which to pay for primary care, drugs, and non-urgent hospital treatment for patients. It has been used in the UK between 1991 and 1999 and reintroduced in 2005. Again, if non-urgent hospital treatment is more expensive than primary care, PCPs have an incentive to limit referrals.

Empirical studies clearly confirm the positive theoretical relationship between capitation and the number of (unnecessary) referrals. Empirical support also exists for associating fundholding with a lower rate of referrals.

As many professionals, physicians might actually be heterogeneous in the way they respond to financial incentives because they are actually heterogeneous in both the ability and sense of professional duty. Under FFS or fundholding, it can be expected that very altruistic yet not very able PCPs refer all patients to specialists. PCPs who are relatively altruistic and very able might decide to either treat or refer according to their diagnostic. The very selfish yet able PCPs might treat all their patients to either maximize their earnings under FFS or minimize their expected expenditure under fundholding. Empirical evidence is needed to eventually confirm these theoretical predictions.

Selection

Incentives need not be uniform for a given population within a health care system. For instance, physicians in the US can work either in a traditional FFS setting or in a managed-care organization with a capitation arrangement. Primary care practices in the UK in the beginning of the 1990s had the choice to adopt the fundholding scheme or not. PCPs in France can voluntarily participate in the Contract for Improving Individual Practice (CAPI) scheme, which pays PCPs a performance payment for satisfying guidelines on prescription and prevention behavior. On the demand side, patients in the US can enroll into either gatekeeping health plans or plans allowing direct access to specialists.

In theory, the ability of selecting one or another type of organization may result in a pooling of individuals with the same profile in each organization type, which potentially results in increased inequalities. It can also increase efficiency, thanks to a better match between individuals and organizations.

Allowing PCPs to select between FFS or capitation can be optimal if savings on specialists' costs are not the main concern of a regulator. Otherwise, all PCPs should be paid on an FFS basis to avoid the incentive that associates capitation to excessive expensive referrals. Limited empirical evidence exists about either the existence or the lack of PCP selection into one or another plan. The French experience with the CAPI system shows no PCP selection according to their profile. Conversely, there is some evidence of British selection concerning groups of PCPs enrolling in the fundholding system in 1991. Evidence about the effects of selection is not available so far.

On the patients' side, patients enrolling into gatekeeping health plans are expected to be less likely to see a specialist than are others in plans with unrestricted access to specialists. There is significant evidence of selection into plans with gatekeeper and/or network selection in the US. Self-selection occurs because individuals, possessing knowledge of their own health attributes and economic constraints, select plans accordingly. These attributes that partly determine the individual's choice of health plans also affect their expected utilization of services. Individuals in plans that require sign-ups with a PCP have more visits to nonphysician providers of care, more surgeries, and hospital stays but substantially fewer emergency room visits.

To sum up, there are three channels through which the choice of remuneration scheme may affect PCPs' output or productivity: First, certain kinds of behavior may be encouraged by the scheme itself (the incentive effect); second, certain kinds of physicians may be attracted to certain types of physician practices, which, in turn, are influenced by the remuneration scheme (the physician selection effect); third, certain kinds of patients may be attracted to certain types of physician practices, which, in turn, are influenced by the remuneration scheme (the patient selection effect).

The Supply of Specialized Care

Gatekeeping systems have developed in countries with a limited supply of specialists, as in the UK. There is also empirical evidence that the supply of specialists is an important system determinant of referrals. Therefore, controlling the market for specialists might help improving the organization of primary care. For instance, in health systems with lots of PCPs and few specialists per medical discipline, specialists enjoy in theory a high level of monopoly power eventually leading to high fees. Therefore, increasing PCPs' qualification may decrease the monopoly power of specialists.

Increasing the number of PCPs makes the primary-care market more competitive too. Together with a capitation payment for a patient-list system, more PCPs may experience a patient shortage from the more intense competition. There exist empirical evidence that this may lead to more referrals.

Intuitively, against more competition, PCPs refer patients more often, responding positively to patient requests. Therefore, the cost-saving effect of the substitution of specialists by PCPs may be weakened by the PCPs' reactions.

Efficiency gains that are usually attributed to gatekeeping cannot be taken for granted. In the short run, better matches between patients and specialists may lead to efficiency gains. However, in the long run, specialists have an incentive to adjust their specialization so that differentiation between specialists increases. This would increase the monopoly power of specialists, which might counteract the positive short run effect.

Conclusion

This article has discussed issues related to the organization of the primary-care sector. The following important relationships have been reported. Gatekeeping arrangements result in lower expenditures and utilization of health care services although no significant effect has been proven on health care outcomes. Concerning PCPs' incentives, capitation is associated with an increase in referrals to specialists, whereas the fundholding scheme seems to limit these expensive referrals. Concerning the choice of a health plan, patients opting for a gatekeeping plan are less likely to see a specialist than are others in plans that allow direct access to specialists. However, when regulating the primary care sector, it is important to anticipate its consequences on the behavior of specialists.

France is now witnessing a movement from solo PCP practice to group practice. This change may result in new incentives and behaviors. An opportunity for relevant research appears there, to follow on the recent interest of economists for group practice and norms.

See also: Access and Health Insurance. Competition on the Hospital Sector. Efficiency and Equity in Health: Philosophical Considerations. Health Insurance and Health. Income Gap across Physician Specialties in the USA. Organizational Economics and Physician Practices. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Physician Market. Physician-Induced Demand. Physicians' Simultaneous Practice in the Public and Private Sectors. Preferred Provider Market. Pricing and User Fees. Sample Selection Bias in Health Econometric Models. Specialists

Further Reading

Aakvik, A. and Holmas, T. H. (2006). Access to primary health care and health outcomes: the relationships between GP characteristics and mortality rates. *Journal of Health Economics* **25**(6), 1139–1153.

Allard M., Jelovac, I. and Léger, P. T. (2010). Physicians selection of a payment mechanism: capitation versus fee-for-service. GATE Lyon St Etienne, Working Paper 1024.

Allard, M., Jelovac, I. and Léger, P. T. (2011). Treatment and referral decisions under different physician payment mechanisms. *Journal of Health Economics* **30**(5), 880–893.

Brekke, K. R., Nuscheler, R. and Straume, O. R. (2007). Gatekeeping in health care. *Journal of Health Economics* **26**(1), 149–170.

Deb, P. and Trivedi, P. K. (2009). Provider network and primary-care signups: do they restrict the use of medical services? *Health Economics* **18**, 1361–1380.

Devlin, R. A. and Sarma, S. (2008). Do physician remuneration schemes matter? The case of Canadian family physicians. *Journal of Health Economics* **27**, 1168–1181.

Forrest, C. B. (2003). Primary care gatekeeping and referrals: effective filter or failed experiment? *British Medical Journal* **326**, 692–695.

Forrest, C. B., Nutting, P. A., von Schrader, S., Rohde, C. and Starfield, B. (2006). Primary care physician specialty referral decision making: patient, physician, and health care system determinants. *Medical Decision Making* **26**, 76–85.

Garcia Mariñoso, B. and Jelovac, I. (2003). GP's payment contracts and their referral practice. *Journal of Health Economics* **22**(4), 617–635.

Gonzalez, P. (2010). Gatekeeping versus direct-access when patient information matters. *Health Economics* **19**, 730–754.

Iversen, T. and Ma, A. (2011). Market conditions and general practitioners' referrals. *International Journal of Health Care Finance and Economics* **11**, 245–265.

Luras, H. (2004). General practice: four empirical essays on GP behavior and individuals preferences for GPs. Thesis for doctoral dissertation, Department of Economics, University of Oslo, Working Paper 2004: 1.

Malcomson, J. M. (2004). Health service gatekeepers. *RAND Journal of Economics* **35**(2), 401–421.

Naiditch, M. and Dourgnon, P. (2010). The preferred doctor scheme: a political reading of a French experiment of Gatekeeping. *Health Policy* **94**, 129–134.

Porteiro, N. (2005). Regulation of specialized medical care with public and private provision. *European Journal of Political Economy* **21**, 221–246.

Schwenkglenks, M., Preiswerk, G., Lehner, R., Weber, F. and Szucs, T. D. (2006). Economic efficiency of gatekeeping compared with fee for service plans: a Swiss example. *Journal of Epidemiology & Community Health* **60**, 24–30.

Velasco Garrido, M., Zentner, A. and Busse, R. (2011). The effects of gatekeeping; a systematic review of the literature. *Scandinavian Journal of Primary Health Care* **29**(1), 28–38.

WHO (2008). *The World Health Report 2008 – Primary Health Care (Now More Than Ever).*

Primer on the Use of Bayesian Methods in Health Economics

JL Tobias, Purdue University, West Lafayette, IN, USA

Introduction

Bayesian econometrics has become an increasingly popular paradigm for the fitting of economic models, since the early 1990s. Although Bayesian efforts in economics existed well before this time – perhaps originating in our specific discipline with the pioneering work of Zellner in the early 1970s – Bayesian applied work before 1990 was relatively scarce and often resorted to approximate or asymptotic posterior analysis in order to make the approach operational.

The rather dramatic upswing in popularity that occurred in the 1990s seemingly, and perhaps surprisingly to many previously embroiled in the Bayes/frequentist debate, had little to do with ideology or a conversion of the masses to the tenets of Bayesian theory but instead was derived from a simulation-based 'revolution' that greatly facilitated Bayesian computation. Although, in principle, posterior distributions could always be obtained on the combination of prior and likelihood, numerical characterization of the posterior and its specific properties offered a daunting – and often insurmountable – computational challenge.

The purpose of this article is to review, in very general terms, two popular simulation-based algorithms that have greatly simplified the practice of Bayesian econometrics: the Gibbs sampling and the Metropolis–Hastings (M–H) algorithms, and to illustrate how these can be used to fit various microeconometric models commonly employed in health economics. The article reviews the basic procedures for implementing these algorithms, discusses strategies for diagnosing their convergence (or nonconvergence) and, finally, illustrates their application in various examples involving wages and their relationship to the Body Mass Index (BMI).

The outline of the article is as follows. The following section reviews the general approach to Bayesian estimation and inference and then provides details of both the Gibbs sampling and M–H algorithms in a representative setting. The authors then apply the Gibbs sampling algorithm in a linear regression model and review issues of convergence diagnostics and posterior prediction within that context. The Section Bayesian Inference in Latent Variable Models extends these ideas to nonlinear settings, thereby providing a generic representation that encompasses a variety of discrete choice models widely used in health applications. This article concludes with a brief summary.

How It Works

To begin, by consider a model $\mathcal{M}$ that yields a parametric likelihood function $L(\theta; y)$. The reader with no previous exposure to Bayes will likely find this portion of the empirical exercise familiar: Distributional assumptions on unobserved components of the model induce a sampling distribution for the data y, denoted as $p(y|\theta)$. Common examples include the assumption of normally distributed errors in linear regression, yielding the classical normal linear regression model (LRM), or extreme value-distributed errors in the context of a binary choice problem, leading to the logit. Regarded as a function of θ given the observed data y, this defines the likelihood function.

A non-Bayesian or frequentist econometrician who is willing to go so far as to impose enough model structure to define a likelihood stops at this point, often proceeding with well-established tools for point estimation and known asymptotic approximations for inference. The Bayesian, however, continues beyond the specification of a likelihood and adds to it a prior, denoted as $p(\theta)$, describing their beliefs regarding values of the parameters before having witnessed the data. In some cases, the adopted prior may be 'informative,' constructed from results obtained from past studies or information offered to the researcher by an expert. What is commonly done in practice, however, is to employ a prior that is proper (i.e., it integrates to unity) yet suitably 'diffuse' or 'noninformative' in the sense that the data information will typically overwhelm whatever information is insinuated through the prior.

It is at this stage of adopting a prior that the frequentist often becomes uncomfortable, fearing that the analysis is no longer objective and that the Bayesian practitioner could and may have derived the results they want simply by choosing the prior accordingly. Although commenting on issues of prior sensitivity falls a little outside the very general goals of this article, such questions will inevitably be posed to any Bayesian practitioner. In light of this, it seems useful to at least make a few brief points on this front, both as a means to motivate the Bayesian approach and also to offer the novice Bayesian a few possible responses to these kinds of queries.

First, it is useful to point out that in smooth, finite-dimensional models – such as all those considered here – the priors employed typically have little effect on posterior results with even moderate data. Although prior sensitivity is certainly a concern, the issue is often overblown and raised by those with an inherent distrust of Bayesian methods and, often, little knowledge of their operation. The influence of the prior is, however, of first-order importance in Bayesian model selection and comparison – an issue which is not addressed in detail here but is discussed in the references provided at the conclusion of this article.

Second, to say that frequentist econometrics is prior free and clearly differentiated from the Bayesian approach by its lack of subjectivity is simply incorrect. The very process of model/variable selection is inevitably personal and subjective: Imagine, for example, two different researchers locked in separate rooms, each in possession of a data set such as the National Longitudinal Survey of Youth or National Health and Nutrition Examination Survey and seeking to use it to answer the same economic question. It does not seem controversial to conclude that these two researchers will almost surely arrive

 doi:10.1016/B978-0-12-375678-7.00708-2

at different final models with resulting summary point estimates. The fact that ordinary least squares (OLS) or some other 'objective' estimator was used in the estimation of parameters in the final models simply masked the appearance of prior information that was used at earlier parts of the modeling process. One researcher will have deemed one subset of covariates as relevant and worthy of potential inclusion and further consideration, whereas the other, via his or her own beliefs and decision making, will likely focus on a different set of explanatory variables. These types of judgments are simply prior beliefs at work, and these types of priors that sculpt the classes of models to be entertained are, in fact, more restrictive than priors over parameters that the Bayesian employs, because the data can at least revise the latter type of belief, whereas no amount of data can revise the former.

Finally, there are a variety of numerical strategies for assessing the sensitivity of posterior estimation results with respect to the prior, and care regarding the prior and its influence should be a component of any serious Bayesian endeavor. Given the goals of this article, these methods cannot be discussed in detail, but the reader is advised to refer again to the references at the conclusion of this article.

Bayesian Computation

Now turn to the practical issue of Bayesian computation. Given the likelihood $L(\theta; y)$ and prior $p(\theta)$, the joint posterior distribution, denoted as $p(\theta|y)$, is obtained via Bayes' Theorem as:

$$p(\theta|y) = \frac{L(\theta;y)p(\theta)}{\int_{\Theta} L(\theta;y)p(\theta)d\theta} \quad [1]$$

The posterior distribution in eqn [1] summarizes beliefs regarding θ after having combined the prior and likelihood and represents the output of any Bayesian analysis. When θ is univariate or bivariate, this output can be summarized by simply plotting the prior times likelihood in the numerator of eqn [1] over different values of θ, thus providing the analyst with a sense of the overall shape of the joint posterior. In most cases, however, $p(\theta|y)$ is high dimensional, rendering this type of visual analysis practically infeasible.

In models of even moderate complexity, moments or specific features of eqn [1] cannot be directly calculated, as the denominator – the normalizing constant of the posterior – cannot be determined analytically. Moreover, even if this normalizing constant were known, calculation of a posterior moment or posterior quantile of interest from eqn [1] will define an additional integration problem that would likely lack a closed-form solution.

Fortunately, recent simulation-based methods like the Gibbs sampler and M–H algorithms offer convenient and powerful numerical strategies for the calculation of statistics and features of eqn [1]. These algorithms deliver a series of draws, say, $\theta^{(0)},\theta^{(1)},\theta^{(2)},\ldots$ that are constructed to converge in distribution to the joint posterior described in eqn [1]. The draws are Markovian, with the value of the draw at iteration $t+1$ depending on the current parameter value $\theta^{(t)}$. Once convergence to the target density is 'achieved' these draws can be used in the same way as one would use direct Monte Carlo integration in order to calculate posterior means, posterior standard deviations, and other posterior statistics. For example, in order to estimate a posterior mean of θ, one can simply take a sample average of the simulated θ draws. In practice, care should be taken to diagnose that the parameter chain has converged to the target density, to discard an initial set of the preconvergence draws (often called a burn-in period), and then to use the postconvergence sample to calculate the desired quantities.

The postconvergence draws that is obtained using these iterative methods will prove to be correlated, as the distribution of, say, $\theta^{(t)}$ depends on the last parameter sampled in the chain, $\theta^{(t-1)}$.

If the correlation among the draws is severe, it may prove to be difficult to traverse the entire parameter space, and the numerical standard errors associated with the point estimates can be quite large.

With this general preview of the methods in place, and broad concerns regarding their performance in mind, the details of two commonly used simulation-based methods for Bayesian estimation and inference will now be discussed.

The Gibbs sampler

Let θ be a $K \times 1$ parameter vector with associated posterior distribution $p(\theta|y)$ and write $\theta = [\theta_1\ \theta_2 \ldots \theta_K]$. The Gibbs sampling algorithm proceeds as follows:

(i) Select an initial parameter vector $\theta^{(0)} = [\theta_1^{(0)}\ \theta_2^{(0)} \ldots \theta_K^{(0)}]$. This initial value could be arbitrarily chosen, sampled from the prior, or perhaps could be obtained from a crude estimation method such as least squares.

(1) Sample $\theta_1^{(1)}$ from the complete posterior conditional density:

$$p(\theta_1|\theta_2 = \theta_2^{(0)}, \theta_3 = \theta_3^{(0)}, \cdots \theta_K = \theta_K^{(0)}, y)$$

(2) Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1 = \theta_1^{(1)}, \theta_3 = \theta_3^{(0)}, \cdots, \theta_K = \theta_K^{(0)}, y)$.

$\vdots$

(K) Sample $\theta_K^{(1)}$ from $p(\theta_K|\theta_1 = \theta_1^{(1)}, \theta_2 = \theta_2^{(1)}, \cdots, \theta_{K-1} = \theta_{K-1}^{(1)}, y)$

(ii) Repeatedly cycle through (1)→(K) to obtain $\theta^{(2)} = [\theta_1^{(2)}\ \theta_2^{(2)}\ \theta_K^{(2)}]$, $\theta^{(3)}$, etc., always conditioning on the most recent values of the parameters drawn (e.g., to obtain $\theta_1^{(2)}$, draw from $p(\theta_1|\theta_2 = \theta_2^{(1)}, \theta_3 = \theta_3^{(1)}, \ldots \theta_K = \theta_K^{(1)}, y)$, etc.). Also note that some groups of parameters can be blocked together (such as a full vector of regression parameters) and the conditionals in (1)–(K) need not be univariate.

To implement the Gibbs sampler, the ability to draw from the posterior conditionals of the model is required. Although the joint posterior density $p(\theta|y)$ may often be intractable, the complete conditionals prove to be of standard forms in many cases, particularly in hierarchical models and latent variable models using data augmentation. For this reason, the Gibbs sampler is now routinely used to fit a variety of popular econometric models. An illustration of the power of the Gibbs algorithm will be provided in the following sections.

The Metropolis–Hastings algorithm

The M–H algorithm is an accept–reject type of algorithm in which a candidate value, say θ_c, is proposed, and then one decides whether to set $\theta^{(t+1)}$ (the next value of the chain) equal to θ_c or to remain at the current value of the chain, $\theta^{(t)}$. Formally, let $P(\theta|\theta^{(t)})$ be an approximating proposal density (where the potential dependence on the current value of the chain is made explicit), and consider generating samples from $P(\theta|\theta^{(t)})$ instead of the target distribution $p(\theta|y)$. Supposing that θ_c is sampled from $P(\cdot|\theta^{(t)})$, set $\theta^{(t+1)}=\theta_c$ with (M–H) probability

$$\min\left\{1, \frac{p(\theta_c|y)}{P(\theta_c|\theta^{(t)})}\frac{P(\theta^{(t)}|\theta_c)}{p(\theta^{(t)}|y)}\right\} \quad [2]$$

and otherwise set $\theta^{(t+1)}=\theta^{(t)}$. In the case of a symmetric proposal density (the original Metropolis algorithm), the above probability of acceptance reduces to $p(\theta_c|y)/p(\theta^{(t)}|y)$, wherefrom candidate draws from regions of higher density are always accepted in the algorithm, and draws from regions of lower density are occasionally accepted.

The Gibbs sampler and the M–H algorithms are often used in combination in a given application. For example, it might be the case that the complete conditionals for $K-1$ of the elements of θ have convenient functional forms, wherefrom the Gibbs sampler can be used to sample from these $K-1$ posterior conditionals. The complete conditional for the remaining parameter, however, may not take a standard form, and for this parameter, one could use the M–H algorithm to generate samples. This type of sampling is often (though mostly inappropriately) referred to as a 'Metropolis-within-Gibbs' step, and in (partially) nonconjugate situations, the use of both algorithms in combination often proves to be computationally attractive.

To illustrate the application of Markov Chain Monte Carlo (MCMC) methods in practice (and the Gibbs sampler in particular), the following section provides a simple example in the context of the LRM.

A Simple Linear Regression Example

To serve as a starting point, consider estimation and posterior prediction in an LRM. The goal in this section is to review a Gibbs sampling algorithm for the basic linear model, to briefly discuss methods for diagnosing convergence of a posterior simulator, and to illustrate how the posterior simulations can be used to calculate a variety of objects of interest. Under the assumption of conditionally normally distributed errors and homoskedasticity, the linear model can be written as:

$$y_i = x_i\beta + u_i, \quad u_i|X \overset{iid}{\sim} \mathcal{N}(0,\sigma^2) \quad [3]$$

where x_i is a $1\times k$ vector of covariate data and $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

The likelihood function is implied by eqn [3], and a Bayesian analysis is completed by specifying prior distributions for the model parameters which, for this model, are β and σ^2. Here, priors that are conditionally conjugate are specified, meaning that they combine naturally with the likelihood and will yield conditional posterior distributions that are of the same distributional families as the priors. Specifically,

$$\beta \sim \mathcal{N}(\mu_\beta, V_\beta) \quad [4]$$

$$\sigma^2 \sim IG(a,b) \quad [5]$$

are chosen, with $IG(a, b)$ denoting an inverse gamma density with parameters a and b. The hyperparameters μ_β,V_β,a and b are chosen by the researcher to accord with their prior beliefs and, often, are specified so that the prior densities will be quite flat over a large region of the parameter space.

To effect a Gibbs sampling scheme for fitting this model, the conditional posterior distributions $\beta|\sigma^2, y$ and $\sigma^2|\beta, y$ are required. With a bit of algebra, one can show that these conditional distributions are normal and inverse gamma, respectively. The conditional posterior mean of β can be expressed as a matrix-weighted average of the prior mean μ_β and the OLS estimate $\hat{\beta}=(X'X)^{-1}X'y$. As the size of the data set (n) grows, this posterior mean places increasing weight on the OLS estimate and, given a fixed prior, equals the OLS estimator in the limit. Similarly, the conditional posterior mean of σ^2, $E(\sigma^2|\beta, y)$, can be written as a simple weighted average of the prior mean and the maximum likelihood estimate (given β), $\hat{\sigma}^2=(y-X\beta)'(y-X\beta)/n$. For fixed a and b it is again the case that as $n\to\infty$, the conditional posterior mean collapses on $\hat{\sigma}^2$.

A Gibbs algorithm for fitting this model, as previously discussed in the Section How It Works, involves iteratively sampling from the conditional multivariate normal density $\beta|\sigma^2, y$ and the inverse gamma density $\sigma^2|\beta, y$. Although routines for sampling from a multivariate normal are well known, sampling from the inverse gamma is probably less familiar, although equally easy to do, as such a simulation can be obtained by inverting a draw from a gamma distribution.

Obesity Example

To fix ideas, a specific LRM using a sample of female data ($n=1782$) from the study of Kline and Tobias (2008) is considered. These authors used data from the British Cohort study and sought to estimate the impact of BMI on labor market earnings. The regression of interest used in this study, therefore, uses log wages as the dependent variable and also includes an obesity indicator (BMI ≥ 30), tenure on the current job (and its square), family income, a high school completion indicator, an indicator for an A-level degree, an indicator for a college degree, and finally, union and marriage indicators as controls.

To fit this model, the Gibbs sampler is run for 5500 iterations and the first 500 of these are discarded as a burn-in period. Recall from the previous Section The Gibbs sampler that Gibbs is an iterative algorithm – eventually the samples that are produced will represent a correlated set of draws from the posterior, although the initial set of simulations may not have converged to the posterior. To mitigate such effects, the first 500 draws are thrown out and the final 5000 simulations are used to calculate posterior means and standard deviations of parameters of interest. For the priors, fairly noninformative choices are made by setting $\mu_\beta=0$, $V_\beta=100I_k$, $a=3$ and $b=2.5$.

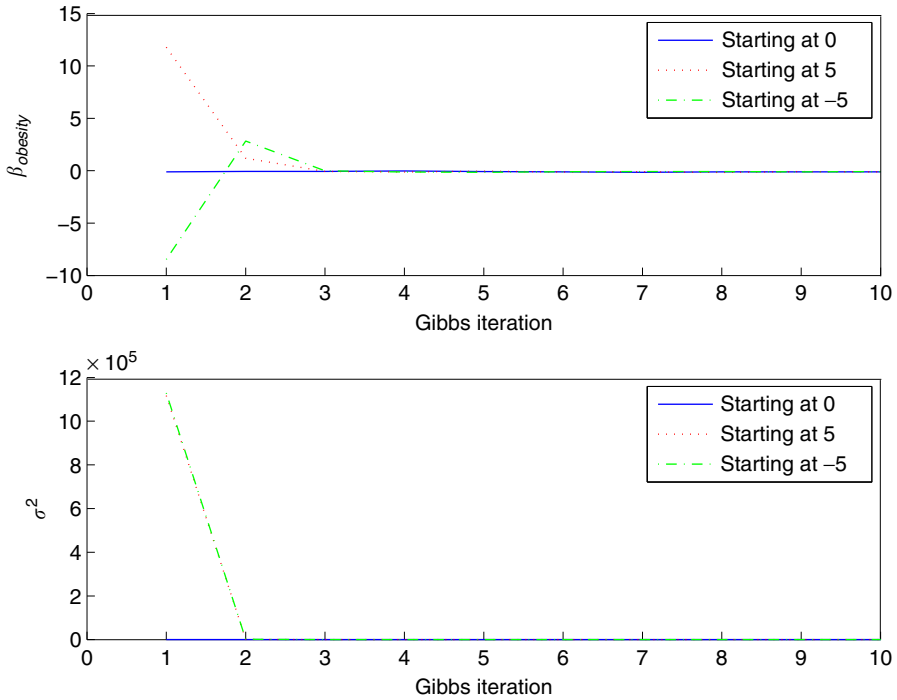


Figure 1 Initial Gibbs simulations from LRM example.

Convergence diagnostics and mixing

Before discussing the estimation results, the issue of convergence of the MCMC sampler used in the study is first considered and it is sought to determine if 500 draws represent an adequate burn-in period for the application. One popular method for assessing convergence is to repeat the Gibbs analysis several times, each time using a different set of starting values, where the starting values are typically chosen to be intentionally 'extreme' so that the progression of the parameter draws can be clearly monitored as they move toward exploring the common posterior surface.

Figure 1 presents graphical results of such an exercise, which is denoted as a trace plot in the literature. Here, three separate Gibbs chains are run: one starting with $\beta=0$ (for all elements of the coefficient vector), a second starting with $\beta=5$, and a third starting with $\beta=-5$. The top portion of the figure plots the path of a representative coefficient – the coefficient on obesity, denoted $\beta_{obesity}$, from all three Gibbs samplers (not plotting the initial condition itself). The lower panel similarly plots the paths of σ^2 for all three chains. If the paths of these parameters from the three separate chains reveal no intersection, this provides evidence that the samplers have not yet converged, that a longer burn-in period is required, and potentially, that the algorithm itself may need to be refined in order to accelerate convergence. However, if it is observed that the simulations quickly move away from their overdispersed (and incorrect) starting values to eventually explore a common region of the parameter space, this provides evidence of convergence within the viewed number of iterations.

As Figure 1 clearly suggests, within just three iterations, the three separate chains appear to settle down and explore the same region of the parameter space. When starting with the unreasonable values of $\beta=5$ or $\beta=-5$, the first few iterations of the sampler produce very large values of the variance parameter σ^2, which is not at all surprising given the extent to which the very early values of β are far from the mass of the posterior density. However, within just three iterations, it appears as if the variance parameter simulations shake off the influence of these starting values and then settle down to explore a common area.

The evidence offered by these graphs is that convergence to the posterior happens very quickly in the example: by discarding the first 500 iterations, one can credibly guard oneself against the problem that the early set of simulations produced by the chosen sampler are not draws from the joint posterior distribution and should not be used in the calculations.

Figure 2 offers a second trace plot, but this time graphs of the first 50 postconvergence simulations obtained from the sampler (i.e., the draws obtained from iterations 501–550 for each of the three chains). The obesity coefficient is provided in the top panel and the variance parameter in the bottom panel. First note the tremendous refinement in scale in Figure 2 relative to Figure 1: The data convey substantial information regarding these parameters, and once convergence has been achieved, the simulations explore the posterior surface and are no longer influenced by initial conditions. Second, the three separate chains appear to have very similar properties, such as means and variances, again providing evidence of convergence. Finally, the graphs also speak of the mixing of the chains. If the draws in a Gibbs algorithm are highly correlated, then the trace plot will reveal very long cycles. The plot in Figure 2, however, does not appear to exhibit any type of strong cyclical behavior and, instead, appears more like that of an electroencephalography, as can be seen under iid sampling from the posterior distribution.

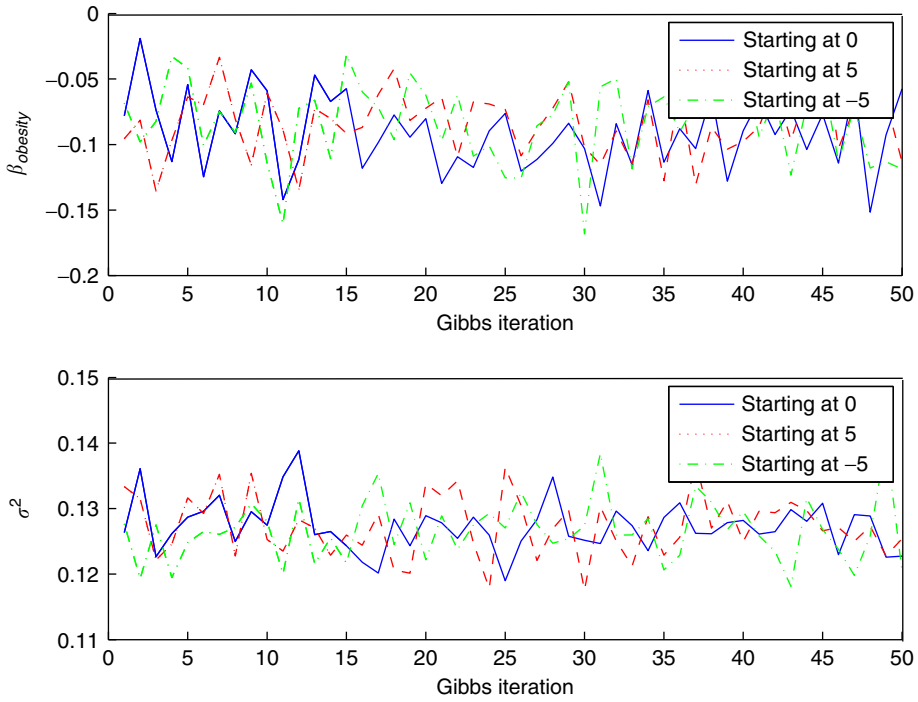


Figure 2 Gibbs iterations 501–550 from LRM example.

A more formal procedure one could use here in order to assess the mixing of the simulations is to report the so-called inefficiency factors for the parameters. These quantify the loss from adopting a Gibbs or M–H sampling approach (whose draws are autocorrelated) relative to an iid sampling scheme. These inefficiency factors can be obtained by calculating:

$$ineff_k = 1 + 2\sum_{j=1}^{J} \rho_k(j) \qquad [6]$$

where $\rho_k(j)$ refers to the lag-j autocorrelation for parameter k. These can be calculated from the simulated output by simply computing the correlation between simulations j iterations apart. The upper limit of the summation, J, is typically chosen in accord with some type of rule of thumb – for example, once the lag-J correlation is smaller than, say, .05, the contribution of lag-t correlations, for all $t > J$.

When $ineff_k$ is near 1, the mixing of the posterior simulations is near the iid ideal. The value of $ineff_k$ is also directly interpretable: If a value equal to $c > 1$ is obtained, this implies that one must run the sampler for $c \times M$ iterations in order to reproduce the numerical efficiency found in M *iid* simulations. For this linear regression example, it is found that the inefficiency factors for all parameters are very close to 1, indicating that the Gibbs algorithm effectively mimics the performance of iid sampling from the joint posterior.

Estimation results

Presented in Table 1 are posterior means, posterior standard deviations, and posterior probabilities of being positive for the parameters of the LRM used in the study. For reference, the first column also reports OLS coefficient estimates, which are seen to be virtually indistinguishable from the reported posterior means, which is to be expected with ($n = 1782$) and fairly noninformative priors. With respect to the key parameter of interest, it is found that obesity clearly has a negative impact on log wages, and this finding can be summarized via the statement: $[\Pr(\beta_{obese} < 0|y) \approx .999]$. In terms of the point estimate, obese females earn approximately 8.2 % less, on average, than women who are not obese. With respect to the quantity $\Pr(\beta_{obese} < 0|y) \approx .999$, note that it offers a very natural interpretation of the evidence at hand: conditioned on the model, the priors, and the observed data, it can be confidently (nearly certain) claimed that obesity has a negative effect on wages.

Table 1 Posterior statistics from LRM

Variable	*OLS*	$E(\cdot\|y)$	$Std(\cdot\|y)$	$\Pr(>0\|y)$
Intercept	1.68	1.68	0.030	1.00
Obese	−0.082	−0.082	0.029	0.001
Tenure	0.025	0.025	0.007	0.999
Tenure^2	−0.001	−0.001	0.0005	0.015
FamInc	0.001	0.001	0.0001	1.00
High school	0.068	0.069	0.0224	0.998
A-level	0.282	0.282	0.0332	1.00
Degree	0.347	0.347	0.0254	1.00
Union	0.251	0.247	0.0195	0.897
Married	−0.019	−0.019	0.0176	0.142

Although the Bayesian posterior probabilities reported in the final column of Table 1 may seem somewhat similar to the frequentist p-value, it is important to recognize that they are vastly different in terms of interpretation. First, such statements are entirely inappropriate in the classical paradigm, as β_{obese} is a fixed parameter and is, therefore, either negative,

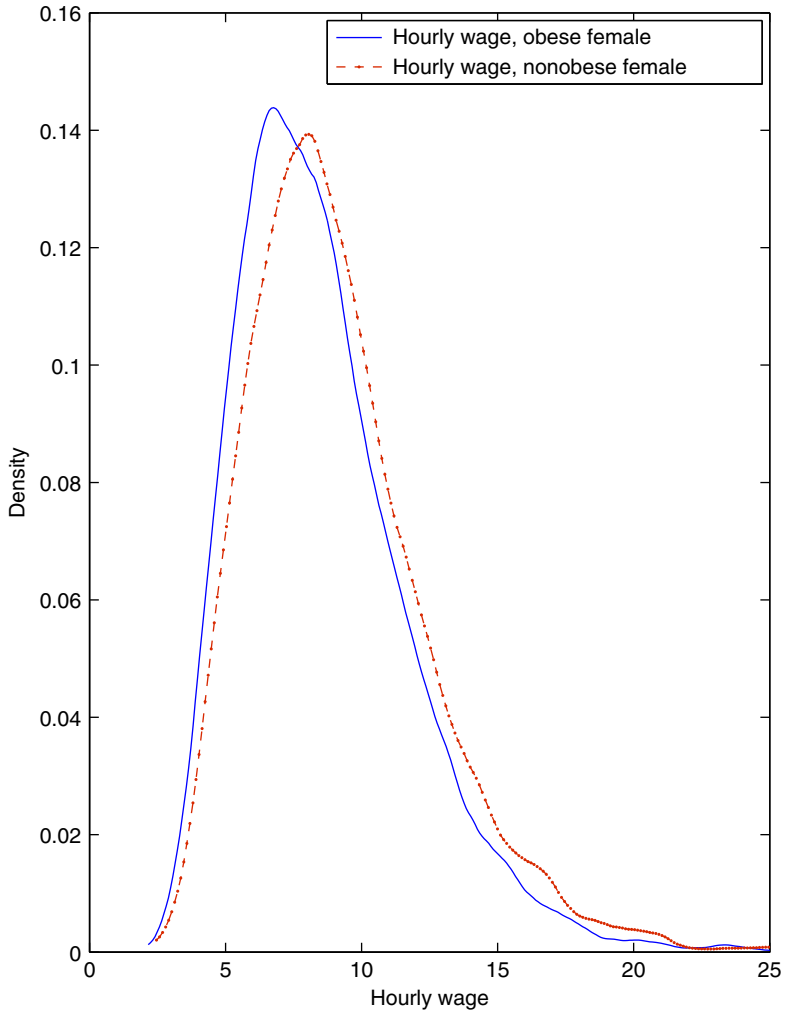


Figure 3 Posterior predictive wage densities for obese and nonobese females.

positive, or zero. The probability statement that the Bayesian is able to provide, however, seems to be exactly what was expected to obtain from the analysis and offers a very useful quantity to offer to medical professionals and/or policy makers. Second, any probabilistic interpretation of effects in the frequentist case is derived from the sampling distribution of the estimator, highlighting the fact that such an approach places primary importance on what estimates one may have obtained had other data sets actually been observed. For the Bayesian, this sense of importance is misplaced, as decisions and recommendations should be based on the data at hand, and averaging over data sets that could have been observed, but were not, is neither advisable nor relevant.

Posterior prediction

The last exercise focuses on posterior prediction. Here, it is sought to move beyond the narrow (and too often terminal) goal of parameter estimation to use the model to make predictions about future outcomes. In the context of the LRM used in this study, let w_f denote the hourly wage (not log wage) of the individual, and let x_f denote a given set of fixed characteristics. Because the predictive outcome $w_f = \exp(y_f)$ is a function of the regression parameters β and σ^2, the posterior predictive distribution of w_f via simulation is numerically approximated. That is, for every $\beta(r)$ and $\sigma^{2,(r)}$ simulation from the posterior, it can be simply calculated as:

$$w_f^{(r)} = \exp[x_f \beta^{(r)} + \sigma^{(r)} \varepsilon^{(r)}] \quad [7]$$

where $\varepsilon^{(r)} \sim^{iid} \mathcal{N}(0,1)$. Proceeding in this way generates a series of draws from the posterior predictive wage distribution – one draw obtained for each (β, σ^2) simulation from the joint posterior.

In Figure 3 this approach is used and posterior predictive wage distributions are reported for obese and nonobese females, keeping the other explanatory variables constant and approximately equal to overall sample mean values. These densities are obtained by kernel-smoothing the simulated draws from the posterior predictive density.

As shown in the figure, the wage density for the nonobese female is right-shifted relative to that of the obese female. To contrast this with a frequentist alternative, note that such an approach would likely proceed by plugging in point estimates of β and σ^2 into an expression for the hourly wage density. Although such an expression can be analytically derived in this

particular example (i.e., the wage density is lognormal), this exercise can often require a messy change of variables. Furthermore, such an approach suppresses parameter uncertainty and instead conditions on values of the parameter point estimates. The Bayesian approach to this exercise is quite appealing by comparison, as simulations can be used in place of analytic derivations, and parameter uncertainty is automatically accounted for, as they are properly integrated out to obtain the posterior predictive distribution $p(w_f|x_f,\gamma)$.

Bayesian Inference in Latent Variable Models

Although useful as a starting point and surely of interest in its own right, the previous treatment of the LRM is far from fully satisfactory, as it does not cover Bayesian model fitting in wide array of alternate models – including nonlinear and limited dependent variable models, for example – that are widely used in health research. In this section, a very general structure that nests many popular nonlinear microeconometric models is, therefore, introduced and posterior simulation for this general case is discussed. To this end, consider the following specification:

$$p(\theta) = p(\beta)p(\Sigma^{-1}) \quad [8]$$

$$z_i|X,\theta \overset{ind}{\sim} \mathcal{N}(X_i\beta,\Sigma), \quad i = 1, 2, \ldots, n \quad [9]$$

$$y_i|z_i = g(z_i), \quad i = 1, 2, \ldots, n \quad [10]$$

Equation [8] introduces a prior for the model parameters, consisting of a set of regression coefficients β and an inverse covariance matrix Σ^{-1}. Equation [9] depicts the generation of a (potentially) multivariate latent variable z_i, whereas eqn [10] links the observed outcomes y_i to the latent data z_i.

The generality of eqns [8]–[10] should not be overlooked. To illustrate, note for a univariate outcome y_i:

- The probit model is a special case of eqns [8]–[10], where z_i is a scalar, Σ^{-1} is absent from the model (i.e., the scalar variance parameter is normalized to unity), and eqn [10] specializes to $y_i = I(z_i > 0)$, where $I(\cdot)$ denotes the standard indicator function.
- The tobit model is produced when z_i is a scalar, $\Sigma = \sigma^2$ is a scalar variance parameter, and eqn [10] specializes to $y_i = max\{0, z_i\}$.
- The ordered probit model is produced when z_i is a scalar, Σ^{-1} is absent from the model, and the observed data is connected to the latent data via a cutpoint vector: α:

$$y_i = j \quad if \quad \alpha_j < z_i \leq \alpha_{j+1}$$

In this final instance, eqn [10] must be generalized to allow the link between latent and observed outcomes to depend on the cutpoint vector α.

The equations outlined in eqns [8]–[10] also cover a wide variety of multivariate models. For example, the seemingly unrelated regressions model, the hurdle and sample selection models, generalized tobit models, multivariate ordinal models, models with continuous and discrete endogenous variables (as discussed in more detail below), and the multivariate probit (MVP) and multinomial probit (MNP) models can be regarded as specific cases of this general structure. As such, knowing how to estimate a system of equations like eqns [8]–[10] enables the applied researcher to estimate a wide variety of popular microeconometric models.

Applying the idea of data augmentation, the latent data z in eqn [9] can be treated like another parameter of the model, giving rise to the augmented joint posterior distribution $p(z,\beta,\Sigma^{-1}|y)$. A Gibbs sampler, then, draws from each of the three constituent posterior conditional distributions.

In some models, the required simulations are almost trivially performed, whereas other specifications will demand refinement of any general scheme. In many cases, β will be sampled from a multivariate normal, Σ^{-1} from a Wishart distribution, and z_i from a truncated normal. Posterior simulation in the probit model, for example, simply involves generation of a multivariate normal random variable for sampling β and a series of (independent) univariate truncated normals for sampling the latent data z. In terms of the latter, the observed binary response y_i serves to truncate the value of the latent variable: $y_i = 1$ indicates that z_i must be positive, whereas $y_i = 0$ restricts z_i to be nonpositive. In the MNP and MVP models, however, additional care must be taken to normalize the associated variance parameter(s) to unity. The general sampling of the latent data will also prove more challenging in some models than in others, and an appropriate method of sampling the latent data must be carefully considered for the model in question.

With these practical implementation details noted, the reader should still recognize the generality and scope of coverage that the simple description of eqns [8]–[10] offers, as well as the fact that in many cases, nonlinear model fitting only requires the ability to sample from standard distributions (such as the normal, truncated normal, and Wishart). Further details and associated references regarding a variety of specific models are provided in the further reading section of this article.

Posterior Simulation with an Endogenous Binary Variable

To provide a specific example of a popular model 'covered' by eqns [8]–[10], the well-studied case of a continuous outcome model with a dummy endogenous variable is considered. In this case, eqn [9] might specialize to:

$$z_{i0} = r_i\gamma + u_{i0} \quad [11]$$

$$y_{i1} = \alpha_0 + \alpha_1 y_{i0} + s_i\alpha_2 + u_{i1} \quad [12]$$

where

$$y_{i0} = I(z_{i0} > 0) \quad [13]$$

Equation [11] is a latent variable equation governing the generation of an endogenous binary outcome y_{i0}. The variable y_1 is a continuous outcome variable, which is completely observed. The variables r and s are assumed to be exogenous, with r containing at least one element that is not in s.

Table 2 Posterior statistics from binary endogenous variable model

Coefficient	*Log wage equation*			*Obesity probit equation*		
	$E(\cdot\|y)$	$Std(\cdot\|y)$	$\Pr(>0\|y)$	$E(\cdot\|y)$	$Std(\cdot\|y)$	$\Pr(>0\|y)$
Constant	1.70	0.033	1.00	−4.91	0.451	0.000
Obese	−0.278	0.086	0.001			
Tenure	0.025	0.007	0.999			
$Tenure^2$	−0.001	0.001	0.015			
FamInc	0.001	0.001	1.00	−0.001	0.001	0.375
HighSchool	0.068	0.023	0.998	−0.001	0.108	0.500
Alevel	0.286	0.034	1.00	0.150	0.162	0.821
Degree	0.334	0.027	1.00	−0.304	0.140	0.015
Union	0.025	0.019	0.912			
Married	−0.014	0.018	0.198	0.065	0.087	0.775
momBMI				0.087	0.011	1.00
dadBMI				0.062	0.014	1.00
ρ_{01}	0.305	0.122	0.986			
σ_1^2	0.131	0.005	1.00			

With a bit of finesse, this bivariate system of equations can also be mapped into the form of the system in eqn [9]. In terms of posterior simulation, the variables can be stacked and the vector of parameters $\beta=[\gamma\ \alpha_0\ \alpha_1\ \alpha_2]$ can be sampled from a multivariate normal. A straightforward reparameterization of the problem enables the simulation of the variance parameter σ_1^2 and the covariance between the errors of eqns [11] and [12], denoted σ_{01}. Finally, the latent data z_{i0} can be sampled from a univariate truncated normal, where the region of truncation is governed by the observed value y_{i0}.

Obesity Example, Revisited

For illustration purposes and to expand the application of Section 3 (where the potential endogeneity of female obesity was ignored), the binary endogenous variable model of Section 1 is fitted and applied to wage/obesity data. The continuous outcome y_{i1} remains the log hourly wage and the binary endogenous variable y_{i0} is the obesity indicator. The covariates in the wage equation remain the same as those employed in Section 2, whereas in eqn [11] maternal and paternal BMI are included as exclusion restrictions (i.e., variables that are assumed to affect the respondent's BMI but have no conditional effect on the respondent's log wage). In addition, this equation also includes a set of education indicators, family income and a marriage indicator. The Gibbs algorithm is run for 10 000 iterations, with the first 1000 discarded as the burn-in. Results of this analysis are provided in Table 2.

Apart from the values of the obesity and union coefficients, the values of the log wage regression coefficients remain very similar to those presented in Table 1. In the obesity equation, consistent with the prior expectations, strong evidence that higher parental BMI leads to a higher likelihood of child obesity and that the college educated have lower rates of obesityis found.

The interesting difference in results relative to those in Table 1 lies in the estimated impact of obesity on log wages and its interpretation. As shown in the second-to-last row of Table 2, ρ_{01} is found to be positive with very high posterior probability. One interpretation of this result, which is also offered by Kline and Tobias, 2008, is that individuals who are very dedicated to their job, perhaps working long hours in order to earn a high conditional wage, may forego investments in health capital in order to earn such a wage. As a result, unobserved factors leading an individual to earn a high conditional wage will also positively correlate with unobservables affecting the production of obesity, consistent with the positive value of ρ_{01} in Table 2. The results in Table 1, then, have presented a conservative estimate of the obesity penalty, as they have combined the actual obesity penalty with an effect arising from unobservable characteristics of productive workers that also contribute to obesity. When separating these effects, as the analysis of Table 2 seeks to do, a much larger estimate of the obesity penalty is seen, as the coefficient has increased (in an absolute sense) more than three-fold, with a posterior mean now equal to −.278.

Although the specific results of this application are of secondary importance, what is most important to appreciate is the relative ease with which these results have been obtained; fitting the model only required the ability to draw from normal, inverse gamma and univariate truncated normal distributions. This simplicity is not specific to the model considered here but applies to many specifications that fall within the class of models described by eqns [8]–[10]. Given the relative ease with which parameter and latent variable simulations can be obtained from the joint posterior, it then becomes easy to move beyond point estimation to calculate marginal effects and to conduct various counterfactual or policy experiments.

Conclusion

This article has reviewed the basics of the Bayesian approach to estimation and inference, focusing in particular on applications of the Gibbs sampler and M–H algorithms. In conjunction with data augmentation, these algorithms greatly ease estimation in a variety of microeconometric models, many of which are commonly employed in health economics applications. Although the material presented here only offers a superficial review of the Bayesian approach, the references

provided contain significantly more information on both the theory and application of these methods.

Reference

Kline, B. and Tobias, J. L. (2008). The wages of BMI: Bayesian analysis of a skewed treatment-response model with nonparametric endogeneity. *Journal of Applied Econometrics* **23**, 767–793.

Further Reading

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Chib, S. (2012). MCMC methods. In Geweke, J., Koop, G. and van Dijk, H. (eds.) *Handbook of Bayesian econometrics*, pp 183–220. Oxford: Oxford University Press.

Gelfand, A., Hills, S., Racine-Poon, A. and Smith, A. (1990). Illustration of Bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*. Hoboken, NJ: John Wiley & Sons.

Geweke, J. and Keane, M. (2001). Computationally intensive methods for integration in econometrics. In Heckman, J. J. and Leamer, E. (eds.) *Handbook of Econometrics*, vol. 5, pp 3463–3568. Amsterdam, The Netherlands: Elsevier.

Greenberg, E. (2008). *Introduction to Bayesian econometrics*. New York: Cambridge University Press.

Koop, G. (2003). *Bayesian econometrics*. Hoboken, NJ: John Wiley & Sons.

Koop, G., Poirier, D. J. and Tobias, J. L. (2007). *Bayesian econometric methods*. New York: Cambridge University Press.

Lancaster, A. (2004). *An introduction to modern Bayesian econometrics*. Malden, MA: Blackwell Publishing.

Lopes, H. and Tobias, J. L. (2011). Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis. *Annual Review of Economics* **3**, 107–131.

Poirier, D. J. (1995). *Intermediate statistics and econometrics: A comparative approach*. Cambridge, MA: MIT Press.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian statistics and marketing*. Hoboken, NJ: John Wiley & Sons.

Train, K. E. (2009). *Discrete choice methods with simulation*. New York: Cambridge University Press.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. Malabar, FL: Krieger Publishing Company.

Priority Setting in Public Health

K Lawson, University of Glasgow, Glasgow, Scotland
H Mason, Glasgow Caledonian University, Glasgow, Scotland
E McIntosh, University of Glasgow, Glasgow, Scotland
C Donaldson, Glasgow Caledonian University, Glasgow, Scotland

Glossary

Cost-benefit analysis A form of economic evaluation by comparing the costs and (money-valued) benefits of alternative courses of action.

Cost-consequence analysis A method of assembling the components of the costs and benefits of a project or investment option, usually in noncomparable units, without any attempt to combine them into a single monetary cost figure or combined artificial construct.

Cost-effectiveness analysis A method of comparing the opportunity costs of various alternative health or social care interventions having the same benefit or in terms of a common unit of output, outcome, or other measure of accomplishment.

Equity Equity is not necessarily to be identified with equality or egalitarianism, but relates in general to ethical judgments about the fairness of the distribution of such things as income and wealth, cost and benefit, access to health services, exposure to health-threatening hazards, and so on. Although not the same as 'equality,' for some people, equity frequently involves the equality of something (such as opportunity, health, and access). There are, however, also fair inequalities. For example, an unequal allocation of health care may be necessary in order to achieve more equal health.

Marginal The additional benefit, health, cost, etc. attributable to a small increase in a factor bringing it about (other things equal).

Multi-criteria decision analysis A technique, akin to cost-effectiveness analysis (CEA), for helping decision makers to take decisions. It differs from CEA by explicitly helping decision makers to consider factors beyond standard welfare or health maximization.

Opportunity cost The value of a resource in its most highly valued alternative use. In a world of competitive markets, in which all goods are traded and where there are no market imperfections, opportunity cost is revealed by the prices of resources: the alternative uses forgone cannot be valued higher than these prices or the resources would have gone to such uses.

Programme budgeting marginal analysis (PBMA) PBMA combines program budgeting with marginal analysis to provide a means of both determining which resources have been allocated to which program goals and analyzing the opportunity cost of marginal changes in the sizes of programs and the mix of inputs in comparison with the consequential changes in goal outcomes.

Quality-adjusted life expectancy Life-expectancy using quality-adjusted life-years rather than years of life.

Introduction

There is an established 'healthy public policy' agenda concerned with the social determinants of health, which recognizes that nonhealth sectors of public policy often have greater impacts on population health and health inequalities than health sector policies. This political agenda has been promoted by the World Health Organization (WHO) since the 1980s (see Box 1), and has led to a widening of the definition of public health interventions to include nonhealth sectors.

In line with this agenda, there are increasing calls for economists to help generate evidence that investing in the social determinants of health, for the explicit purpose of improving health and tackling health inequality, can represent good value for money. To date, however, health economic research in this area is relatively limited. This article reviews existing economic principles and methods of priority setting, and discusses how they can most fruitfully be applied to support the 'healthy public policy' agenda. The discussion focuses on supporting this agenda at the local government level (e.g., city or state level), which has important impacts on the social determinants of health, yet has hitherto received particularly limited attention by health economists. However, the economic principles and methods reviewed can of course also be applied to public policy making at national and supranational levels.

The section The Scope of the Challenge: The Social Determinants of Health illustrates the challenges faced inpublic health, where interventions that have major impacts on health originate from multiple policy sectors and are led by decision makers are primarily motivated to deliver specific nonhealth

Box 1 The WHO 'Adelaide Recommendations' on healthy public policy

"Healthy Public policy is characterized by an explicit concern for health and equity in all areas of policy and by an accountability for health impact. The main aim of healthy public policy is to create a supportive environment to enable people to lead healthy lives. Such a policy makes health choices possible or easier for citizens. It makes social and physical environments health-enchancing. In the pursuit of healthy public policy, government sectors concerned with agriculture, trade, education, industry, and communications need to take into account health as an essential factor when formulating policy. These sectors should be accountable for the health consequences of their policy decisions. They should pay as much attention to health as to economic considerations"

Second International Conference on Health Promotion, Adelaide, South Australia, 5–9 Apr 1988

outputs such as new housing. Intersectoral impacts are a common feature of most public policies; health is not a special case in this respect. Therefore, the wider challenge is how to encourage different policy sectors to consider and value all major intersectoral impacts of importance to society as a whole, including health.

The section How to Encourage Intersectoral Alignment: the Role of Economic Evaluation describes how economics is well-placed to encourage intersectoral coordination by developing the economic evidence in order to identify, measure and value the intersectoral spillovers of policies and the overall impact on social welfare. It begins by offering a brief primer on the economic way of thinking about priority setting. It emphasizes that a distinctive advantage of the economic way of thinking is its ability to adopt a broad societal perspective, which considers how best to allocate scarce resources to improve overall social welfare, making appropriate tradeoffs between different and potentially conflicting social objectives. Economics is thus not constrained to a narrow focus on one particular outcome or objective, such as improving health or reducing health inequality, but is capable of combining multiple objectives into the same analysis. Therefore, the scope of 'healthy public policy' fits naturally within an economic approach. The practical challenge is to engage with decision makers both to address sector specific concerns and to identify areas where further coordination can improve social welfare.

The article then considers appropriate approaches to economic evaluation and contends that the best approach is to combine cost consequence (CCA), cost effectiveness (CEA), and cost benefit analysis (CBA). This would account for and report all major intersectoral impacts, including health, and value these impacts in terms of net social welfare. It then reviews standard economic tools for priority setting and discuss their potential role in helping to coordinate different policy sectors, frame the decision process, make explicit stakeholder objectives, and translate economic evidence into policy. Taken together, the sections on economic evaluation and priority setting tools lay out an 'integrated societal framework' to enable all impacts to be accounted for, valued, and taken into consideration by decision makers.

The section Translating Evidence into Policy: The Role of Priority Setting Tools briefs how economic evidence can then be used by decision makers. It reviews standard economic tools for priority setting and discusses their potential role in helping to coordinate different policy sectors, frame the decision process, make explicit stakeholder objectives, and translate economic evidence into policy. Taken together, the sections on economic evaluation and priority setting tools lay out an 'integrated societal framework' to enable all impacts to be accounted for, valued, and taken into consideration by decision makers.

The Scope of the Challenge: The Social Determinants of Health

The Rise of the 'Healthy Public Policy' and 'Social Determinants of Health' Agendas

Alongside the political agenda for 'healthy public policy', there is also an established research agenda regarding the social determinants of health, with contributions from a number of eminent scholars from different disciplines. Much of this research from outside the discipline of economics has been usefully collated in the 2008 report of the WHO Commission on the Social Determinants of Health; though this report does not offer comprehensive coverage of economic contributions to theory and evidence on this topic.

Public health interventions can impact on health directly through the provision of public goods, such as water and sanitation; or through changes in legislation, such as environmental standards or food industry regulations. Other sectors also impact on health indirectly by influencing the willingness and ability of communities and individuals to invest in health, for example, through behaviors like healthy eating. For instance, education in childhood influences aspirations and future adult employment, which may in turn provide the means for individuals to invest in themselves and their children. Further, community interventions such as housing and regeneration may offer an incentive to invest in and protect neighborhoods from crime and antisocial behavior, which may be especially harmful for childhood development. Figure 1 attempts to summarize the wide array of drivers of population health. The diagram is limited in that the interaction between the various drivers is not captured and nor is the lifecourse, where an individual's early years development can influence their future adult outcomes. Nonetheless, it is a commonly used and helpful illustration that health is not simply the product of healthcare.

As a consequence, a 'healthy public policy' agenda has emerged, which may be crudely described as an advocacy movement. The contention is that interventions from nonhealth sectors should be thought of as 'upstream' public health interventions, which can have larger impacts on population health and health inequalities in the long run than health sector interventions.

A particular concern of the 'healthy public policy' agenda is that health impacts are still not being (fully) considered when nonhealth sectors make decisions. Further, in times of fiscal tightening, when public sector budgets are being reined in, policymakers understandably concentrate on short-term priorities such as the provision of amenities that are currently and visibly in high demand from voters and interest groups, which may then be detrimental to long run goals of improving population health and reducing health inequalities.

In recent years, there have been growing calls for economists to engage more fully with this agenda to generate economic evidence on which investments in the social determinants of health represent the best value for money.

Engaging with the 'Healthy Public Policy' Agenda: The Need for Intersectoral Alignment

In considering how best economists can engage with the 'healthy public policy' agenda, there are three important observations. The first is that decision makers are primarily incentivized to deliver sector specific outputs. Health outcomes, if considered at all, are byproducts and not the main priority of nonhealth sectors. Second, intersectoral impacts are a common feature of most policies. Third, health is

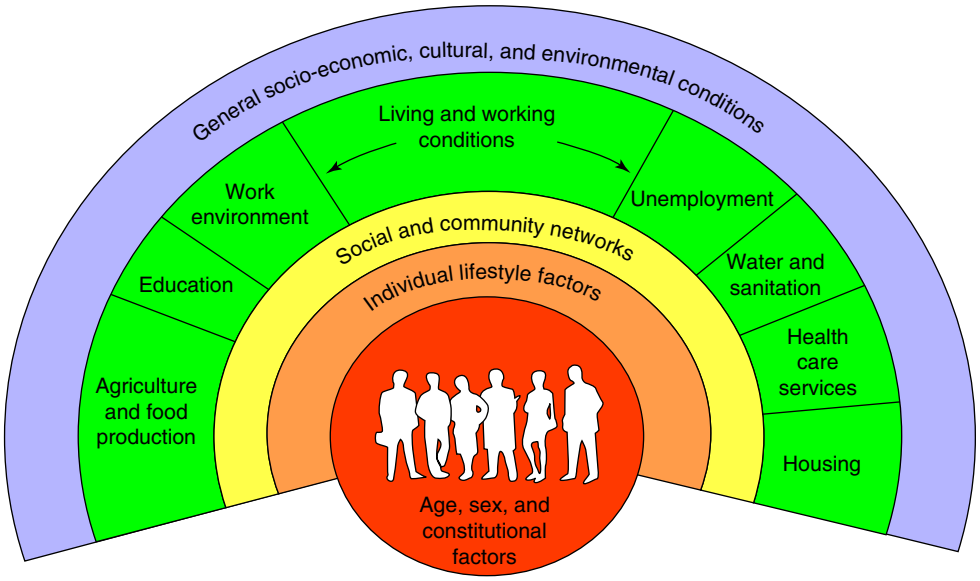


Figure 1 The drivers of population health. Reproduced from Dahlgren, G. and Whitehead, M. (1991). *Policies and strategies to promote social equity in health*. London: Institute of Fiscal Studies. Available at: http://ideas.repec.org/p/hhs/ifswps/2007_014.html (accessed 01.11.12).

nevertheless an outcome of particular importance for social welfare, which people value both as a consumption good and as an investment good that allows them to lead flourishing lives and act as productive members of their family and of the wider economy and society.

Taken together, these observations suggest that simply advocating to nonhealth sectors the importance of considering the health consequences of policies may not be enough to influence decision makers. There is a need to generate a '*quid pro quo*' to incentivize decision makers to consider health impacts. A productive way forward may be for all sectors to be encouraged to move beyond narrow sector perspectives to consider all major intersectoral impacts. Consequently, the challenge is not simply 'how can we persuade nonhealth sectors to produce health outcomes?' but 'how can we help coordinate and align sectors where health is valued as one important input toward achieving the common overall aim of increasing social welfare?'

Opportunity for Intersectoral Alignment: Local Decision-Making

Before considering how economics can help the ability of decision makers to coordinate, it is important to identify whether decision makers have the willingness to do so. This short section discusses where economists may have the most immediate opportunities to improve policy coordination.

It can be helpful to distinguish multiple policy 'levels' that drive health, including: international, national, regional, and local. However, these levels are not necessary strictly demarcated, and there are interactions between them. At the international level, major drivers include global warming and trade legislation. At the national level, the fiscal allocation of resources to spending departments is an important issue, and there is a dialog at the national level, which is synchronized with political cycles where spending departments bid for funds. Government economists are typically closely involved in this process. These national and international issues are important but not the focus of the present discussion. Rather, the focus is on the local level of decision-making (e.g., local authority, state, and city), an area that health economics has paid relatively little attention to, thus far.

Local decision makers increasingly have a culture where different sectors and agencies operate together in partnership working. This provides a real opportunity for 'horizontal' coordination across sectors (and ideally 'vertical' coordination between levels of the system). For instance, the Public Health Agency of Canada has expanded its remit from coordination of interventions to prevent infectious disease outbreaks, to developing a vision that seeks to harness the social determinants of health to protect, maintain, and improve population health, more generally.

Policy coordination needs supporting institutional structures. There are promising examples from across the world that this is happening. For instance, in the UK, the Department for Health in England has devolved responsibility for public health to local levels with budgets being transferred accordingly. Public health can now be considered along with the wide range of other public sector concerns. Existing decision-making forums, such as 'One Place,' can also facilitate the development of common objectives that all sectors work toward. Another example is an innovation in Scotland called the Single Outcome Agreement, which provides joint targets that policies need to demonstrate progress against. A third example is the 2009 Australian National Partnership Agreement on Preventive Health. This illustrates the use of a national-level institutional structure to support local public health policy coordination, encouraging both horizontal coordination between sectors and vertical coordination between local and national levels.

The emerging notion of a 'systems approach' to public health policy-making recognizes the multiple social determinants

of health, and the interaction between policies, and looks for better ways to coordinate. For instance, health inequalities are the result of a system of influences, in which social disadvantages often cluster on the same groups in society, such as poor education, low employment, poor housing, and high rates of crime. It has therefore been argued that policies to reduce health inequalities should be developed (and evaluated) as 'multisectoral packages' rather than 'sector specific interventions.'

How to Encourage Intersectoral Alignment: The Role of Economic Evaluation

The Need to Take a Societal Perspective in Evaluation

From first principles, the fundamental problem economics is concerned with is scarcity: wants are infinite, but means are finite. The overarching purpose of normative economics is to inform the allocation of scarce resources with a view to improving 'social welfare' (variously known in other disciplines and contexts as 'social value', or 'the social good', or 'the public interest').

Prioritization is inevitable and normative economics is concerned with how to do this explicitly and rationally, given the information available. According to one of the founding fathers of health economics, Anthony Culyer, economic analysis ought to:

> "...identify relevant options for consideration; enumerate all costs and benefits to various relevant social groups; quantify as many as can be sensibly quantified; not assume the unquantified is unimportant; use discounting where relevant to derive present values; use sensitivity analysis to test the response of net benefits to changes in assumptions; and look at the distributive impact of the options"

Therefore, the challenge of greater intersectoral alignment, discussed previously, is congruent with the first principles of health economics.

An Integrated Approach to Economic Evaluation: Combining 'Broad' and 'Narrow' Approaches

The approach outlined by Culyer fits most comfortably within the general method of economic evaluation known as 'CBA,' which seeks to quantify all relevant social costs and benefits, and value them in the common currency of money. However, the practical reality is that decision makers in particular policy sectors are not primarily incentivized to make decisions purely on the basis of increasing general social welfare. Therefore, simply reporting net social welfare impacts is not ideally suited for decision makers. Therefore, the challenge is to take an approach to economic evaluation that satisfies the immediate sector specific concerns of decision makers, but also demonstrates the wider impacts of decisions and pinpoints where better coordination can lead to improvements in social welfare.

It is suggested here that an integrated approach to economic evaluation should be taken, where different approaches can be seen as complementary, rather than necessary competing. The starting point could be to undertake a CCA, which is essentially a social accountancy exercise, detailing the major impacts that result from an intervention. Economists can then use the CCA to develop the outcomes that different end-users are interested in **Figure 2**.

First, as the funding sector(s) may be primarily interested in the provision of specific outputs or amenities, the economic evaluation can take a 'narrow' CEA approach and simply report cost per unit of output – for example, cost per unit of social housing of a required standard. This evidence can be used to inform technical efficiency; to identify interventions that provide a certain output, at least cost.

Second, the consequential impacts on health can be evaluated within a CEA that focuses on a generic measure of health. The CCA could include a relevant generic health related quality of life (HRQoL) questionnaire, such as SF-12 or EQ-5D measuring how HRQoL has changed following an intervention(s) on a ratio scale between 0 (for zero health) and 1 (for full health). Responses are then weighted, by population preferences regarding the desirability of different states, to generate a single score of 'preference-weighted' HRQoL. This summary score is sometimes referred to as a 'health utility' and can be used on its own in evaluation, or used to weight length of life to generate quality-adjusted life-years, or disability adjusted life expectancy. The suggested reading list at the end includes an encyclopedia entry which discusses the methods used to generate the weights in deriving the health utility score (e.g., time tradeoff and standard gamble).

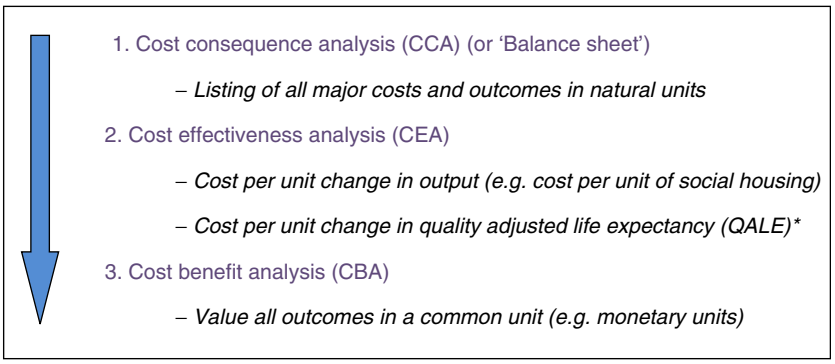


Figure 2 Integrated approach to economic evaluation: combining approaches.

Third, a CBA can then be conducted, which values all major outcomes detected in the CCA to then estimate net social benefit. The most common approach in CBA is to value all outcomes in financial terms. The suggested reading list includes an encyclopedia entry that discusses the methods used to attach financial values to outcomes (such as contingent valuation or discrete choice experiments). This 'broad' approach is then intended to estimate the overall social worth of alternative interventions.

Overall, combining approaches (CCA, CEA, and CBA) would have the strength of explaining how different stakeholder interests are related to one another, and how they are valued as part of overall social welfare. In this respect, the approaches of CCA and CEA can be seen as 'nested' within the overall framework of CBA.

It is important to recognize that producing economic evidence is only the first step in how economists can influence priority setting. The next step is to help decisionmakers to translate evidence into policy.

Translating Evidence into Policy: The Role of Priority Setting Tools

This penultimate section first explains why economic evaluation, while necessary, is rarely sufficient for decisionmakers when setting priorities: economic evidence is only one consideration. It then discusses how economists can help frame the priority setting process, helping to bring stakeholders together to articulate objectives, intervention options and value judgements. Crucially, this process should articulate decision-making criteria, so that economic evidence can then be used systematically alongside other relevant considerations.

The Two Key Principles in Priority Setting

There are two key economic principles that underlie priority setting from an economics perspective. The first is 'opportunity cost': when investing resources in one area, the most relevant cost for the decisionmaker to consider is the opportunity for benefit that is forgone because those resources are not invested elsewhere. The second is that of the 'margin': when changing the resource mix, the most relevant costs and benefits for the decisionmaker to consider are the marginal costs and benefits resulting from the proposed change in the resource mix, rather than the average or total costs and benefits of all the historical resources used. The concept of the margin is important regardless of whether budgets are changed or remain the same. If additional resources are made available, the key is to use the evidence to invest in the options offering best value. If the budget is decreasing, then the challenge is disinvestment, and budget should then be taken from interventions that provide the least value. Even with static budgets, there may be scope for reallocation to produce outcomes more efficiently. Economic evaluation is intended to provide this information to make explicit the costs and benefits of alternative courses of action, and the impacts of shifting resources at the margin. However, rarely will economic evidence be immediately translated into decisions. It is important to discuss why this is the case.

Economic Evidence Is Necessary but Not Sufficient for Priority Setting

For economic evidence to be sufficient for policymakers to use as the sole basis for making policy decisions, five conditions would need to be satisfied. First, the decisionmakers involved would need to clearly articulate and agree on the policy objective(s). This then would allow economists to develop an appropriate generic outcome measure, which all alternative courses of action can be measured against. Second, the valuation of outcome measure(s) would need to incorporate all relevant ethical considerations for decision-making, including considerations of fairness or equity as well as considerations of efficiency in maximizing the sum total of social benefits net of social opportunity costs. Third, the methods used by economic evaluation would need to be fully trusted by decisionmakers. Fourth, there would need to be no additional political constraints on decisionmakers beyond considerations of efficiency and equity. Fifth, economic evidence would need to be available to use. If these conditions were satisfied, then the role of policymakers would be largely passive. That is, once the initial objectives were articulated, economic evaluation could then produce a final all-thing-considered policy recommendation, which could determine the policy decision.

These conditions are unlikely to hold in practice. First, given that the scope of public health is multisectoral, there are likely to be competing objectives and value judgements. Different stakeholders are incentivized to produce different outputs. Second, equity has not been properly addressed in economic evaluation so far. Rather, units of benefit are typically valued equally regardless of which groups in society are affected. This is clearly a problem for public health, where many interventions are in fact delivered primarily in an attempt to reduce inequality (e.g., social housing for deprived communities). This suggests that a unit of benefit for a deprived individual can in some contexts be valued higher than the same unit of benefit for a less deprived individual. The issue of equity is becoming a key research focus for health economics, and the reading list refers to relevant articles in the encyclopedia. Third, decisionmakers typically lack the specialist expertize to fully understand economic evaluation methods and findings, and as a result may view economic evidence with suspicion. This is an area of contention, however. Distilling the impacts of an intervention into a single index is important to enable direct comparison of the impacts of different interventions, but perhaps economists need to improve the communication with decisionmakers to foster greater trust and reliance on the academic peer review process to ensure that methods are appropriate and fit-for-purpose. Fourth, decisionmakers often have to balance economic concerns with institutional and political concerns. Political concerns may result in certain decisions being taken largely in the absence of evidence, driven by opinions, values, and political constraints. Institutional concerns relate to both 'inhouse politics' and 'lags' in policy-making, where resources can rarely be instantly transferred between uses. For instance, services often involve a precommitment to funding over certain time periods and involve contracted and skilled staff who may not be easily transferred to alternative uses. This slows the process of improving allocative efficiency. Fifth, there is a general lack

of economic evidence regarding investments in the social determinants of health. To date, health economics has overwhelmingly concentrated on interventions within the health sector.

Furthermore, there are real difficulties in developing robust economic evidence. Many public health interventions are complicated in the sense that they are often multicomponent, making it difficult to identify active ingredients and to distinguish between a good intervention and poor implementation. Interventions are also complex in the sense that they can interact with local context (e.g., past and present interventions). In effect, context is an effect-modifier. Also, interventions may have the greatest impact over the long-term and even intergenerationally (e.g., urban regeneration). These common features can cause difficulties in establishing both causality (e.g., the opportunity for randomized trials is limited) and the generalizability of evidence, given that context can vary substantially between settings. The interaction of economic evidence with context provides an opportunity for interdisciplinary research in the future in order to improve both the generation and generalizability of evidence. Further, economics has a distinctive role to play in addressing the importance of context by analyzing how individual and organizational choices and behavior are likely to respond to changes in context-specific incentives and constraints.

Priority Setting as a Management Process

Priority setting is essentially a management process. This process needs to balance a wide range of concerns given that stakeholders often have competing objectives and diverse political and institutional constraints, and given the difficulties of generating robust evidence regarding likely policy outcomes. Owing to lack of evidence, the prioritization process is often fundamentally driven by value judgements. The priority setting process grows even more challenging when we consider the challenge of coordination between multiple policy sectors.

Priority Setting Tools: Framing the Decision Process

The priority setting process can often lack transparency and accountability. Policymakers themselves have expressed frustration regarding a lack of priority setting frameworks that they can use to guide decisions and enhance the credibility of resource allocation decisions. There is an opportunity for economists to help apply (and further develop) frameworks to steer decisionmakers through the process of priority setting, in addition to the generation of economic evidence.

There are a variety of priority setting tools that have been developed over the past 40 years, often as part of an interdisciplinary process. So despite the frustrations of policymakers, the issue may be more of awareness and application of existing tools – from both policymakers and perhaps economists too. Two of the most common tools are program budgeting marginal analysis (PBMA) and multicriteria decision-making. The rationale for using such tools is similar: to make explicit and improve the transparency and accountability of the priority setting process. For illustration, PBMA can be discussed briefly.

Table 1 Program budgeting marginal analysis (PBMA)

The five key steps in PBMA
1. What resources are available in total?
2. In what ways are these resources currently spent?
3. What are the main candidates for more resources and what would be their effectiveness and cost?
4. Are there any areas which could be provided to the same level of effectiveness but with the less resources, so releasing those resources to fund candidates from (3)?
5. Are there areas which, despite being effective, should have less resources because a proposal from (3) is more effective (for s spent)?

Source: Adapted from Mitton, C. and Donaldson, C. (2004). Health care priority setting: Principles, practice and challenges. *Cost Effectiveness and Resource Allocation* **2**(1), 3.

PBMA has been used in mainly in the UK, Canada, Australia, and New Zealand. Further, it has mainly been applied within the health sector. However, PBMA could in principle be used in any priority setting process and to coordinate multiple sectors.

An advisory panel is normally established with key stakeholders to identify the aims and scope of the priority setting exercise. Thereafter, five simple steps are in the process that is aimed at ultimately identifying areas for investment and disinvestment. The first step is to articulate the resources available for consideration in a reallocation exercise. The second step is to map out how existing resources are currently spent. The third step is to then identify the main candidates for more resources that offer the greatest value for money. The fourth step is to look at ways by which existing resources can be spent more efficiently to free up resources for Step 3 (Table 1).

The fifth step is to then make comparisons across spending areas and transfer resources if the interventions identified in Step 3 offer greater value.

Ideally, economic evidence would exist and be comprehensive enough to inform steps 3–5. However, often this is not the case or there are additional political or institutional concerns. Therefore, a key issue is for stakeholders to develop explicit decision-making criteria. This may include things such as health gain, access, innovation, sustainability, staff retention/recruitment, and system integration. This is where multi-criteria decision analysis (MCDA) can be nested within a PBMA approach to help decisionmakers choose between options. MCDA essentially involves four main steps: identifying interventions; identifying evaluation criteria; measuring interventions against the criteria; and combining criteria scores using a weighting to produce an overall assessment of each intervention.

Applying Priority Setting Tools

In effect, there are three potential applications of priority setting tools. The first is to determine the initial funding to individual sectors. If health improvement and tackling health inequality are priorities, then funding nonhealth sector interventions may be a more productive approach than health

sector interventions. Second, there is scope to use tools for the reallocation of funding within sectors to improve the efficiency of delivering outputs. This is particularly important when budgets may be under pressure due to a tighter fiscal environment. Third, these tools can in principle be used to coordinate sectors. This is where it becomes important for economic evidence to take an 'integrated approach' (combining different approaches to economic evaluation) and demonstrate to policymakers in different sectors the impacts of their decisions on one another, and the implications for health and overall social welfare. Through an explicit priority setting process, sectors can either compensate each other for the impacts of policies on one another, or ideally coordinate policies to create synergies and promote overall social value.

Conditions for Successful Priority Setting

For priority setting exercises to be successful certain conditions are required. Key amongst these is leadership. There needs to be willingness and commitment by leaders within organizations to the process and to ensure that resource reallocation can and does actually take place. Without leadership, the process can lack credibility. Priority setting exercises can also be time-consuming and involve senior staff in organizations. The opportunity to undergo these exercises may be limited to the start to the next budget cycle. Further, there needs to be a willingness to repeat the exercise, as experience has shown that organizations need a learning-by-doing period for the priority process to improve and develop credibility.

Overall, priority setting is inherently a messy process, involving economic, political, and institutional concerns. Priority setting tools can help willing participants to make explicit the decision process, articulate all issues and improve the rationality and accountability of resource allocation decisions. Tools such as PBMA can shape the decision process to accord with economic principles, to use the available economic evidence and make gains in technical and allocative efficiency.

Conclusions

This article has considered how economists can best engage with the 'healthy public policy' agenda, which is concerned with the social determinants of health where nonhealth sectors are considered to have significant impacts on population health and health inequalities. The practical challenge is how to generate and then translate economic evidence into decision-making in nonhealth policy sectors where, at present, health impacts are often considered as 'byproducts'. Given intersectoral impacts are a common feature of most policies, the wider challenge is to facilitate the process of intersectoral coordination and alignment toward improvements in overall social welfare, where health is just one important element.

By taking an integrated societal approach, economists can produce a consistent body of evidence that is commensurate both with the most pressing objectives of funding sectors and with the first principles of economics. The aim is to help policymakers move beyond narrow sector-specific perspectives and take decisions to improve overall social welfare. An integrated approach can begin with a CCA and then convert outcomes into relevant cost effectiveness measures (both cost per unit of output and cost per generic unit of health gain), and then value all outcomes consistently within a CBA. In this sense, the seemingly different approaches of economic evaluation can be viewed as complimentary, where CCA and CEA are 'nested' within an overall CBA.

Priority setting tools can then be used to facilitate the translation of evidence into decision-making. Tools such as PBMA are consistent with a societal approach and can help frame the scope of the priority setting exercise, facilitate stakeholders coming together, and make explicit the decision-making criteria so that evidence can be used systematically.

Overall, economics has much to offer public health to help facilitate intersectoral alignment so that decisionmakers take account of wider social determinants of health. Equally, public health has much to offer health economics; providing an opportunity to rediscover the societal approach where the ultimate aim of economics is to allocate scarce resources for the improvement of overall social welfare, wherein health is just one important element.

See also: Disability-Adjusted Life Years. Economic Evaluation of Public Health Interventions: Methodological Challenges. Equality of Opportunity in Health. Ethics and Social Value Judgments in Public Health. Health and Its Value: Overview. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Public Choice Analysis of Public Health Priority Setting. Public Health: Overview. Quality-Adjusted Life-Years. Valuing Health States, Techniques for. Willingness to Pay for Health

Reference

Mitton, C. and Donaldson, C. (2004). Health care priority setting: Principles, practice and challenges. *Cost Effectiveness and Resource Allocation* **2**(1), 3.

Further Reading

Bernier, N. F. (2007). Breaking the deadlock: Public health policy coordination as the next step. *Healthcare Policy* **3**(2), 117–127.

Coast, J. (2004). Is economic evaluation in touch with society's health values? *British Medical Journal* **329**(7476), 1233–1236 (and the responses to this article).

Cookson, R. and Claxton, K. (eds.) (2012). *The humble economist: Tony Culyer on health, health care and social decision making. York*: University of York and London: Office of Health Economics.

Dahlgren, G. and Whitehead, M. (1991). *Policies and strategies to promote social equity in health.* London: Institute of Fiscal Studies. Available at: http://ideas.repec.org/p/hhs/ifswps/2007_014.html (accessed 01.11.12).

Donaldson, C. (2011). *Credit crunch health care: How economics can save our publicly funded health services.* Chicago: University of Chicago Press.

Dionne, F., Mitton, C., Smith, N. and Donaldson, C. (2009). Evaluation of the impact of program budgeting and marginal analysis in Vancouver Island Health Authority. *Journal of Health Services Research and Policy* **14**(4), 234–242.

Hauck, A., Smith, P. C. and Goddard, M. (2003). The economics of priority setting for health care: A literature review. Washington: World Bank; 2003. Available at: http://siteresources.worldbank.org/HEALTHNUTRITIONANDPOPULATION/Resources/281627-1095698140167/Chapter3Final.pdf (accessed 01.11.12).

Joffe, M. and Mindell, J. (2004). A tentative step towards healthy public policy. *Journal of Epidemiology and Community Health* **58**(12), 966–968.

Kelly, M. P., McDaid, D., Ludbrook, A. and Powell, J. (2005). Economic appraisal of public health interventions. Briefing paper; Health Development Agency (London: part of the UK's National Institute of Clinical Excellence). Available at: http://www.cawt.com/Site/11/Documents/Publications/Population%20Health/Economics%20of%20Health%20Improvement/Economic_appraisal_of_public_health_interventions.pdf (accessed 01.11.12).

Leischow, S. J., Best, A., Trochim, W. M., et al. (2008). Systems thinking to improve the public's health. *Americal Journal of Preventive Medicine* **35**(2 Suppl), S196–S203.

Marsh, K., Dolan, P., Kempster, J. and Lugon, M. (2012). Prioritizing investments in public health: A multicriteria decision analysis. *Public Health* 1–7.

McQueen, D., Wismar, M., Lin, B., Jones, C. M. and Davies, M. (eds.) (2012). *Intersectoral governance for health in all policies*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Mortimer, D. (2010). Reorienting programme budgeting and marginal analysis (PBMA) towards disinvestment. *BMC Health Services Research* **10**, *288*.

Shiell, A., Hawe, P. and Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *British Medical Journal* **336**(7656), 1281–1283.

Sibbald, S. L., Singer, P. A., Upshur, R. and Martin, D. K. (2009). Priority setting: what constitutes success? A conceptual framework for successful priority setting. *BMC Health Services Research* **9**, 43.

Wanless, D. (2004) *Securing good health for the whole population*. UK: HM Treasury. Available at: http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4074426 (accessed 01.11.12).

Relevant Websites

http://www.ncchpp.ca/en/
Canadian National Collaborating Centre for Healthy Public Policy.

http://www.gcph.co.uk/latest/news/296_new_blog_series-economics_of_public_health
Glasgow Centre for Population Health (GCPH).

http://www.scotland.gov.uk/Topics/Government/local-government/delperf/SOA
Scottish Government.

http://oneplace.audit-commission.gov.uk/Pages/default.aspx
UK Government.

http://www.local.gov.uk/web/guest/media-centre//journal_content/56/10171/3374673/NEWS-TEMPLATE
UK Government.

http://fuseopenscienceblog.blogspot.co.uk/2012/06/happy-birthday-health-economics-from.html
University of Newcastle.

http://www.who.int/healthpromotion/conferences/previous/adelaide/en/index1.html
WHO Adelaide Recommendations on Healthy Public Policy.

http://www.who.int/sdhconference/declaration/en/index.html
WHO Rio Political Declaration on the Social Determinants of Health.

Private Insurance System Concerns

K Simon, Indiana University, Bloomington, IN, USA, and National Bureau of Economic Research, Cambridge, MA, USA

Introduction

As compared to public insurance systems, private systems face a unique set of market problems that have occupied a central position in health economics research. These ethical and efficiency problems that arise in market transactions plague national insurance systems to a much lesser degree and involve different solutions. This article first considers the reasons why private insurance systems exist in their current forms; next, it examines the ways in which private and public systems attempt to solve market failures; finally, it ends with a discussion of regulations related to perceived problems in the private market.

Healthcare appears in some ways a very well-suited market for private insurance transactions, as it involves high and unpredictable costs for a large segment of the population. However, numerous problems can potentially lead to classic market failures or otherwise societally unacceptable outcomes in a private health insurance system. The efficiency related problems of such a system include adverse selection due to a lack of perfect information and moral hazard due to price responsiveness in the demand for medical care. There are also several ethical questions that surround the affordability of insurance for the poor and high premiums for the sick, which lead to private-system regulations that borrow features of a public system. As expected, private systems maintain a delicate balance between ethical and efficiency concerns: various practices adopted by private insurance markets in response to adverse selection and moral hazard problems have heightened ethical concerns, while government policies adopted within private systems designed to alleviate ethical concerns have sometimes had market failure consequences.

Purely public systems, described in other articles in this Encyclopedia in terms of the history of health insurance in developed countries (John Murray) and a comparison between such insurance systems (Victor Rodwin), are distinct from the largely private system that exists in the US (see the article by Tim Jost on the history of health insurance in the US). Yet, as a thought exercise, one could imagine a private market, that under certain conditions recreates the essential elements of social insurance for healthcare, providing cradle-to-grave mandatory insurance that is financed by progressive taxation.

A Private-System Equivalent of Standard Public Insurance

Even with private insurers selling health insurance, it is theoretically possible that all policies offered are based on lifetime contracts, purchased at birth and priced uniformly. What leads such a scenario to be infeasible? For one, there is no mechanism in the private market to retain the feature of equity present in a public system through progressive taxation based financing. Even if lifetime policies were available, parents would have to purchase them for their children based on private funds. As long as there was an unequal distribution of incomes, there would be some level of uninsurance.

Even among families whose lifetime expected incomes could pay for lifetime insurance policies for their children, liquidity constraints and borrowing market imperfections would preclude them from paying for a policy that covers a long timeframe and thus would prevent the system from being considered universal and mandatory. One could design a lifetime policy with periodic payments, as is done for long-term care (see the article by Tamara Konetzka) and life insurance. A version of this long-term policy proposal was first considered by Cochrane. For several reasons including lack of contract enforceability, long-term insurance contracts are not available. Requiring periodic payments incentivizes the customers to not delay purchase until their health deteriorates and creates incentives on the part of insurers to reinterpret or rescind sales to those whose health has deteriorated since the initial contract was established. This leads to selection issues, which are one main form of difficulty experienced by a private system.

Government regulations are another reason that long-term policies do not exist in health insurance. Some regulations in current-day private systems attempt to create an equitable financing system by providing subsidies and supplemental public systems that pay for those with lower incomes. In other ways, community rating and guaranteed-issue-type regulations redistribute wealth from the healthy to the sick and could actually have regressive elements, as age and health status are correlated negatively, whereas age and income are correlated positively. Because of the risk of insurance companies refusing to cover those who experience a negative health shock after years of being continuously covered (revision of risk), regulations such as guaranteed renewability, protection of preexisting conditions coverage, and portability laws have arisen to protect consumers. These regulations address equity problems but could themselves lead to efficiency concerns. For example, guaranteed issue and community rating without a strong mandate for purchasing coverage could lead to worsened adverse selection and instability in insurance markets, as has occurred in the individual health insurance market in New York in the past two decades. The availability of publicly financed health insurance for low-income families could lead to reductions in the private provision of health insurance, as has been pointed out in the case of Medicaid expansions in the US.

Some solutions to problems affiliated with a private health insurance system come from the private market itself rather than from regulation. The fact that employer groups are the main organizing form of health insurance provision in the US helps mitigate problems with adverse selection, especially among large employers who provide stable pools for insurers. Private insurance plans also impose a guarantee issue period in a plan year; enrollees must select coverage within a certain window of time within the US or else forgo coverage for that

entire plan year. If employees were able to select coverage at any point in the year, the system would suffer from greater adverse selection.

A Model of a Pure Private Health Insurance System Absent Market Problems

In addition to considering the reasons why a private system replica of the standard public insurance model does not exist, it is important to consider why problems arise in a private system for health insurance. In theory, private markets could solve the risk inherent in medical care demand provided the correct conditions exist. John Nyman's article in this Encyclopedia on theory of demand provides further details on the welfare implications of health insurance. Suppose that health insurance is available each year to risk-averse individuals who face identical risks and are inelastic in their demand for medical care. In such a setting, optimal risk protection would be full insurance, and the price of insurance would be above or at an actuarially fair price, depending on the degree of customer risk aversion, loading costs, and degree of competition among sellers of insurance.

Problems of Imperfect Information

The world described above does not match reality in many ways. For one, individuals are not identical in their risk profiles, and insurance sellers cannot easily discern this information. At any price, insurers would find that those whose probability of needing coverage is higher than the population average would be more likely to buy insurance, causing insurers to experience 'adverse selection death spirals.' Seminal work addressed the problems of asymmetric information in insurance markets. Other economists developed insurance models to consider the equilibria that could exist in health insurance markets where hidden information is held by two groups of insurance customers; they concluded that the information imperfection could lead to loss of welfare. If insurers are allowed to offer different insurance contracts, a separating equilibrium could occur in which the group with the lower but unverifiable risk gets partial insurance and the high-risk group gets full insurance. The high-risk group thus imposes an externality on the low-risk group. If insurers are only able to offer one insurance contract, the market could fail to exist altogether because of the unsustainability of a pooling equilibrium.

The importance of the insights from seminal authors writing on information problems in economics in general was recognized when in 2001 Joseph Stiglitz and two other economists jointly received the Nobel Prize in Economics. However, this early work left the reader with a rather pessimistic view of possible solutions to the information problem in the health insurance market. A large literature since then has discussed the conditions under which equilibria may exist, including cases in which consumers differed in risk aversion as well as in risk probabilities. Most recently, researchers have introduced the possibility that there could be advantageous selection in insurance markets whereby those who are lower risk or more risk averse purchase more insurance than those who are higher risk. This literature has shown that advantageous selection exists in the case of Medicare supplemental plans in the US. However, other empirical investigations have also found evidence in favor of adverse selection.

Private market solutions to adverse selection problems

As Arrow noted several decades ago, private markets may be able to solve information failures by finding ways to convey information to sellers. In the US, health insurance is provided by employers as the predominant form of insurance to nonelderly individuals. This results in more stable insurance markets than those available for individuals, because insurers understand that health is not the primary reason for the formation of employer groups. Thereby, both the fact that employment is a signal of one's health and the fact that the heterogeneity of the production function is a signal that a firm was not just formed for the purpose of purchasing health insurance enable insurers to be less on guard regarding adverse selection.

Health insurance tied to the workplace: How does this affect job mobility?

An issue related to the labor market that public systems need not consider is the relationship between the labor market and health insurance. In the US, it is plausible that because of the important role played by employers, those who value health insurance may not leave their job for one that pays higher wages but does not pay for health insurance. One way in which private markets lessen this problem is by 'portability' laws that make it easier for someone to change jobs and obtain insurance without serving new waiting periods.

Insurer practices to guard against adverse selection

Insurers are aware of their information disadvantage and attempt to gather as much information as possible on their customers' health profiles. Insurers aim to base their prices on the information gathered, and they sometimes refuse outright to sell a policy to someone with past health problems in a practice known as 'red lining.' Another insurer practice used to guard against adverse selection is the refusal to cover preexisting conditions, defined as conditions that were diagnosed or treated in the past so many years or months, for the first so many years or months of a policy.

Although in one way these practices can be seen as necessary for the functioning of insurance markets, these practices are often also seen as unfair because they cause those who are unhealthy to pay more for health insurance. As a result, the US private insurance system has undergone several changes to address the concern that insurer practices guarding against adverse selection lead to inequities that society does not find fully acceptable. The latest set of such changes is being made through the Affordable Care Act, but there is long history in the US of regulating terms of sale.

The ability of insurers to place preexisting condition exclusions and other antiadverse selection restrictions on policies sold in the American individual and small group insurance markets has been regulated heavily by states since the early 1990s. The most progressive state in this topic, New York, has since 1992 prohibited insurers from charging different premiums, differentiating plan characteristics, or denying the sale of insurance based on any health or demographic factors,

except to allow slightly different prices according to whether one lives in the northern or southern part of the state. This approach is known as pure community rating, and it is supplemented by guaranteed issue, guaranteed renewability, and limitations on preexisting conditions exclusions. Most other states have taken what is known as a modified community rating approach to the pricing restrictions, in which they allow some adjustments for age or gender. Although these practices are most prevalent in the individual and small-group markets, laws of this nature have been strengthened and applied to all health insurance markets by the federal US government through the Health Insurance Portability and Accountability Act of 1996 and further through the Affordable Care Act of 2010.

Long-Term Contracts for Insurance

As described when delineating the private market replica of the public system, even if insurance markets do not suffer from information problems, they may still have efficiency problems in the real world because of the relatively short term of insurance contracts. If all insurance was to be sold on a long-term basis (e.g., if it was purchased by parents on behalf of children at the point of conception or birth), adverse selection problems would be mitigated because there would be relatively little known information about the health of the customer at this point in their life. Time consistent health insurance asks why insurance contracts offered in the current market are 1 year in length at most, which causes difficulties in purchasing insurance the year after an illness manifests itself. The lack of long-term insurance contracts is suboptimal because it makes it impossible for people to insure against reclassification risk. That is, risk-averse individuals may want to protect themselves not just against the unpredictable costs of insurance in the next year, but also against the unpredictability in risk-rated premiums in the future that could result from unpredictability in health. Early research in the field acknowledged that the current lack of long-term contracts is rooted in several logistical issues, including the lack of court enforcement, the likelihood of contracts being reinterpreted after illness occurs, and regulation. The literature concluded that regulation is the main impediment to the existence of such markets.

Reclassification Risk

Even though long-term health insurance contracts do not exist, many policies are written with a clause called guaranteed renewability, which means that the policy will be available the next year, too. Of course, the risk protection offered depends on the extent to which insurers can change an individual's premiums if their health status changes. Recent work has addressed the value of guaranteed renewability clauses in allowing individuals who purchase insurance to be protected against the risk that they might become worse health risks over time.

Portability and Preexisting Condition Exclusions

Even if insurance policies are written to have a within-policy guarantee of issue through a guaranteed renewability clause, this may not be enough protection against reclassification risk in employment-based insurance systems such as the one in the US. That is, once an individual who has been insured through the policy of one employer decides to switch jobs, the insurer of the next employer may treat him or her as a new entrant. Any reclassification risk the customer had enjoyed may disappear. This is especially the case if the new form of employment is self-employment, in which case the new insurer may be the individual market. A closely related practice is preexisting condition exclusion. Those who are switching between insurers lose their risk protection because of clauses that reassess their premiums based on current health status and because policies can exclude coverage for preexisting conditions.

Moral Hazard

In another departure from the quintessential perfectly competitive insurance market, real-world healthcare demand is not insensitive to price. This behavior, known as moral hazard, simply means that we consume more when we are faced with lower prices; we react to economic incentives. Although full insurance would be the optimal insurance contract if demand for care were completely price inelastic, insurers use cost sharing to reduce the inefficient overuse (use beyond the point at which marginal benefit equals marginal cost to society) of medical care by those who are insured. The most compelling data on the price responsiveness of medical care consumption come from the RAND Health Insurance Experiment, which placed individuals randomly into plans with different cost-sharing structures and discovered that individuals adjusted their healthcare use accordingly.

In the standard economic model, moral hazard is a source of inefficiency, as it creates dead weight loss. In the optimal design of an insurance contract, one must weigh the benefits of risk protection against the costs of inducing moral hazard and dead weight loss. As RAND researchers have pointed out, the optimal contract will depend on risk aversion parameters as well as the elasticity of demand for different forms of medical care. Other work pointed out that there is an interesting parallel between optimal cost sharing to avoid moral hazard and the optimal taxation rules set out by Ramsey. In optimal taxation, we attempt to avoid the dead weight loss of induced behavioral changes by taxing inelastic behaviors more. By analogy, cost sharing could be lower in inelastically demanded care, implying that insurance would lower their prices to a greater extent without inducing moral hazard. This works especially well if services that are inelastically demanded are also the ones in which we most value risk protection.

An alternative view of moral hazard holds that in the face of income constraints it is not all welfare reducing. Writers have raised the concern that moral hazard may act to transfer income from those who are healthy to those who are sick in ways that we could consider efficient. This would happen to the extent that income constraints cause insured individuals to respond by using coverage and to the extent that society as a whole would want an individual to purchase care even when the out-of-pocket cost of doing so is lower than the price paid by society.

Public System Solutions to Moral Hazard Problems

Although in theory adverse selection concerns are mitigated in public systems that offer a single mandatory and free insurance option through taxpayer-based financing, the risk of moral hazard after becoming insured is possibly as high in such systems as it is in private systems. A pure public system might not have as much ability as a private system to pursue solutions that rest on cost sharing. Private systems rely on many forms of cost sharing in the form of fixed or variable copays and (increasingly) high deductibles. Insurance plan formularies for medication place higher-cost drugs with more competitors into high cost-sharing tiers. Although the tiering of inpatient-service cost sharing is not prevalent, private systems are not prevented from such tiering and may decide to develop such systems in the future to mitigate cost growth.

In contrast, it is generally believed that public systems use nonprice methods to ration care (in the economic sense of the word). For example, one may have to be on a waiting list for services that are considered non essential. Researchers have compared waiting times for elective surgery in 12 Organization for Economics Cooperation and Development nations and found that mean wait times more than 3 months are quite common. These delays typically result from the fact that hospitals are paid according to a global budget and use wait times to adjust supply to demand in the absence of a pricing system. The use of a global budget with either a cap or a target is a tool by which public systems institute a nonmarket reimbursement scheme that would control costs. In a target setting, there is a fixed quota of services with fixed fees but there are some additional (lower) fees provided for producing above the quota. The cap system establishes a total budget for a set period and reimbursements rates are calculated *ex post* depending on the quantity of services that were provided. Fan *et al*. provide a theoretical explanation for why a cap is more effective than a target at controlling the quality of services provided. Regardless of the specific type, global budgets are similar in spirit to the provision of capitated reimbursements to providers under managed care within a private insurance system, also global budgets apply at the population level, whereas capitation in managed care usually applies at the individual patient level.

An additional way by which public systems counteract moral hazard on the supply side is by using cost effectiveness in decisions to cover one type of medical technology over the other, when close substitutes are available. Public systems use a ratio of expected benefits to expected costs to prioritize decisions such as whether to place a certain drug on the national formulary, whereas within a private system a formulary usually uses differences in out of pocket costs to steer consumers toward certain medications. Within private systems, managed-care organizations also tend to use this type of information to some degree, but its use is much more widespread in public systems.

Moral hazard is also tempered within public systems by the fact that often only basic care is covered by the public system whereas more expensive treatments, and perhaps treatments with more price-elastic demand, are covered by private and voluntary supplemental systems. Often the supplemental system will also cover cost sharing imposed by the public system. The literature noted that in France, private payments account for a quarter of national spending. They show that within a mixed system, supplemental coverage may increase spending for the public system as well. Even if a supplemental plan were to only cover services not provided by the public system, that coverage could stimulate the additional use of public covered services, too, through complementarities. Robust evidence of spillover effects from supplemental private coverage to the public system has been shown to exist in the case of Medicare and Medigap plans in the US as well.

Moral hazard also affects the supply side in both private and public insurance systems. Some of this can be interpreted as supplier-induced demand, where information asymmetry allows providers to misrepresent the benefits of healthcare (Arrow) even more if that care is paid for by insurance companies rather than the patient; other forms of moral hazard can be seen in rises in technology adoption due to the reduced cost sharing introduced by insurance. The advent of Medicare in the US introduced a vast array of medical technology; this may be viewed as a dynamic response to moral hazard incentives.

As mentioned earlier, both private and public systems have used managed care type arrangements (capitation or global budgets) to counteract moral hazard. Another way that public and private systems can reduce moral hazard on the part of suppliers is by extending the Ramsey rule related insight on optimal patient cost sharing to optimal provider reimbursement setting. Researchers have presented a theory of optimal pricing for regulators, who can consider using principles of Ramsey pricing to reduce oversupply in settings where physicians can extend demand. This suggests that regulators should consider setting lower reimbursements for procedures with more demand inducement possibilities.

The above discussion has focused on *ex post* moral hazard, the behavior of consumers and providers after they are covered by insurance. *Ex ante* moral hazard refers to a dynamic form of response that occurs at the individual level when one knows that consequences of risky health behaviors may be mitigated by insurance coverage. In fact, public systems face greater risk of *ex ante* moral hazard because of their inability to price according to health status, regardless of whether those health conditions result from health behavior choices that provided instantaneous gratification at the expense of worse health later in life. Once again, the waiting times for elective procedures may be seen as partially counteracting this form of moral hazard, as does the banning of direct to consumer ads to consumers. Seeing attractive images associated with lifestyle drugs may increase awareness and demand among consumers to these drugs. Aside from the US, only one country (New Zealand) allows the advertising of medications to consumers.

Coverage for those who are unable to pay for private insurance

The largest social issue faced by private systems is probably the question of payment. For many reasons, a private market may not reach universal coverage. This is not an outcome predicted by insurance theory, where there should be an insurance policy by which everyone is fully insured. Assuming no moral hazard and no asymmetric information, the market should provide an opportunity for risk-averse insurance buyers and

risk-neutral sellers to find mutually beneficial policies that reflect at least the actuarial price of coverage.

This textbook model assumes that the risks faced do not exceed available resources; relaxing just this one assumption leads to a world in which individuals will go uninsured and presumably less than fully insure, incurring debt or receiving inadequate care when they fall ill. Additional reasons for lack of insurance also include myopia and other threats to rationality. Insurance involves payment now for problems that could occur in the future, and theories of limited rationality suggest there will be under investment in such expenses leading to *ex post* regret when someone is injured while uninsured. The literature on health insurance affordability considers approximately one-fourth to three-fourth of uninsured adults to be able to afford insurance, using various measures of affordability.

Other reasons for uninsurance have to do with whether free care is provided as the 'outside good.' The US provision that hospitals shall provide stabilizing care regardless of pay, the availability of charity care, and other public provisions could influence an individual's decision to remain uninsured, avoiding the payment of insurance premiums. It could also be that due to regulations or market institutional reasons, an individual may not find the policy described by the textbook model where price is relevant to that individual's risk profile and coverage provided is relevant to the risk events they face. For example, young adults may not find prices that reflect their low statistical probability of illness because policies are based on community rates and do not vary by age, or because insurance laws mandate that the policy include coverage for types of care that young customers are unlikely to use.

When the available policy is too expensive for an individual because of their level of income, because of imperfect rationality, because of the availability of charity care, or because of the lack of a full spectrum of insurance policy options, uninsurance results and becomes a societal problem leading to concerns of equity and to financial and psychological stress rooted in inadequate access to healthcare. A private insurance system must then decide whether certain populations will be placed into a public system as well as the means by which such care will be paid. In reality, such a public system may include substantial out of pocket costs, as in the case of the US Medicare.

Summary

Public and private systems that aim to insure individuals against the uncertain need for medical care face different issues. Specifically, private systems are much more likely than public systems to suffer from adverse selection and equity concerns. Both sectors risk moral hazards but take different approaches to solving them. This article starts with the question of why a private system does not mimic a standard public insurance system and achieve universal coverage through equitable methods of financing. Issues such as affordability, selection, moral hazard, and legal contract enforceability lead to private systems providing coverage on terms that are often viewed as socially unacceptable. Both regulations and private market solutions exist to counteract these problems. However, regulations could themselves exacerbate efficiency problems in the private market, leading to a delicate balance between intended and unintended consequences, and between risk pooling and moral hazard.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance in Developed Countries, History of. Health Insurance in the United States, History of. Health Insurance Systems in Developed Countries, Comparisons of. Long-Term Care Insurance

Further Reading

Arrow, K. (1963). Uncertainty and the welfare economics of medicare care. *American Economic Review* **53**, 941–973.

Besley, T. J. (1988). Optimal reimbursement health insurance and the theory of Ramsey taxation. *Journal of Health Economics* **7**(4), 321–336.

Buchmueller, T., Couffinhal, A., Grignon, M. and Peronnin, M. (2004). Access to physician services: Does supplemental insurance matter? Evidence from France. *Health Economics* **13**(7), 669–687.

Cochrane, J. (1995). Time consistent health insurance. *Journal of Political Economy* **103**(3), 445–473.

Ehrlich, I. and Becker, G. S. (1972). Market insurance, self-insurance, and self-protection. *Journal of Political Economy* **80**, 623–648.

Fan, C-P., Chen, K.-P. and Kan, K. (1998). The design of payment systems for physicians under global budget – An experimental study. *Journal of Economic Behavior & Organization* **34**, 295–311.

Fang, H., Keane, M. and Silverman, D. (2008). Sources of advantageous selection: Evidence from the Medigap. *Journal of Political Economy* **116**(2), 303–350.

Finkelstein, A. (2007). The aggregate effects of health insurance. *Quarterly Journal of Economics* **122**(1), 1–37. doi:10.1162/qjec.122.1.1.

Herring, B. and Pauly, M. V. (2006). Incentive-compatible guaranteed renewable health insurance premiums. *Journal of Health Economics* **25**(3), 395–417.

Madrian, M and Brigitte, C. (1994). Employment-based health insurance and job mobility: Is there evidence of job-lock? *Quarterly Journal of Economics* **109**(1), 27–54. MIT Press.

Manning, W. G. and Marquis, M. S. (1996). Health insurance: The tradeoff between risk pooling and moral hazard. *Journal of Health Economics* **15**(5), 609–640.

Nyman, J. A. (2003). *The theory of demand for health insurance*. Stanford, CA: Stanford University Press. ISBN: 9780804744881.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**(3), 531–537.

Rothschild, M. and Stiglitz, J. E. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**(4), 630–649.

Siciliani, L. and Hurst, J. (2005). Tackling excessive waiting times for elective surgery: A comparison of policies in twelve OECD countries. *Health Policy* **72**, 201–215.

Problem Structuring for Health Economic Model Development

P Tappenden, University of Sheffield, Sheffield, UK

Glossary

Conceptual model The abstraction and representation of complex phenomena of interest in some readily expressible form, such that individual stakeholders' understanding of the parts of the actual system, and the mathematical representation of that system, may be shared, questioned, tested, and ultimately agreed.

Design-oriented conceptual model Conceptual models which are focused on the consideration of alternative potentially acceptable and feasible quantitative model designs, to specify the model's evidence requirements and to provide a basis for comparison and justification against the final implemented model.

Disease logic model A problem-oriented conceptual model which sets out the disease-specific events and processes within the system in which the decision problem exists.

Problem-oriented conceptual model A form of conceptual model which is developed to understand the decision problem and the system in which that problem exists.

Problem structuring methods A set of formal methods developed within the field of Operational Research intended to develop consensus, structure, and make sense of complex or messy problems.

Service pathways model A problem-oriented conceptual model which sets out the elements of the service which are relevant to the system in which the decision problem exists.

Introduction

The economic evaluation of health care is a general framework for informing decisions about whether particular health-care technologies represent a cost-effective use of health-care resources. Commonly, the evidence required to inform a decision about the cost-effectiveness of a given set of competing health technologies is not available from a single source. The use of mathematical modeling can be used to support this decision-analytic framework thereby allowing the full range of relevant evidence to be synthesized and brought to bear on the decision problem (Briggs *et al.*, 2006). The process of developing a decision-analytic model is generally seen as being iterative, and requires the model developer to make a substantial number of choices about what should be included in a model and how these included phenomena should be related to one another. These choices take place at every stage of the model development process, and include choices about the comparators to be assessed, choices about which health states and sequences of events will comprise the model's structure, choices about which evidence sources should be used to inform the model parameters, and choices about statistical methods for deriving the model's parameters, to name but a few. Importantly the absence of perfect information through which to comprehensively validate a model means that there is rarely a definitive means through which to prospectively determine whether these choices are right or wrong. Instead, model development choices are made on the basis of subjective judgments, with the ultimate goal of developing a model which will be useful in informing the decision at hand.

Therefore, model development is perhaps best characterized as a complex process in which the modeler, in conjunction with other stakeholders, determines what is relevant to the decision problem (and at the same time, what can reasonably be considered irrelevant to the decision problem). This notion of relevance has a direct bearing on the credibility of a model and on the interpretation of results generated using that model. Failure to account for the complexities of the decision problem may result in the development of models which are "mathematically sophisticated but contextually naïve" (Ackoff, 1979). The development of useful mathematical models therefore requires more than mathematical ability alone: first, it requires the model developer to understand the complexity of the real system that the model will attempt to represent, and the choices available for translating this understanding of complexity into a credible conceptual and mathematical structure. It is perhaps surprising that while much has been written about the technical aspects of model development, for example, the statistical extrapolation of censored data and methods for synthesizing evidence from multiple sources, there is a comparative dearth of practical guidance surrounding formal processes through which an appropriate model structure should be determined. It is this complex and messy subject matter that forms the focus of this article.

The purpose of this article is not to rigidly prescribe how model development decisions should be made, nor is it intended to represent a comprehensive guide of 'how to model.' The former would undoubtedly fail to reflect the unique characteristics of each individual decision problem and could discourage the development of new and innovative modeling methods. Conversely, the latter would inevitably fail to reflect the sheer breadth of decisions required during model development. Rather, the purposes of this article are threefold:

1. To highlight that structural model development choices invariably exist;
2. To suggest a generalizeable and practical hierarchical approach through which these alternative choices can be prospectively exposed, considered and assessed; and

 doi:10.1016/B978-0-12-375678-7.01410-3

3. To highlight key issues and caveats associated with the use of certain types of evidence in informing the conceptual basis of the model.

The article is set out as follows. The article begins by introducing concepts surrounding the role and interpretation of mathematical models in general, and attempts to highlight the importance of conceptual modeling within the broader model development process. Following on from this, existing literature surrounding model structuring and conceptual modeling is briefly discussed. The article then moves on to suggest a practicable framework for understanding the nature of the decision problem to be addressed in order to move toward a credible and acceptable final mathematical model structure. A series of potentially useful considerations is presented to inform this process.

The Interpretation of Mathematical Models

A mathematical model is a "representation of the real world… characterized by the use of mathematics to represent the parts of the real world that are of interest and the relationships between those parts" (Eddy, 1985). The roles of mathematical modeling are numerous, including extending results from a single trial, combining multiple sources of evidence, translating from surrogate/intermediate endpoints to final outcomes, generalizing results from one context to another, informing research planning and design, and characterizing and representing decision uncertainty given existing information (Brennan and Akehurst, 2000). At a broad level, mathematical or simulation models in Health Technology Assessment (HTA) are generally used to simulate the natural history of a disease and the impact of particular health technologies on that natural history in order to estimate incremental costs, health outcomes, and cost-effectiveness.

All mathematical models require evidence to inform their parameters. Such evidence may include information concerning disease natural history or baseline risk of certain clinical events, epidemiology, resource use and service utilization, compliance/participation patterns, costs, health-related quality of life (HRQoL), survival and other time-to-event outcomes, relative treatment effects, and relationships between intermediate and final endpoints. However, the role of evidence is not restricted to informing model parameters. Rather, it is closely intertwined with questions about which model parameters should be considered relevant in the first place and how these parameters should be characterized. The consideration of how best to identify and use evidence to inform a particular model parameter thus first requires an explicit decision that the parameter in question is 'relevant,' the specification or definition of that parameter, and some judgment concerning its relationship to other 'relevant' parameters included in the model. This often complex and iterative activity is central to the process of model development and can be characterized as a series of decisions concerning (1) what should be included in the model, (2) what should be excluded, and (3) how those phenomena that are included should be conceptually and mathematically represented.

The need for these types of decisions during model development is unavoidable, rather it is a fundamental characteristic of the process itself. Although this activity already takes place in health economic model development, it is often unclear how this process has been undertaken and how this may have influenced the final implemented model. In practice, the reporting of model structures tends to be very limited (Cooper *et al.*, 2005) and, if present, usually focuses only on the final model that has been implemented. In such instances, the reader may be left with little idea about whether or why the selected model structure should be considered credible, which evidence has been used to inform its structure, why certain abstractions, simplifications, and omissions have been made, why certain parameters were selected for inclusion (and why others have been excluded), and why the included parameters have been defined in a particular way. This lack of systematicity and transparency ultimately means that judgments concerning the credibility of the model in question may be difficult to make. To produce practically useful guidance concerning the use of evidence in models, it is first important to be clear about the interpretation of abstraction, bias, and credibility in the model development process.

Credibility of Models

A model cannot include every possible relevant phenomenon; if it could it would no longer be a model but would instead be the real world. The value of simplification and abstraction within models is the ability to examine phenomena which are complex, unmanageable, or otherwise unobservable in the real world. As a direct consequence of this need for simplification, all models will be, to some degree, wrong. The key question is not whether the model is 'correct' but rather whether it can be considered to be useful for informing the decision problem at hand. This usefulness is directly dependent on the credibility of the model's results, which is, in turn, hinged on the credibility of the model from which those results are drawn. Owing to the inevitability of simplification and abstraction within models, there is no single 'perfect' or 'optimal' model. There may, however, exist one or more 'acceptable' models; even what is perceived to be the 'best' model could always be subjected to some degree of incremental improvement (and indeed the nature of what constitutes an improvement requires some subjective judgment). The credibility of potentially acceptable models can be assessed and differing levels of confidence can be attributed to their results on the basis of such judgments. The level of confidence given to the credibility of a particular model may be determined retrospectively – through considerations of structural and methodological uncertainty *ex post facto*, or prospectively – through the *a priori* consideration of the process through which decisions are made concerning the conceptualization, structuring, and implementation of the model.

Defining Relevance in Models

The purpose of models is to represent reality, not to reproduce it. The process of model development involves efforts to reflect those parts of reality that are considered relevant to the decision problem. Judgments concerning relevance may differ

between different modelers attempting to represent the same part of reality. The question of 'what is relevant?' to a particular decision problem should not be judged solely by the individual developing the model; rather making such decisions should be considered as a joint task between modelers, decision-makers, health professionals, and other stakeholders who impact on or are impacted on by the decision problem under consideration. Failure to reflect conflicting views between alternative stakeholders may lead to the development of models which represent a contextually naïve and uninformed basis for decision-making.

The Role of Clinical/Expert Input

Clinical opinion is essential in understanding the relevant facets of the system in which the decision problem exists. This clinical opinion is not only relevant, but essential, because it is sourced from individuals who interact with this system in a way that a modeler cannot. This information forms the cornerstone of a model's contextual relevance. However, it is important to recognize that health professionals cannot fully detach themselves from the system in which they practice; their views of a particular decision problem may be to some degree influenced by evidence they have consulted, their geographical location, local enthusiasms, their experience, and expertise, together with a wealth of other factors. Understanding why the views of stakeholders differ from one another is important, especially with respect to highlighting geographical variations. As such, the use of clinical input in informing models and model structures brings with it the potential for bias. Bias may also be sourced from the modeler themselves as a result of their expertise, their previous knowledge of the system in which the current decision problem, and the time and resource available for model development. Whenever possible, potential biases should be brought to light to inform judgments about a model's credibility.

Problem Structuring in Health Economics and Other Fields

It is important at this stage to note that although related to one another, there is a distinction between problem structuring methods (PSMs) and methods for structuring models. The former are concerned with understanding the nature and scope of the problem to be addressed, eliciting different stakeholders' potentially conflicting views of the problem and developing consensus, exploring what potential options for improvement might be available, and even considering whether a problem exists at all. There exist a number of methods to support this activity which have emerged from the field of 'soft' Operational Research; these include Strategic Options Design and Analysis (SODA) and cognitive mapping, Soft Systems Methodology (SSM), Strategic Choice Approach, and Drama Theory to name but a few. All stakeholders are seen as active 'problem owners' and each of their views are considered important. The emphasis of PSMs is not to identify the 'rationally optimal' solution, but rather to lay out the differing perceptions of the problem owners to foster discussion concerning potential options for improvement to the system. The value or adequacy of the PSMs is gauged according to whether they usefully prompt debate, with the intended endpoint being some agreement about the structure of the problem to be addressed and the identification and agreement of potential improvements to that problem situation. They do not necessarily assume that a mathematical model is appropriate or required. These methods are not discussed further here, but the interested reader is directed to the excellent introductory text by Rosenhead and Mingers, 2004.

Conversely, formal methods for model structuring, which relates principally to developing a conceptual basis for the quantitative model, remain comparatively underdeveloped, both in the context of health economic evaluation as well as in other fields. A recent review of existing conceptual modeling literature (Robinson, 2008) concluded that although conceptual modeling is 'probably the most important element of a simulation study,' there remains for the most part, a vacuum of research in terms of what conceptual modeling is, why it should be done, and how it may be most effectively implemented. Where formal conceptual modeling viewpoints have emerged, there is little consensus or consistency concerning how this activity should be approached.

This problem is particularly applicable in the field of health economics. Recently, a qualitative research study was undertaken to examine techniques and procedures for the avoidance and identification of errors in HTA models (Chilcott *et al.*, 2010). Interviewees included modelers working within Assessment Groups involved in supporting NICE's Technology Appraisal Program as well as those working for outcomes research groups involved in preparing submissions to NICE on behalf of pharmaceutical companies. A central aspect of these interviews involved the elicitation of a personal interpretation of how each interviewee develops models. These descriptions were synthesized to produce a stylized model development process comprising five broad bundles of activities (**Box 1** and **Figure 1**).

One particular area of variability between interviewees concerned their approaches to conceptual model development. During the interviews, respondents discussed the use of several approaches to conceptual modeling including

Box 1 Main stages in the model development process

1. Understanding the decision problem: Activities including immersion in research evidence, defining the research question, engaging with clinicians, decision-makers, and methodologists, and understanding what is feasible.
2. Conceptual modeling: Activity related to translating the understanding of the decision problem toward a mathematical model-based solution (Robinson, 2008).
3. Model implementation: Implementation of the model within a software platform.
4. Model checking: Activity to avoid and identify model errors. This includes engaging with experts, checking face validity, testing values, structure and logic, checking data sources, etc.
5. Engaging with decision: Model reporting and use by the decision-maker(s).

Source: Adapted from Chilcott, J. B., Tappenden, P., Rawdin, A., et al. (2009). Avoiding and identifying errors in health technology assessment models. *Health Technology Assessment* **14**(25), i–135, and Robinson, S. (2008). Conceptual modelling for simulation Part I: Definition and requirements. *Journal of the Operational Research Society* **59**, 278–290.

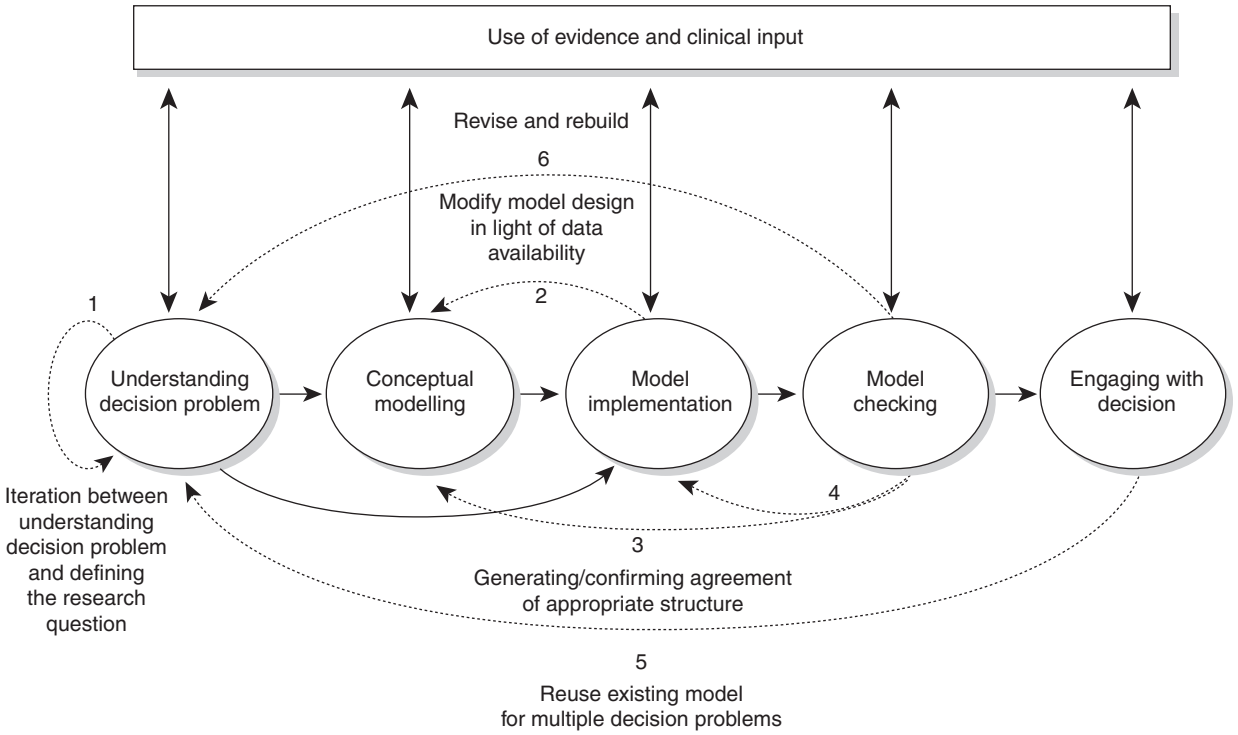


Figure 1 Stylized model development process. Reproduced from Chilcott, J. B., Tappenden, P., Rawdin, A., et al. (2010). Avoiding and identifying errors in health technology assessment models. *Health Technology Assessment* **14**(25), i–135.

documenting proposed model structures, developing mock-up models in Microsoft Excel, developing sketches of potential structures, and producing written interpretations of evidence. For several respondents, the model development process did not involve any explicit conceptual modeling activity; in these instances, the conceptual model and implementation model were developed in parallel with no discernable separation between the two activities. This is an important distinction to make with respect to model credibility and validation (as discussed in Section Definition and Purpose of Conceptual Modeling) and the processes through which evidence is identified and used to inform the final implemented model.

Definition and Purpose of Conceptual Modeling

Although others have recognized the importance of conceptual modeling as a central element of the model development process, it has been noted that this aspect of model development is probably the most difficult to undertake and least well understood (Chilcott *et al.*, 2010; Law, 1991). Part of the problem stems from inconsistencies in the definition and the role(s) of conceptual modeling, and more general disagreements concerning how such activity should be used to support and inform implementation modeling. The definition and characteristics of conceptual modeling are dependent on the perceived purposes of the activity. For the purpose of this document, conceptual modeling is taken as: "the abstraction and representation of complex phenomena of interest in some readily expressible form, such that individual stakeholders' understanding of the parts of the actual system, and the mathematical representation of that system, may be shared, questioned, tested, and ultimately agreed."

Although there is inevitable overlap associated with processes for understanding the decision problem to be addressed, conceptual modeling is distinguishable from these activities in that it is targeted at producing tangible outputs in the form of one or more conceptual models. In the context of health economic evaluation, conceptual model development may be used to achieve a number of ends, as highlighted in Box 2. Broadly speaking, these roles fall into two groups: (1) those associated with developing, sharing, and testing one's understanding of the decision problem and the system in which this exists and (2) those associated with designing, specifying, and justifying the model and its structure. Therefore it seems sensible to distinguish between problem-oriented conceptual models and design-oriented conceptual models; this distinction has been made elsewhere outside of the field of health economics (Lacy *et al.*, 2001). The characteristics of these alternative types of conceptual model are briefly detailed below. Both of these types of model may be useful approaches for informing the relevant characteristics of a health economic model.

Problem-oriented conceptual models: This form of conceptual model is developed to understand the decision problem and the system in which that problem exists. The focus of this model form concerns fostering communication and understanding between those parties involved in informing, developing, and using the model. In health economic evaluation, this type of conceptual model is primarily concerned with developing and agreeing a description of the disease and treatment systems: (1) to describe the current clinical

Box 2 The roles of conceptual modeling in health economic model development

Problem-oriented conceptual models

- To ensure that health professionals understand how the model will capture the impact of the interventions under consideration on costs and health outcomes.
- To ensure that the proposed model will be clinically relevant – that all relevant events, resources, costs, and health outcomes have been included and that these reflect current knowledge of disease and treatment systems.
- To ensure that the proposed model will meet the needs of the decision-maker.
- To provide a reference point during model implementation.
- To highlight uncertainty and variation between health-care practitioners.

Design-oriented conceptual models

- To provide a common understanding amongst those involved in model development regarding model evidence requirements before model implementation.
- To provide an explicit platform for considering and debating alternative model structures and other model development decisions before implementation (including the *a priori* consideration of structural uncertainties).
- To provide a reference point during model implementation.
- To provide the conceptual basis for reporting the methods and assumptions employed within the final implemented model.
- To provide a basis for comparison and justification of simplifications and abstractions during model development.

understanding of the relevant characteristics of the disease process(es) under consideration and important events therein; and (2) to describe the clinical pathways through which patients with the disease(s) are detected, diagnosed, treated, and followed-up. This type of conceptual model is therefore solely concerned with unearthing the complexity of the decision problem and the system in which it exists; its role is not to make assertions about how those relevant aspects of the system should be mathematically represented. The definition of 'what is relevant?' for this type of conceptual model is thus primarily dependent on expert input rather than the availability of empirical research evidence. In this sense, this type of conceptual model is a problem-led method of inquiry.

Design-oriented conceptual models: This form of conceptual model is focused on the consideration of alternative potentially acceptable and feasible quantitative model designs, to specify the model's anticipated evidence requirements, and to provide a basis for comparison and justification against the final implemented model. To achieve these ends, it draws together the problem-oriented conceptual views of relevant disease and treatment processes and interactions between the two. The design-oriented conceptual model sets out a clear boundary around the model system, defines its breadth (how far down the model will simulate certain pathways for particular patients and subgroups) and sets out the level of depth or detail within each part of the model. It therefore represents a platform for identifying and thinking through potentially feasible and credible model development choices before actual implementation. Within this context, the definition of 'what is relevant?' is guided by the problem-oriented models and therefore remains problem-led, but is mediated by the question of 'what is feasible?' given the availability of existing evidence and model development resources (available time, money, expertise, etc.).

Conceptual modeling activity, however defined, is directly related to model credibility and validation (Sargent, 2004). The absence of an explicit conceptual model means that a specific point of model validation is lost. As a model cannot include everything, an implemented model is inevitably a subset of the system described by the conceptual model. This hierarchical separation allows simplifications and abstractions represented in the implemented model to be compared against its conceptual counterpart, thereby allowing for debate and justification (Robinson, 2008). However, in order to make such comparisons, conceptual model development must be overt: the absence or incomplete specification of a conceptual model leads to the breakdown of concepts of model validation and verification. Without first identifying and considering the alternative choices available, it is impossible to justify the appropriateness of any particular model. Furthermore, without first setting out what is known about the relevant disease and treatment processes, the extent or impact of particular assumptions and simplifications cannot be drawn out explicitly. Therefore, the benefit of separating out conceptual modeling activity into distinct problem-oriented and design-oriented components is that this allows the modeler (and other stakeholders) to first understand the complexities of the system the model intends to represent, and then to examine the extent to which the simplifications and abstractions resulting from alternative 'hard' model structures will deviate from this initial view of the system. Figure 2 shows the hierarchical relationship between the real world, the problem- and design-oriented conceptual models, and the final implemented model.

Practical Approaches to Conceptual Modeling in HTA

This section suggests how conceptual modeling could be undertaken and which elements of model development activity should be reported. Practical considerations surrounding conceptual model development are detailed below with reference to a purposefully simple model to assess the cost-effectiveness of adjuvant treatments for a hypothetical cancer area. These considerations are intended to be broadly generalizeable to economic analysis within other diseases and conditions. It should be noted that the illustrative model is only intended to suggest how the alternative conceptual models forms may be presented and used. The problem-oriented model is divided into two separate conceptual model views: a disease logic model and a service pathways model.

Problem-Oriented Conceptual Modeling – Disease Logic Models

Figure 3 presents a simple example of a conceptual disease logic model for the hypothetical decision problem. The focus of this type of model is principally on relevant disease events and processes rather than on the treatments received. At each

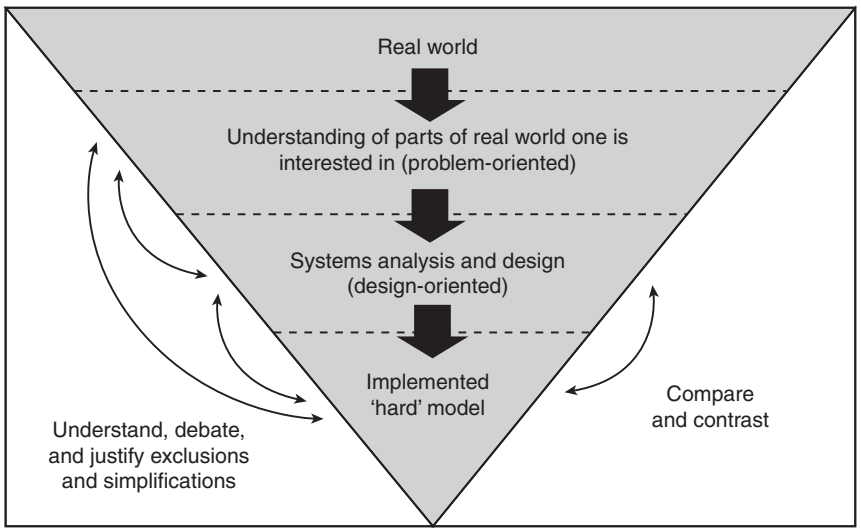


Figure 2 A hierarchy of models.

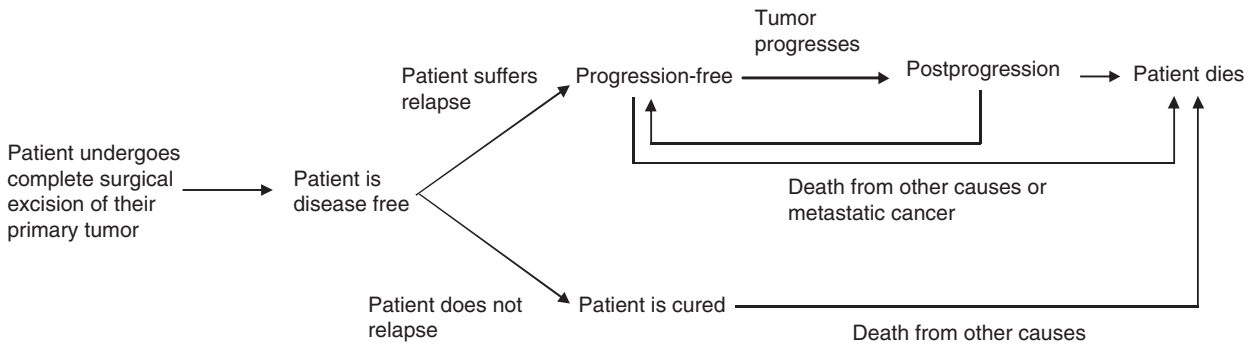


Figure 3 Illustrative disease logic model.

point in the pathway, the focus should therefore relate to an individual patient's true underlying state rather than what is known by health-care professionals at a particular point in time. It should be reiterated that this type of conceptual model does not impose or imply any particular decision concerning modeling methodology or appropriate outcome measures; it is solely a means of describing the relevant clinical events and processes within the system of interest. It should also be noted that such conceptual models should be accompanied by textual descriptions to support their interpretation and to capture and factors or complexities which are not represented diagrammatically.

The following nonexhaustive set of issues and considerations may be useful when developing and reporting this type of problem-oriented conceptual model:

Inclusion/Exclusion of Disease-Related Events

- What are the main relevant events from a clinical/patient perspective? Does the conceptual model include explicit reference to all clinically meaningful events? For example, could a patient experience local relapse? Or could the intervention affect other diseases (e.g., late secondary malignancy resulting from radiation therapy used to treat the primary tumor)?
- Can these relevant events be discretized into a series of mutually exclusive biologically plausible health states? Does this make the process easier to explain?
  - If so, which metric would be clinically meaningful or most clinically appropriate? Which discrete states would be clinically meaningful? How do clinicians think about the disease process? How do patients progress between these states or sequences of events?
  - If not, how could the patient's preclinical trajectory be defined?
- Do alternative staging classifications exist, and if so can/should they be presented simultaneously?
- Are all relevant competing risks (e.g., relapse or death) considered?
- For models of screening or diagnostic interventions, should the same metric used to describe preclinical and post-diagnostic disease states?
- Is the breadth of the conceptual model complete? Does the model represent all relevant states or possible sequences of events over the relevant patient subgroup's lifetime?

- What are the causes of death? When can a patient die from these particular causes? Can patients be cured? If so, when might this happen and for which states does this apply? What is the prognosis for individuals who are cured?

Impact of the Disease on HRQoL and Other Outcomes

- Is there a relationship between states, events, and HRQoL? Which events are expected to impact on a patient's HRQoL?
- Does the description of the disease process capture separate states in which a patient's HRQoL is likely to be different?
- Does the description of the disease process capture different states for prognosis?

Representation of Different-Risk Subgroups

- Is it clear which competing events are relevant for particular subgroups?
- Does the description of the disease process represent a single patient group or should it discriminate between different subgroups of patients?
- Are these states/events likely to differ by patient subgroup?

Impact of the Technology on the Conceptualized Disease Process

- Have all competing technologies relevant to the decision problem been identified?
- Can the conceptual model be used to explain the impact(s) of the technology or technologies under assessment? Do all technologies under consideration impact on the same set of outcomes in the same way?
- Are there competing theories concerning the impact(s) of the technology on the disease process? Can these be explained using the conceptual model?
- Does the use of the health technology result in any other impacts on health outcomes that cannot be explained using the conceptual disease logic model?

Problem-Oriented Conceptual Modeling – Service Pathways Models

Figure 4 presents an illustrative service pathways model for the hypothetical decision problem. In contrast to the disease logic model, the focus of the service pathways model is principally concerned with the health-care interventions received based on what is known or believed by health-care practitioners at any given point in time. Again, such conceptual models should be accompanied by textual descriptions to ensure clarity in their interpretation and to retain any complexity which is not or cannot be captured diagrammatically.

The following issues and considerations may be useful when developing and reporting this type of conceptual model:

Relationship between Risk Factors, Prognosis, and Service Pathways

- Is it clear where and how patients enter the service? Is it clear where patients leave the service (either through discharge or death)?
- Does the model make clear which patients follow particular routes through the service?
- Are there any service changes occurring upstream in the disease service which may influence the case-mix of patients at the point of model entry? For example, if surgical techniques were subject to quality improvement might this change patient prognosis further downstream in the pathway?
- Does the model highlight the potential adverse events resulting from the use of particular interventions throughout the pathway? What are these? Do they apply to all competing technologies under consideration?
- Are there any potential feedback loops within the system (e.g., resection → follow-up → relapse → re-resection → follow-up)?
- Which patients receive active treatment and which receive supportive care alone? What information is used to determine this clinical decision (e.g., fitness, patient choice)?

Distinction between What Is True and What Is Known

- How does the pathway change on detection of the relevant clinical events, as defined in the conceptual disease logic model? For example, at what point may relapse be detected?
- Is the occurrence of certain events likely to be subject to interval censoring?

Geographical Variations

- How do the service pathways represented in the model vary by geographical location or local enthusiasms? What are these differences and which parts of the pathway are likely to be affected most?

Nature of Resource Use

- What are the relevant resource components across the pathway and what is the nature of resource use at each point of intervention? For example, routine follow-up dependent on relapse status, once-only surgery (except for certain relapsing patients), cycle-based chemotherapy, doses dependent on certain characteristics, dose-limited radiation treatment, etc.
- Does the conceptual service pathways model include all relevant resource components?
- Which resources are expected to be the key drivers of costs?

Impact of the Technology on the Service Pathway

- Which elements of the conceptual model will the intervention under assessment impact on? For example, different costs of adjuvant treatment, different mean time in follow-up, different numbers of patients experiencing

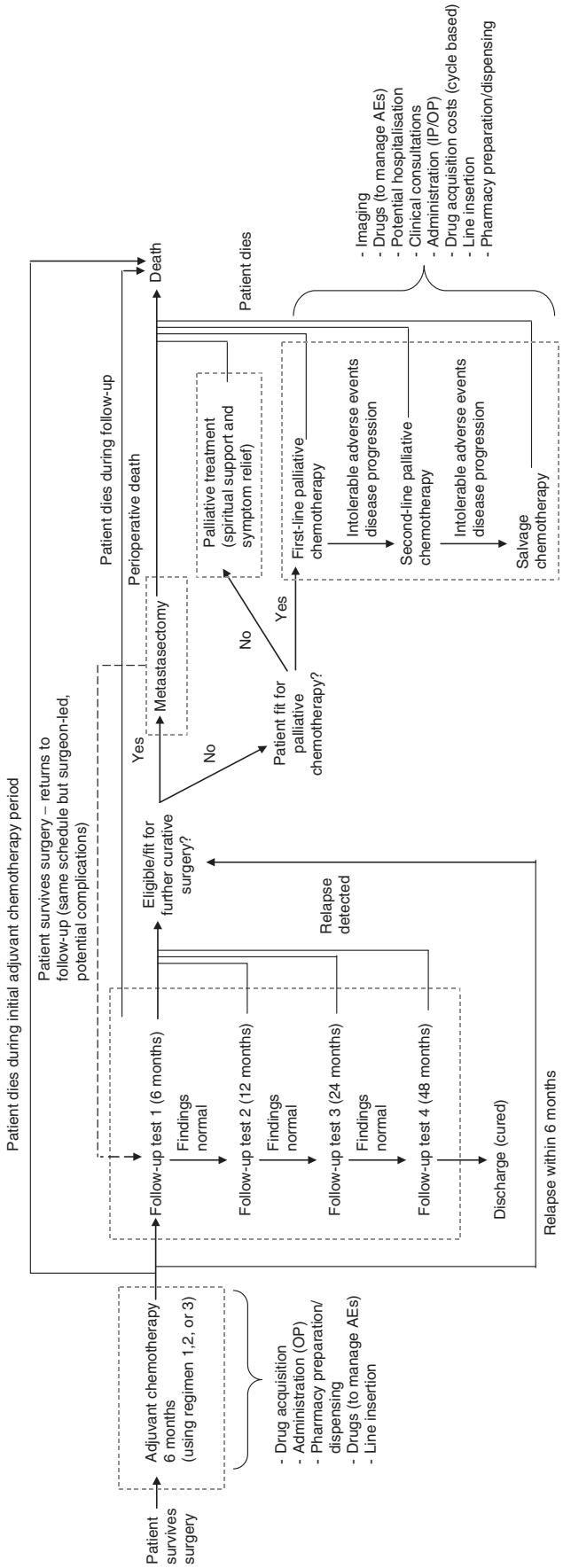


Figure 4 Illustrative service pathways model. IP, inpatient; OP, outpatient.

Box 3 Recommendations for practice – problem-oriented models

1. Develop the structure of the problem-oriented conceptual model using clinical guidelines and health professionals.
2. Use other health professionals not involved in model development to provide peer review and to check understanding of the conceptual models.
3. The precise graphical approach for presenting the conceptual models is important only in that they should be easily understood by health professionals and other decision stakeholders.
4. For the sake of clarity, it may be beneficial to present the model in both diagrammatic and textual forms using nontechnical, nonmathematical language.
5. Develop the problem-oriented models before developing the design-oriented model. The feasibility and acceptability of the design-oriented conceptual model should have no bearing on the adequacy of the problem-oriented conceptual models.

metastatic relapse? What are expected to be the key drivers of costs?

Box 3 presents recommendations for developing and reporting problem-oriented conceptual models.

Practical Considerations – Design-Oriented Conceptual Models

Figure 5 presents an example of a design-oriented conceptual model for the hypothetical decision problem (again, note that this is not intended to represent the 'ideal' model but merely illustrates the general approach). This type of model draws together the problem-oriented model views with the intention of providing a platform for considering and agreeing structural model development decisions. By following this general conceptual approach it should be possible to identify the anticipated evidence requirements for the model at an early stage in model development.

Anticipated evidence requirements to populate the proposed illustrative model are likely to include the following types of information:

- Time-to-event data to describe sojourn time/event rates and competing risks in States 1–4 for the current standard treatment.
- Relative effect estimates for the intervention(s) versus comparator (e.g., hazard ratios or independent hazards time-to-event data).
- Information relating to survival following cure.
- HRQoL utilities for cancer and cured states.
- Estimates of QALY losses or utility decrements and duration data for adverse events.
- Information concerning the probability that a relapsed patient undergoes active/palliative treatment.
- Survival and other time-to-event outcomes for relapsed patients.
- Resource use and costs associated with:
  - Chemotherapy (drug acquisition, administration, pharmacy/dispensing, drugs to manage adverse events, line insertion);
  - Resource use and unit costs for follow-up;
  - Supportive care following relapse; and
  - Active treatments following relapse.

It may be helpful to consider the following issues when developing design-oriented conceptual models.

Anticipated Evidence Requirements

- What clinical evidence is likely to be available through which to simulate the impact of the new intervention(s)? How should these parameters be defined and what alternatives are available? Should independent or proportional hazards be assumed?
- Are all relevant interventions and comparators compared within the same trial? If not, is it possible for outcomes from multiple trials to be synthesized? How will this be done?
- What evidence is required to characterize adverse events within the model? What choices are available?
- Beyond the baseline and comparative effectiveness data relating to the technology itself, what other outcomes data will be required to populate the downstream portions of the model (e.g., progression-free survival and overall survival by treatment type for relapsed patients, survival duration in cured patients)?
- Will any intermediate–final relationships be modeled? What external evidence is there to support such relationships? What are the uncertainties associated with this approach and how might these be reflected in the model?
- Which descriptions of HRQoL states are possible and how will these parameters be incorporated into the final model?
- Will all model parameters be directly informed by evidence or will calibration methods (e.g., Markov Chain Monte Carlo) be required? Which calibration methods will be used and why should these be considered optimal or appropriate?
- What premodel analysis will be required to populate the model? Which parameters are likely to require this?

Modeling Clinical Outcomes

- Which outcomes are needed by the decision-maker and how will they be estimated by the model?
- How/should trial evidence be extrapolated over time?
- If final outcomes are not reported within the trials, what evidence is available concerning the relationship between intermediate and final outcomes? How might this information be used to inform the analysis of available evidence?
- How will the impact(s) of treatment be simulated? How will this directly/indirectly influence costs and health outcomes? What alternative choices are available?

Modeling Approach

- Which methodological approach (e.g., state transition, patient-level simulation) is likely to be most appropriate? Why?
- Is the proposed modeling approach feasible given available resources for model development?

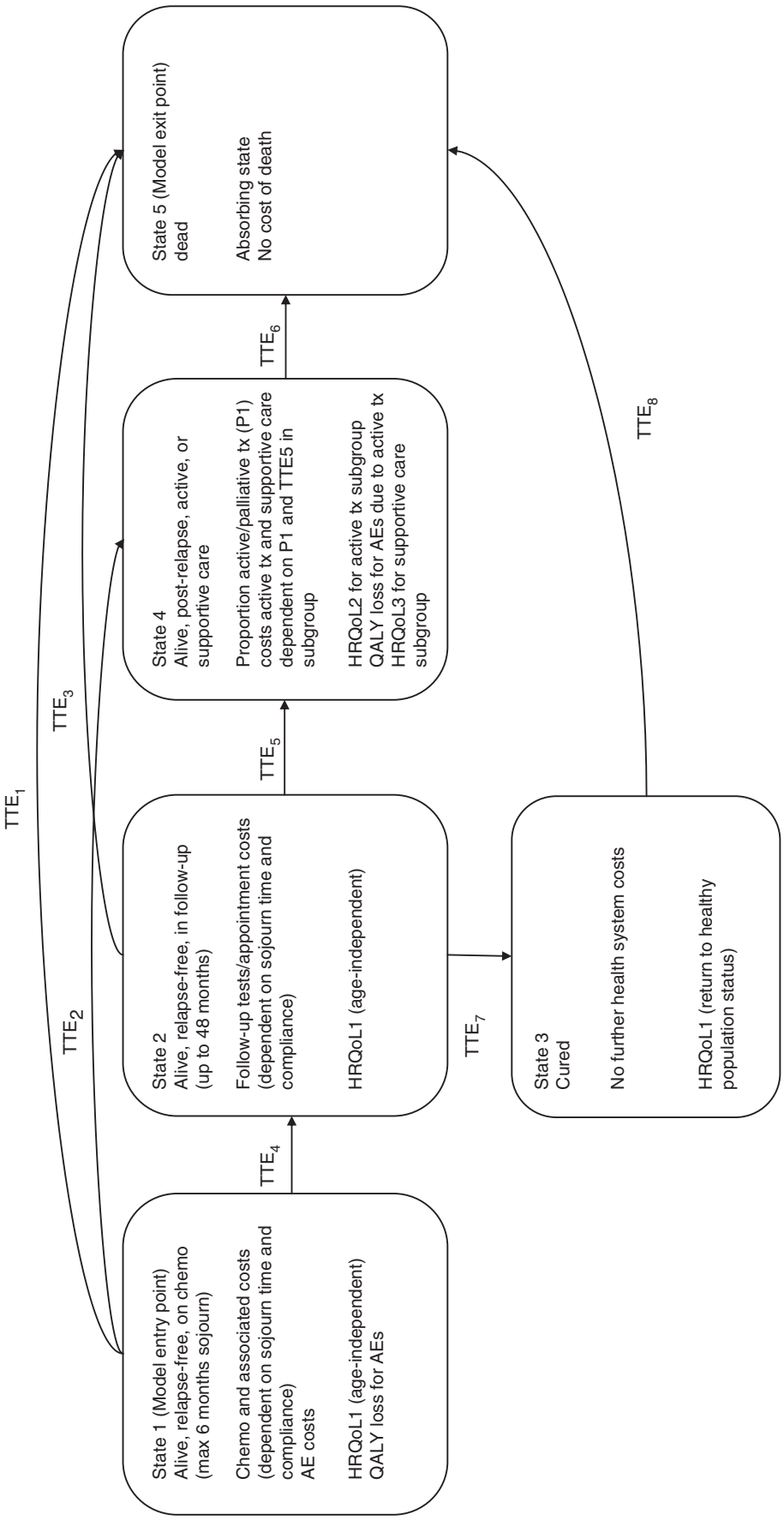


Figure 5 Illustrative design-oriented conceptual model. AE, adverse event; TTE, time to event.

Box 4 Recommendations for practice – design-oriented conceptual models

- The design-oriented conceptual model should be developed initially before the development of the final implementation model. It may, however, be revisited and modified within an iterative process during the development of the quantitative model.
- Model development involves making a large number of decisions and judgments. Not every decision or judgment made during model development will be important. The key decisions are likely to be those whereby the implemented model clearly deviates from the problem-oriented models (e.g., a part of the system is excluded) or whereby several alternative choices exist but none of which are clearly superior (i.e., structural uncertainties). These decisions should be clearly documented and reported.
- The sources of evidence used to inform model structure and the methods through which this information is elicited should be clearly reported.
- Where possible, alternative model development choices drawn out at this stage should be later tested using the quantitative model to assess their impact on the model results. This will not, however, always be possible or feasible.

- How does the approach influence the way in which certain parameters are defined? What alternatives are available (e.g., time-to-event rates or probabilities)?
- Does the proposed modeling approach influence the level of depth possible within certain parts of the model?

Adherence to a Health Economic Reference Case

- Will the proposed model meet the criteria of the reference case specific to the decision-making jurisdiction in which the model will be used? If not, why should the anticipated deviations be considered appropriate?

Simplifications and abstractions

- Have any relevant events, costs or outcomes been purposefully omitted from the proposed model structure? Why? For what reason(s) may these omissions be considered appropriate?

Table 1 Roles and concerns regarding the use of evidence to inform alternative model structures

Existing economic evaluations/ models	*Expert input (including clinicians and potentially patients/service users)*	*Clinical guidelines/ previous TA guidance/ local treatment protocols*	*Empirical clinical studies and reviews (e.g., RCTs, cohort studies)*
Principal role(s) in conceptual model development			
• To apply previously developed model structure to the current decision problem under consideration • To use existing economic analyses to highlight key evidence limitations • To identify possible options for model development decisions • To identify relevant treatment pathways	• To inform problem-oriented conceptual model development • To scrutinize the credibility of alternative model structures • To elucidate uncertainty regarding geographical variation	• To identify existing treatment/management pathways • To highlight gaps in the existing evidence base	• To identify available evidence to inform relationships between intermediate and final endpoints • To investigate what evidence is available
Issues and caveats associated with use			
• Existing models should not be relied on without considerable scrutiny. • Publication or other forms of dissemination of an existing model does not guarantee that the previous model was either appropriate or credible. • Advances in knowledge may render an existing model redundant • There may exist a gap between the decision problem that the model was developed to address and the current decision-problem under consideration	• Seek input from more than one health professional to capture the spectrum of clinical opinion • Use multiple experts located in different geographical locations • There exists a trade-off between seeking support from individuals with considerable expertise and standing (may not have much time but more experience/knowledge) and less experienced clinicians (may have more time to engage but lesser knowledge of evidence base). • Health professionals cannot be completely objectively detached from the system the model intends to represent • May be difficult to distinguish between conflict and geographical variations • Potential conflicts of interest • Potential ethical restrictions	• Current practice may have evolved since publication of guidance • Such evidence sources may not provide sufficient detail to inform the current decision problem • Local protocols may not reflect geographical variations between centers • Local protocols and guidelines may not be evidence-based • There may exist a gap between what should happen and what does happen in clinical practice	• Potential reliance on the availability of evidence rather than the structure of the problem • Differences between studies may suggest competing theories regarding (1) the nature of the disease process and (2) the relevance of particular events. This is not a problem as such but should be drawn out during conceptual model development • Treatments and comparators may reflect historical rather than current or best practice

Abbreviations: RCT, randomized controlled trial; TA, technology appraisal.

- Are there any parts of the disease or treatment pathways that have been excluded altogether? Why?
- What is the expected impact of such exclusion/simplification decisions? Why?
- What are the key structural simplifications? How does the design-oriented model structure differ from the problem-oriented conceptual models? Why should these deviations be considered appropriate or necessary? What is the expected direction and impact of these exclusions on the model results?

Box 4 presents recommendations for developing and reporting design-oriented conceptual models.

Evidence Sources to Inform Conceptual Models

A number of potential evidence sources may be useful for informing these types of conceptual model. Although the evidence requirements for any model will inevitably be broader than that for traditional systematic reviews of clinical effectiveness, the task of obtaining such evidence should remain a systematic, reproducible process of inquiry. Possible sources of evidence to inform conceptual models include: (1) clinical input; (2) existing systematic reviews; (3) clinical guidelines; (4) existing efficacy studies; (5) existing economic evaluations or models; and (6) routine monitoring sources. Table 1 sets out some pragmatic concerns which should be borne in mind when using these evidence sources to inform conceptual model development.

Acknowledgment

The general framework presented within this article has been adapted from a Technical Support Document funded by the National Institute for Health and Care Excellence to support their Technology Appraisal Program. Thanks to Alec Miners, Luke Vale, Rob Anderson, and Chris Hyde for their useful comments on the original draft of this work. The views expressed within this article are those of the author and do not necessarily reflect those of NICE.

See also: Adoption of New Technologies, Using Economic Evaluation. Economic Evaluation, Uncertainty in

References

Ackoff, R. L. (1979). The future of operational research is past. *Journal of the Operational Research Society* **30**, 93–104.

Brennan, A. and Akehurst, R. (2000). Modelling in health economic evaluation: What is its place? What is its value? *Pharmacoeconomics* **17**(5), 445–459.

Briggs, A., Claxton, K. and Sculpher, M. (2006). *Decision modelling for health economic evaluation*. New York: Oxford University Press.

Chilcott, J. B., Tappenden, P., Rawdin, A., et al. (2010). Avoiding and identifying errors in health technology assessment models. *Health Technology Assessment* **14**(25), i–135.

Cooper, N. J., Coyle, D., Abrams, K. R., Mugford, M. and Sutton, A. J. (2005). Use of evidence in decision models: An appraisal of health technology assessments in the UK to date. *Journal of Health Services Research and Policy* **10**(4), 245–250.

Eddy, D. M. (1985). *Technology assessment: The role of mathematical modelling. Assessing medical technology*. Washington DC: National Academy Press.

Lacy, L., Randolph, W., Harris, B., et al. (2001). Developing a consensus perspective on conceptual models for simulation systems. *Proceedings of the 2001 Spring Simulation Interoperability Workshop*.

Law, A. M. (1991). Simulation model's level of detail determines effectiveness. *Industrial Engineering* **23**(10), 16–18.

Robinson, S. (2008). Conceptual modelling for simulation Part I: Definition and requirements. *Journal of the Operational Research Society* **59**, 278–290.

Rosenhead, J. and Mingers, J. (2004). *Rational analysis for a problematic world revisited: Problem structuring methods for complexity, uncertainty and conflict* (2nd ed). England: Wiley.

Sargent, R. G. (2004). *Validation and verification of simulation models. Proceedings of the 2004 Winter Simulation Conference*. Available at: http://www.medicine.mcgill.ca/epidemiology/courses/EPIB654/Summer2010/Modeling/Validation%20paper%20modsim.pdf (accessed 06.08.13).

Production Functions for Medical Services

JP Cohen, University of Hartford, West Hartford, CT, USA

Glossary

Average Product Total output rate divided by the amount of a variable input used in its production.

Cobb–Douglas production function The Cobb–Douglas production function has the form $Y = AK^a L^b$, where Y is the output rate A, a and b are positive constants, A is a variable broadly representing 'technology', and K and L are capital and labour services respectively.

Data envelope analysis A linear programming technique that uses empirical evidence of the most efficient producers of outputs to locate an envelope that predicts the maximum outputs achievable with a variety of different inputs.

Leontief production function A Leontief production function has fixed input proportions, implying zero substitutability between the various inputs: There is no increase in one input rate that would compensate for a reduction in the rate of use of another (keeping output constant).

Marginal product The increase in output associated with a small increase in an input.

Production function A technical relationship between inputs and the maximum outputs or outcomes of any procedure or process. Also sometimes referred to as the 'technology matrix'. Thus it may relate to the maximum number of patients that can be treated in a hospital over a period of time to a variety of input flows like doctor- and nurse-hours, and beds.

q-Complementarity Two inputs in a production function are termed q-complements when an increase in the rate of one raises the marginal product of the other.

q-Substitutability Two inputs in a production function are termed q-substitutes when an increase in the rate of one lowers the marginal product of the other.

Stochastic frontier analysis An empirical method of estimating the maximum outputs obtainable from given resources and, hence, the degree to which actual operations fall short of the most efficient way of operating.

Translog production function Translog is an abbreviation of 'transcendental logarithmic', a form of production function having greater generality than the Cobb–Douglas form of the function.

Introduction: Distinction between Production Functions for Medical Care versus Production Functions for Health

Production function studies in health economics have taken three divergent approaches. Some of these studies focus on the production function for general and regress health (such as reduced mortality) against a variety of factors. Another strand examines the technological relationship between medical care and the inputs that are used to produce medical care. A third approach examines medical care production efficiency more specifically through stochastic frontier or data envelopment analysis techniques. This article describes all three approaches with a primary focus on production functions and on efficiency analysis.

The production of health approach involves a more general specification of the production process, including a variety of societal factors as inputs, such as consumption of medical care, technology, demographics, and personal health habits. Findings in this literature include a positive relationship between medical care and health; demographics (such as education and income) and health; and avoiding risky behaviors (such as smoking and other substance abuse) and health. One recent example of this approach is production function estimation for health in the Organization for Economic Co-operation and Development (OECD) countries, which postulates that life expectancy at the age of 65 years depends on health expenditures, medical technology, and lifestyle. Health expenditures significantly affect life expectancy at the age of 65 years in OECD countries.

In contrast, two other major strands of production function estimation have examined the technology and efficiency associated with production of medical care, which is the primary focus of this article. One approach examines the technological relationship between inputs (such as employment of different types of health care workers, physical capital, and possibly other inputs) and output or outputs, which can include client counts (admissions or discharges), relative value units, or others. More recently, interest has focused on the relationship between inputs and the quality of output, but this is an area deserving much greater attention. The outputs have centered around a variety of different health care services, ranging from hospitals, to physicians and specialty care treatment centers.

Hospital Production Functions

The literature on hospital production functions is quite extensive. In the general hospitals literature on production functions, researchers examine individual hospitals (or other medical entities such as practices) which maximize utility rather than minimize costs, where utility may be defined as a function of effort and leisure. Effort is defined as the number of discharges or admissions. The hospitals maximize their utility, given a budget, labor market conditions, and a

 doi:10.1016/B978-0-12-375678-7.01010-5

production function. The production function for each hospital describes the technological relationship between the capital and labor inputs, and the process by which capital and labor are translated into output.

A typical hospital production function can be estimated by least squares regression techniques, after adding an 'error' term.

It is noteworthy that 'output' can be represented by admissions, discharges, and/or relative value units; the stock of physical capital can be measured by beds, and/or value of equipment and structures; the supply of labor can include either full-time equivalents, number of total employment hours, or one of these measures for several separate labor categories such as physicians, nurses, etc. The labor variable may include one type of labor and focus on aggregate hours or full-time equivalents, whereas alternatively separate labor variables can be included for different types of hospital workers (i.e., physicians, nurses, clinicians, clerical workers, etc.). The latter can be advantageous in assessing the substitutability of different types of workers. Often a vector of client mix or client demographic variables that affect the position of the production function is included in the statistical estimation.

The derivative of output with respect to each input is denoted as the marginal product of the input; it describes how output changes when there is a small change in the amount of one input, while holding constant all other factors of production. Specifically, in the hospital context, the marginal product of physicians is the additional patients who can be treated when there is a slight increase in the number of physicians. This marginal product is required to be positive, and the value of output should be zero in the presence of zero inputs (i.e., if there is only one labor input, hospital full-time equivalents, then zero hospital full-time equivalents implies zero patients treated). If there are several labor inputs (i.e., physicians, nurses, clinicians, clerical workers, etc.), it is permissible that some (but not all) of these labor inputs equal zero. Also, the production function should increase at a decreasing rate – in other words, the marginal product decreases as more physicians (or nurses) are added.

Functional Forms

Before estimating the medical care production function with regression analysis, a functional form must be specified. There are several common functional form assumptions that have been used in the literature, including Cobb–Douglas, translog, and generalized Leontief. A Cobb–Douglas production function is perhaps the most straightforward because of its linear structure in logarithms. A convenient feature of the Cobb–Douglas is that the regression parameter estimates are also elasticities. In assessing the marginal product of labor, for instance, this would be the elasticity of output with respect to labor, times the output level divided by the employment level.

Although the Cobb–Douglas has some advantages due to computational simplicity, and it diminishes the potential for multicollinearity because of a lack of interaction terms, a disadvantage is that it does not allow the elasticities of substitution among different types of hospital workers (or among a particular type of hospital employee and number of beds, for instance) to be different from unity. In other words, a hospital production function can generate information on how the facilities are able to substitute capital for labor by examining the elasticity of substitution, but the Cobb–Douglas production function assumes this elasticity is constant and equal to one at all levels of input use. For most hospitals, this assumption is quite restrictive and unrealistic. So, some researchers have considered an alternative that is more flexible, known as the translog. The translog production function is a generalization of the Cobb–Douglas – in other words, it builds on the Cobb–Douglas by adding interaction terms (in logarithms) for all of the possible combinations of inputs.

One advantage of the translog, compared with the Cobb–Douglas production function, is that the translog allows for the possibility of elasticities of substitution between physicians and nurses to be different from unity, and these elasticities can vary across hospitals. This is a desirable feature of the translog because it provides valuable information that can be useful in policy recommendations. But a potential problem with the translog arises when there is a zero in one or more of the inputs for some hospitals. For instance, if the labor inputs in the model include physicians, nurses, clinicians, and clerical workers, and if some hospitals have no clinicians, then this will be problematic because the log of zero is undefined. As an alternative, researchers have considered a generalized Leontief production function. Typically, the generalized Leontief hospital production function models admissions and/or discharges as a function of total labor and capital inputs and other shift variables. Alternatively, several types of labor can be included along with one type of capital and several shift factors. The generalized Leontief allows for interaction of the square roots of each variable (i.e., every type of labor, capital, and other shift variables).

Interpretation of Production Function Estimates

Average and Marginal Products

There are a couple of possible scenarios that may be evident with the production function estimates. First, hiring additional workers may provide resources for clerical workers to perform more administrative tasks while allowing physicians and clinicians to specialize in treating patients, leading to higher average numbers of patients treated per employee. Alternatively, there may be a sufficiently large number of employees at clinics so that having additional workers might lead to office overcrowding, physicians and clerical workers getting in each others' way, and possibly more difficulty in getting reimbursements for treatment because of additional bureaucratic layers within these organizations. If the regression estimates of the production function support this second scenario, having fewer employees at clinics would be expected to result in higher average number of patients treated per employee. The marginal product of labor is the additional clients that can be treated when an additional worker is hired, while the average number of patients treated by workers is the average product of labor. If the marginal product of labor is greater (less) than the average product, then the average product of labor rises (falls) as more workers are hired.

Elasticities of Complementarity and Substitutability

Clinicians or physicians may work better when they have access to additional equipment, such as computers that may help with record keeping and billing. If so, the additional physical capital may allow the physicians to focus on the tasks they are trained to perform (that is, treating patients), while physical capital can be used for other tasks. This scenario is called q-substitutability; else, would physicians be able to treat greater numbers of patients if the hospitals were to hire additional support workers (such as nurses, clerical staff, or other employees)? This scenario is called q-complementarity. In other words, through the production function regressions one can address the question of whether workers and physical capital or two individual types of labor, are q-complements or q-substitutes.

The technological relationships between any two factor inputs can be assessed by examining the production function for q-complementarity. Here, capital and labor will be q-complements (q-substitutes) if an increase in capital increases (decreases) the marginal product of labor. More generally, the Hicks elasticity is defined as the product of output and the change in marginal product of labor resulting from a change in capital, all divided by the product of the marginal product of capital and the marginal product of labor. The Hicks elasticity measures the relative ease by which one factor can be substituted for another, while keeping admissions or discharges constant, so if the Hicks elasticity is positive (negative), this implies capital and labor are q-complements (q-substitutes). Nonphysician labor has been found to significantly impact physician productivity in the context of hospitals and a translog production function. Other translog production function analysis explains the relationships between physicians and other employees in health care settings. A common finding is that physicians and nonphysician employees are q-complements. In the context of hospital efficiency with a generalized Leontief production function, capital and physicians have been found to be q-complements; capital and technicians/aides have been found to be q-complements as well.

Stochastic Frontier Estimation and Data Envelopment Analysis

The third type of health production function studies focus on efficiency analysis. These studies, which also estimate production functions for medical care, are based on stochastic frontier models, and data envelopment analysis.

Stochastic frontier models (sometimes referred to as 'frontier models') are based on the assumption that the regression error term distribution does not follow a normal distribution. Because the production function models described in the section Hospital Production Functions are actually ideal production function models when there is no inefficiency, whenever the error term is nonzero there is inefficiency. In other words, technological inefficiency for any given hospital occurs when the error term is nonzero. Also, frontier models are assumed to have a typical production function functional form (such as Cobb–Douglas, translog, or generalized Leontief), as well as two error terms. One of these error terms is because of 'technical efficiency' and is assumed to be negative (because of the inability of a hospital to reach the frontier), and another error component is because of unobservables and other measurement difficulties. For any individual hospital, the frontier model can be written as a term that includes the production function as well as the inefficiency error term, and a separate error component from unobservables. Each hospital's deviation from the mean efficiency can be calculated, to assess how inefficiently individual hospitals are operating. The model can be estimated by assuming the errors follow half of a standard normal distribution, and then using maximum likelihood estimation techniques.

Multiproduct Adjustments

When there are multiple outputs, such as different services provided in each hospital, a distance function stochastic frontier approach is appropriate. A stochastic frontier distance function to assess efficiency in Australian hospitals with several outputs has found a range of efficiency scores between 0.7 and 0.75, whereas there were a handful of outliers with efficiency scores that were close to 1. A distance function approach is crucial in the context of hospitals because most hospitals produce many different outputs, including inpatient and outpatient services. But a more disaggregated approach to address the efficiency of an array of many different hospital services necessitates a distance function approach. This can be done by specifying a 'netput' (or net input) transformation function, where a function of vectors of inputs and outputs are set equal to zero.

An alternative that can address the multiple outputs issue is a hospital cost function approach. Duality implies that profit maximization, through input choice for the production function, yields the 'same' result as cost minimization, where inputs are chosen to minimize costs. The optimal cost function depends on input prices and outputs, which can include several outputs. In this respect, a cost function approach to efficiency can be estimated using least squares regression techniques, to test for the presence of economies of scale; or, a cost function approach to represent technology can also be implemented as part of a stochastic frontier estimation.

Data envelopment analysis is a nonparametric approach to measure hospital efficiency, after considering several outputs and inputs. It uses the approach of linear programming, to estimate a model. Here, an 'expansion factor' for each hospital must be chosen so that the 'expanded' output of each type at any hospital must be no greater than the weighted average of all other firms' output of that same type. At the same time, this hospital must use each input in such a manner that it is no less than the weighted average of all other firms' input usage of that same input type. In this approach, which essentially estimates a production possibilities frontier, hospitals' inefficiencies lead to deviations from the frontier. Some of the advantages of this approach are that no functional form needs to be imposed, and also it is possible that in some situations firms can produce outside the production possibilities frontier. In other words, if efficiency is defined as the inverse of the optimal value of this linear

programming model, then the hospital is efficient relative to the other hospitals in the sample if the inverse of the optimal value equals 1. If the inverse of the optimal value is less than 1 for a given hospital, then that hospital is inefficient relative to the other hospitals in the sample. This measure of efficiency can be obtained by solving this optimization problem for each hospital in the sample.

Specific Applications: Specialty Care

Although there are few known applications of production function estimation for substance abuse treatment, there are more mental health applications. One application evaluates how changes to mental health workforce levels, composition, and degree of labor substitutability, affect practice output (measured as relative value unit's) at US Department of Veterans Affairs mental health practices. This estimates the q-complementarity/q-substitutability of mental health workers, using a generalized Leontief production function, examining many labor types, including residents, and then estimates the marginal product for each labor type as well as the substitutability and complementarity of physicians and other mental health workers. Among 28 unique labor–capital pairs, 17 are q-complements and 11 are q-substitutes. Complementarity among several labor types provides evidence of a team approach to mental health service provision at these providers.

Another application studies the efficiency of nursing homes in the state of Connecticut, USA, using a data envelopment analysis approach and finds that among the 140 nursing homes in the state, nearly 100 are more efficient than the mean. The mean efficiency score for Connecticut nursing homes is approximately 0.90, whereas similar studies of nursing homes in other locations have found a range of efficiency scores approximately from 0.57 to 0.93.

A general practice dentistry application estimates a translog production function for a sample of approximately 29 000 dentists in the US, and finds that dentists tend to use dental assistance in practices where they are 'profitable' in terms of their marginal products. Also, dentists in the age range of the mid-40s tend to be the most productive among all age ranges in the sample.

Conclusion

Since the 1970s, production functions in efficiency and productivity studies have been pervasive in a wide variety of applications. Future research should focus on how to assess quality of care in addition to quantity. This would be helpful to practitioners who might rely on marginal product estimates in compensation decisions and to governments in pay-for-performance calculations.

See also: Cost Function Estimates

Further Reading

Baltagi, B., Moscone, F. and Tosetti, E. (2012). Medical technology and the production of health care. *Empirical Economics* **42**(2), 395–411.

Chattopadhyay, S. and Heffley, D. (1994). Are for-profit nursing homes more efficient? Data envelopment analysis with a case-mix constraint. *Eastern Economic Journal* **20**(2), 171–186.

Cohen, J. P. and Catherine, M. P. (2008). Agglomeration and cost economies for Washington state hospital services. *Regional Science and Urban Economics* **38**, 553–564.

Gaynor, M. and Gertler, P. (1995). Moral hazard and risk spreading in partnerships. *RAND Journal of Economics* **26**(4), 591–613.

Greene, W. (2012). *Econometric analysis* (7th ed.). Upper Saddle River, New Jersey: Pearson–Prentice Hall.

Hicks, J. R. (1970). Elasticities of substitution again: Substitutes and complements. *Oxford Economic Papers* **20**, 289–296.

Morrison Paul, C. J. (2002). Productive structure and efficiency of public hospitals. In Fox, K. (ed.) *Efficiency in the public sector*, ch. 8, pp. 219–248. Norwell, Massachusetts: Kluwer.

Reinhardt, U. (1972). A production function for physician services. *Review of Economics and Statistics* **54**, 55–66.

Scheffler, R. M. and Kushman, J. E. (1977). A production function for dental services: Estimation and economic implications. *Southern Economic Journal* **44**(1), 25–35.

Stefos, T., Burgess, J. F., Cohen, J. P., Lehner, L. and Moran, E. (2012). Dynamics of the mental health workforce: Investigating the composition of physicians and other health providers. *Health Care Management Science*. doi:10.1007/s10729-012-92031.

Thornton, J. and Eakin, B. K. (1997). The utility-maximizing self-employed physician. *Journal of Human Resources* **32**(1), 98–128.

Thurston, N. K. and Libby, A. M. (2002). A production function for physician services revisited. *Review of Economics and Statistics* **84**(1), 184–191.

Public Choice Analysis of Public Health Priority Setting ☆

K Hauck, Centre for Health Policy, Imperial College Business School, London, UK
PC Smith, Imperial College, London, UK

Glossary

Bureaucrat A bureaucrat is a nonelected government official, usually a member of an administrative policy-making group within an institution of the government. Street-level bureaucracy is the subset of members of a public agency or government institution who carry out and enforce the actions required by laws and public policies. It is accompanied by the idea that these individuals vary the extents to which they enforce the rules and laws assigned to them.

Cost-effectiveness analysis (CEA) An economic evaluation in which the costs and consequences of alternative interventions are expressed as cost per unit of health outcome. CEA is used to determine technical efficiency; i.e., comparison of costs and consequences of competing interventions for a given patient group within a given budget.

Decision-making The process of making a selective intellectual judgment when presented with several complex alternatives consisting of several variables, which defines a course of action or an idea.

Externalities An externality is a cost or benefit that results from an activity or transaction and that affects an otherwise uninvolved individual(s) who did not choose to incur that cost or benefit. For example, if majority of a community is vaccinated against an infectious disease, the resulting herd immunity benefits those who have not been vaccinated.

Interest group A voluntary association that seeks to publicly promote and create advantages for its cause.

Median voter The voter (or pair of voters) in the exact middle of a ranking of voters along some issue dimension, e.g. from the most left-wing to the most right-wing.

Priority setting The planning for equitable allocation, apportionment, or distribution of available health resources.

Public choice It is concerned with the study of political behavior. In political science, it is the subset of positive political theory that models voters, politicians, and bureaucrats as mainly self-interested. In particular, it studies such agents and their interactions in the social system either as such or under alternative constitutional rules.

Public good A good or service is both nonexcludable and nonrivalrous in which individuals cannot be effectively excluded from use and when used by an individual does not reduce the availability to others. Examples of public goods include fresh air, knowledge or national defense.

Public health The activities that society undertakes to assure conditions in which people can be healthy. These include organized community efforts to identify, prevent and counter threats to the health of the public.

Resource allocation The societal or individual decisions about the equitable distribution of available resources.

Many public health interventions are extremely good value for money. Advice from doctors to give up smoking, vaccinations against communicable diseases, or improved access to clean water in low-income countries are often relatively low cost interventions that produce substantial health gains. Where evidence on cost-effectiveness is available, many preventive and public health interventions fare very well when compared with conventional healthcare interventions. So why are not more public funds invested in public health? And why in some situations has it been so difficult to implement common-sense public health interventions such as sewage treatment, vaccinations, or taxes on cigarettes?

The entries 'Economics of public health: overview,' 'Infectious disease externalities,' 'Health behavior externalities' and 'Public health priority setting' discuss the welfare economic theories of public goods and externalities that are relevant to most public health interventions and that support the case on theoretical grounds for their public provision. But these theories also describe the unique characteristics of public health interventions that inevitably put them at a disadvantage when compared with other investments, in particular investments in health care. Although health care is directly consumed by the individual patient and can offer large, immediate, and certain health benefits, public health actions to mitigate externalities typically offer only small, delayed, and uncertain benefits to particular individuals – even though this may add up to large and certain benefits at the population level. In summary, as Glied (2008) put it: "Public health, economic theory says, is most useful and beneficial when nobody can observe cash savings because of the actions of public health; when public health activities don't even try to reduce taxes; when the potential benefits of public health actions are unclear; and when the potential beneficiaries of public health activities aren't even born yet!" These are the sort of reasons why robust evidence of causal effects from controlled trials and natural experiments is often not available, and so it is difficult to demonstrate convincingly the societal value of public health interventions using conventional economic evaluation tools.

The entry 'Economic evaluation of public health interventions: methodological challenges' discusses problems affecting the assessment of the cost-effectiveness analysis (CEA)

☆Some material in this article has been published in Goddard, M., Hauck, K. and Smith, P. (2006). Priority setting in health – A political economy perspective. *Health Economics, Policy and Law* **1**, 79–90.

 doi:10.1016/B978-0-12-375678-7.00411-9

of public health interventions, and how they could be overcome. But can the authors necessarily conclude that it is poor evidence on the value of public health interventions that made policy makers shy away from making these investments? Did the scientific community fail to produce the kind of evidence that would convince policy makers to do the right thing? Or are there more fundamental structural impediments to securing acceptance of the value of such investments?

The authors argue in this article that it is the realities of the political decision-making process that militate against political backing for public health investments, rather than the methodological shortcomings of CEA. Although economic evaluation offers a powerful rational approach to setting priorities, there may be alternative perspectives from which it is rational for decision makers to disregard the recommendations. The authors use three economic models of public choice – the interest group, majority voting, and bureaucratic decision-making models – to explain why it may be rational even for benevolent social welfare maximizing decision makers to diverge from the traditional economic evaluation approach and take into account public choice theory in order to assess the various political constraints on the decision options available to them, and the likely unintended consequences of alternative policies due to the predicted behavioral responses of key stakeholder groups, and possibly even responses of their social decision-making colleagues in other branches of government who unlike themselves may not behave like benevolent social welfare maximizer. The models help us move from the normative approach to priority setting, based on what should be done to maximize some concept of social welfare, into the realm of positive approaches that attempt to understand what happens in practice.

Models of public choice assume that the same behavioral model that can be used to explain decision making in ordinary markets can also be applied to decision making in the public sector. Public policy makers are not necessarily benevolent maximizers of social welfare, but may be motivated by their own self-interest. Firms seek to maximize profits, consumers seek to maximize utility, and policy makers seek to maximize political support or their own personal gain. The models further assume that, although policy errors are certainly possible, it is more informative to assume that the intended effects of a policy can be deduced from the observed effects, especially when such policies persist over time. In doing so, the authors do not of course argue that such behavior is desirable, or that it happens in all circumstances. Of course individuals have various motivations and the authors acknowledge that many decision makers will act with predominantly altruistic and welfare maximizing considerations in mind. The aim is merely to offer a framework for explaining apparently perverse actions on the part of decision makers.

Interest Groups

The interest group model is a powerful way of explaining why policy making diverges from the recommendations of cost-effectiveness analysis. It demonstrates that some groups of the population are more successful in promoting their interests than other groups and seeks to explain the impact this has on priority setting, resource allocation, redistribution of wealth, and even the survival of governments. The 'capture' theory describes interest groups as 'capturing' the regulatory power of the state to achieve a redistribution of wealth between different groups of the population in the form of transfers that may be cash or favors (Stigler, 1971). Generally, small groups with a clearly defined common objective – for example, the pharmaceutical industry – have lower costs in organizing themselves, securing cohesion, and effectively lobbying decision makers to their advantage, at the expense of the larger population whose interests may be more diffuse and experience higher costs of organizing. Some interest groups have privileged access to information that gives them a comparative advantage. The authors shall discuss examples where powerful minority groups with the interest, means, and opportunity to organize themselves have influenced political decisions to their advantage.

In low-income countries (LICs), the interest group model can explain why expenditures have often focused on healthcare services for richer areas or social groups at the expense of preventive and public health services for the poor, even where the latter offers greater cost-effectiveness. Poorer groups and populations based in rural areas may be less informed, less literate, and have an underdeveloped infrastructure for the dissemination of information compared to wealthier groups or those based in the urban areas where access to information resources is less limited. Groups in formal employment and with greater wealth also tend to be concentrated in urban areas, in particular in low-income Asian countries, see **Figure 1**.

Taxpayers represent an important interest group, especially in LICs with high levels of informal employment and a tax base that is highly dependent on a small minority of wealthy citizens. These citizens tend not to suffer to nearly the same extent from the communicable diseases and chronic conditions suffered by poorer citizens. In a democratic system, as the proportion of poor in the overall electorate is relatively large, most tax-financed healthcare expenditure would be devoted to illnesses of the poor in order to secure support of the majority of voters. However, such a policy choice would imply very large financial transfers, through the tax regime, from the rich to the poor. In short, the rich may have to make big tax contributions to public interventions that do not benefit them greatly. This may lead to resistance among the rich, tax evasion, increased collection costs, or even emigration.

An extreme form of tax evasion is illicit export of capital, especially prevalent when systems of governance are weak. It is estimated that the world's poorest countries lose USD$900 bn each year through illicit flows of capital. **Figure 2** shows the top 20 developing country exporters of illicit capital in declining order of average annual outflows, with China being the top exporter by far. Despite these high levels of tax evasion in some developing countries, however, economic theory of tax compliance predicts that global tax evasion should be much greater than it actually is, given the low extent of deterrence in most countries – conceptualized as the product of the probability of being detected and the size of the fine imposed. Empirical studies allude to an effect known as 'fiscal exchange': the more governments provide public services according to the preferences of taxpayers in exchange for a

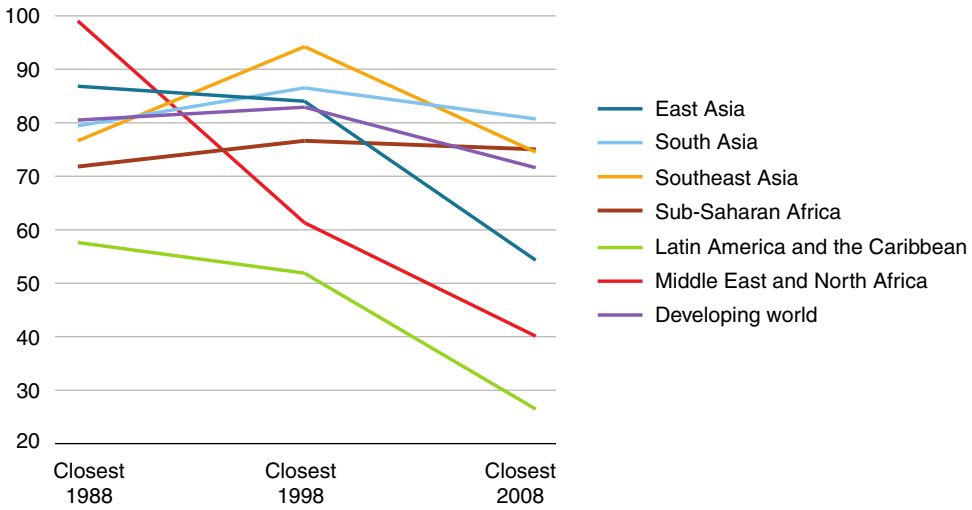


Figure 1 Rural share of total poverty (rural people as a percentage of those living on less than USD$1.25 per day). Reproduced from Figure 2 in International Fund for Agricultural Development (IFAD) (2011) *Rural Poverty Report 2011*, p. 47. Available at: http://www.ifad.org (accessed 11.06.13).

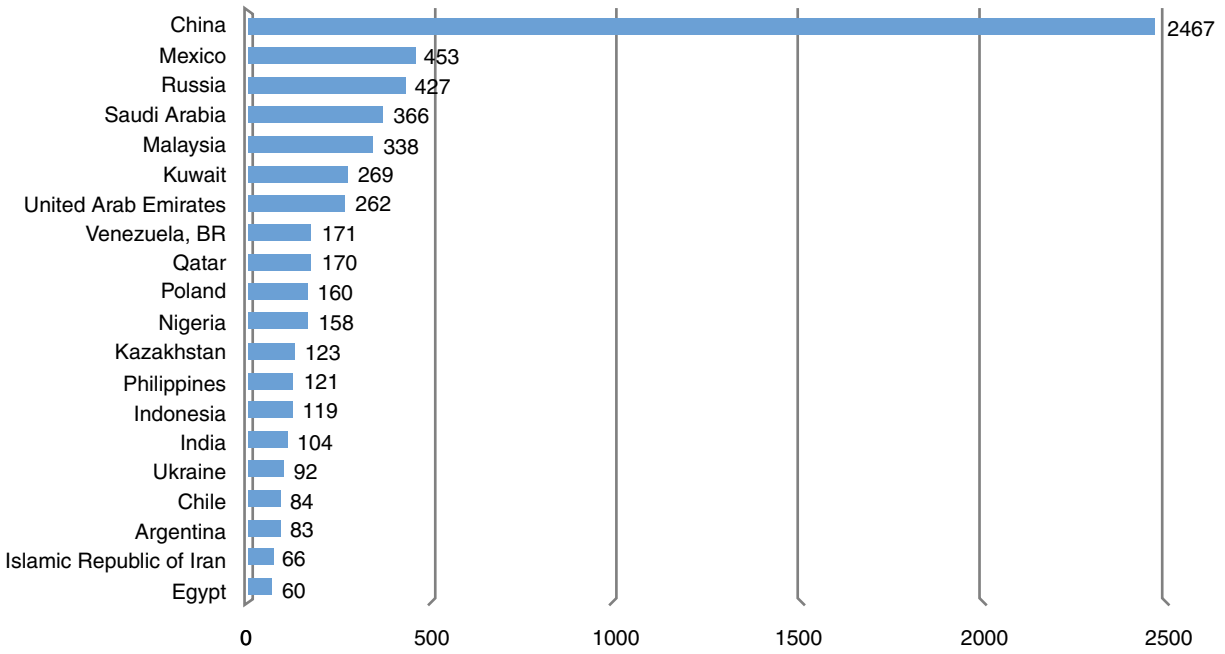


Figure 2 Top 20 countries' cumulative normalized illicit flows in nominal terms: 2000–2009 (billions of US dollars). Reproduced from Kar, D. and Freitas, S. (2011). *Illicit Financial Flows from Developing Countries Over the Decade Ending 2009*. Washington, DC: Global Financial Integrity. Available at: http://www.gfintegrity.org/ (accessed 11.06.13).

reasonable tax price, the more taxpayers comply with the tax laws. In the mid-nineteenth century Switzerland, a voluntary school tax in the canton of Glarus provided sufficient revenue to finance education services, whereas a voluntary welfare tax to redistribute income in the canton of Appenzell i. Rh. had to be quickly turned into coercive taxation. This concept of 'fiscal exchange' has important implications for healthcare priority setting. To limit tax resistance of this kind among the rich, the government may feel constrained to include some provision for the healthcare needs of the rich in the essential package of care in order to retain the viability of the tax base, even when the associated treatments do not qualify for inclusion on strict cost-effectiveness criteria.

Even in well-functioning democracies, powerful interest groups affect resource allocation decisions. Patient associations have successfully lobbied governments to fund drugs publicly, even if there is doubt about their cost-effectiveness, or even clinical efficacy and safety, for example, the breast cancer drug Herceptin in the English National Health Survey (NHS). Patients with long-term chronic illnesses such as

HIV/AIDS have a clear advantage in organizing themselves over the recipients of public health interventions. As public choice theory predicts, illnesses with comparably low prevalence are at an advantage, at least partly due to the fact that costs of organizing are lower. The preventive nature of many public health interventions implies that there is no clearly defined patient group that is lobbying in their favor, and public health has to rely on individuals or groups with altruistic motivation for support. Further, governments often prioritize sensitive political issues concerning highly visible aspects of healthcare services, at the expense of investments in public health. For example, some countries place a high priority on tackling waiting times for elective surgery, which affect a relatively small group of patients. This preoccupation could be interpreted as a response by politicians to the more media-friendly interests of waiting patients when compared with interventions aimed at the whole population, or large and difficult to delineate subgroups of individuals at risk, such as preventive screening or healthy lifestyle campaigns.

Providers of health services – the healthcare professions – form a crucial interest group in many countries, and governments are often wary of alienating doctors who are in a strong position to mobilize opposition to chosen priorities. For example, in the USA, health professionals and their associations are major lobbyists. In 2012, they have together spent more than USD$40 M to influence directly or indirectly decisions made by Congress and federal agencies, an amount that is nearly 6 times more than the tobacco industry. Although providers would certainly not want to be seen to actively work against the interest of patients, it is unlikely that investments in public health feature prominently among their priorities, especially if such investments are made at the expense of traditional healthcare interventions or even reduce demand for their services.

Doctors may have credible threats that can undermine the implementation of policy shifts, ranging from overt threats such as quitting the workforce to subtle noncooperation and adherence to traditional patterns of care. For example, it has been argued that the retreat from traditional models of managed care in the USA has in large part been due to pressure from physician and consumer groups in a backlash against government and insurers' attempts to cut costs through limiting access and rationing care. In some LICs, doctors may have a preference for high-technology medicine and be alienated by policy changes that seek to develop public health interventions and cost-effective care in community settings. Such alienation may have profoundly important consequences, for example, in the form of shifting employment from the public to the private sector or workforce emigration. Although migration is of course due to a multitude of reasons, it is noteworthy that in some LICs more than 50% of highly trained health workers leave for job opportunities in higher income countries.

The pharmaceutical industry is another powerful interest group that may favor healthcare interventions and drug treatment over public health investments. For example, The Council of the Europe Assembly quite openly voiced the suspicion that the pharmaceutical industry has influenced the World Health Organization's response to the H1N1 flu pandemic. The Council accused WHO of exaggerating the seriousness of the epidemic, which resulted in large amounts of public funds being spent on vaccines and antivirals that were never needed. But the interests of the pharmaceutical industry are not always in conflict with public health or preventive policies. Modern antiretroviral treatment (ART) may prevent secondary HIV infections because HIV patients receiving ART may have a significantly reduced risk of passing on the virus to sexual partner (treatment as prevention), and ART given to healthy persons may reduce their risk of acquiring the virus (preexposure prophylaxis). Using ARTs as prevention would open up a vast new market for pharmaceutical companies, potentially comprising all HIV negative individuals at risk of infections.

Other commercial companies that are driven by economic interests have formed powerful interest groups. **Table 1** summarizes the – strikingly similar – strategies adopted by the tobacco and food-producing industries, and historically water companies, to influence public health decision making to their advantage. Many of the strategies the tobacco industry adopted to frustrate public health actions came to light only when an extensive library of internal tobacco industry documents was released publicly as a result of the 1998 settlement agreement. For decades before, the tobacco industry successfully preempted efforts to limit advertising and sale of cigarettes, by publicly disputing evidence that smoking cigarettes damages health, for example, with the infamous 'A Frank Statement to Cigarette Smokers', half-hearted self-regulation such as the Cigarette Advertiser Code, and public messages and advertising that emphasized individual responsibility to deflect blame from the industry.

Table 1 Popular strategies by interest groups to influence political decisions to their advantage

- Emphasizing consumer's personal responsibility and condemning governmental interventions as autocratic
- Distorting scientific evidence with selective reviews or in the extreme, distribution of false evidence
- Adopting financial tactics including setting up, sponsoring, or otherwise developing associations with foundations or organizations that support the corporations agenda
- Exerting political influence and lobbying to support politicians favorable to the industry
- Adopting legal and regulatory tactics, for example, engaging in half-hearted self-regulation efforts or getting industry lobbyists appointed to governmental regulatory agencies
- Adopting legal and regulatory tactics, for example, engaging in half-hearted self-regulation efforts or getting industry lobbyists appointed to governmental regulatory agencies
- Advertising that connects the image of the corporation with worthwhile or popular causes

Source: Adapted from Brownell, K. D. and Warner, K. E. (2009). The perils of ignoring history: Big tobacco played dirty and millions died. How similar is big food? *Milbank Quarterly* **87**(1), 259–294.

"We are proud of the industry's record with respect to cigarette advertising generally and youth in particular. We submit that the record is one of unparalleled restraint and responsibility" – Horace Kornegay, Chairman, the Tobacco Institute. From the US Subcommittee on Health and the Environment's Report to the Committee on Energy and Commerce 'Advertising of tobacco products', House of Representatives, Serial No. 99–167, 18 Jul, 1 Aug 1986.

"Evidence is now available that the 14 to 18 year old group is an increasing segment of the smoking population. RJR must soon establish a successful new brand in this market if our position in the industry is to be maintained over the long term." – RJR's Secret planning assumptions and forecasts for the period 1976–1986 from the RJ Reynolds Tobacco Company Research Department, 18 Mar 1976.

Figure 3 Public statements and internal communications by the tobacco industry. Reproduced from http://legacy.library.ucsf.edu/legal.jsp

Public statements stood in stark contrast to internal communications (**Figure 3**).

Some public health experts now fear that history will repeat itself by comparing the tobacco industries' strategies with current efforts by the food-producing industry to deny the contribution of their products, in particular soft drinks and highly processed snack food, to the obesity epidemic (Brownell and Warner, 2009). They accuse the food industry of following distraction strategies, for example, playing up the importance of physical activity over nutrition, publishing biased reviews of scientific studies, or playing up a relatively harmless health impact (such as tooth decay) to divert attention from the serious one (such as obesity).

The food industry has made highly visible pledges to curtail children's food marketing, sell fewer unhealthy products in schools, and label foods in responsible ways, but has been criticized for their efforts (Sharma et al., 2010). For example, the School Beverage Guidelines developed by various charitable foundations in a partnership with the soft drink industry have been found to be implemented with far less restrictions in high schools, where much of the sugared-beverage intake occurs, than in elementary schools where little intake occurs (for more information visit the Clinton Foundation, http://www.clintonfoundation.org/main/our-work/by-initiative/alliance-for-a-healthier-generation/programs/industry-initiatives/school-beverage-agreement.html).

It is telling that while contributions to federal candidates and political committees from the tobacco industry fell drastically from USD$10.6 M in 1996 when legal battles were at their peak to USD$3.2 M in 2010; over the same period, contributions from the food processing and food retail industry have tripled from USD$10 M to just under USD$30.5 M in 2010. This is possibly attributable to increased congressional action on issues that affect the industry such as food safety, labeling regulations, soft drink taxes, and other anti-obesity initiatives.

There are, of course, differences between food and tobacco as substances. Unlike smoking, eating is necessary to maintain health and life. The associated public health messages are therefore more subtle, seeking to change the types and amounts of food eaten rather than promote abstinence. There is overwhelming evidence that smoking is addictive and damaging to health, whereas research on the addictive properties of food and its impact on human health is only now maturing. Smoking imposes harmful externalities on others through passive smoking, whereas in principle eating an unhealthy diet only harms the eater. The fight against tobacco coalesced around a single product made by a few companies, whereas the food industry is far more complex because it is fragmented, involving an immense array of products made by thousands of companies worldwide.

There are also historic examples of how commercial interests have shaped priority setting in public health. In fact, the public health movement that developed during the industrial revolution in the nineteenth century did so, at least partly, as a reaction to the commercial interests of private water companies. The water companies used their influence to dispute emerging evidence that the poor water quality they provided was responsible for the cholera epidemics and other illnesses that led to the appalling drop in life expectancy in English cities during the industrial revolution, using similar tactics to the tobacco industry 150 years later (Szreter, 2003). The development of germ theory and arrival of microscopic water analysis gave the public health movement the scientific backing to lobby for the construction of publicly funded and maintained sanitary infrastructure (see **Figure 4**).

Employers form an interest group that has traditionally supported public health interventions if they improve and preserve the productivity of their workforce. Historically, investments in public health in low-income countries were driven by the economic interests of colonial countries, and the need to guarantee the health of the workforce seconded to work there. For example, in Britain, in the 1890s, the Colonial Secretary Chamberlain was aware that the poor health of the native workers and the officials sent to serve in the Colonies was a threat to Britain's growing empire. Mortality among officials in some parts of the world, particularly the Gold Coast of West Africa, was soaring and to compensate, salaries were sometimes 100% higher than those of colleagues elsewhere. The economic significance of the control of tropical disease led to establishment of institutions and schools of tropical medicine, such as the London School of Hygiene and Tropical Medicine (UK) and the Pasteur Institute (France). Nowadays, some mining companies are providing free Anti retroviral therapy to their HIV positive employees, for example, Anglo-American. As HIV/AIDS predominantly affects working age adults, companies are possibly motivated by a combination of humanitarian interests and the commercial interest to preserve the human capital established in their workforce.

Even organizations seeking purely 'technocratic' solutions to priority setting by providing scientific evidence on the cost-effectiveness of interventions (e.g., health technology assessment agencies) may be influenced by interest groups in the selection of interventions chosen for assessment. Seeking public involvement in such activities risks capture by interest groups but can be seen as an attempt to increase political

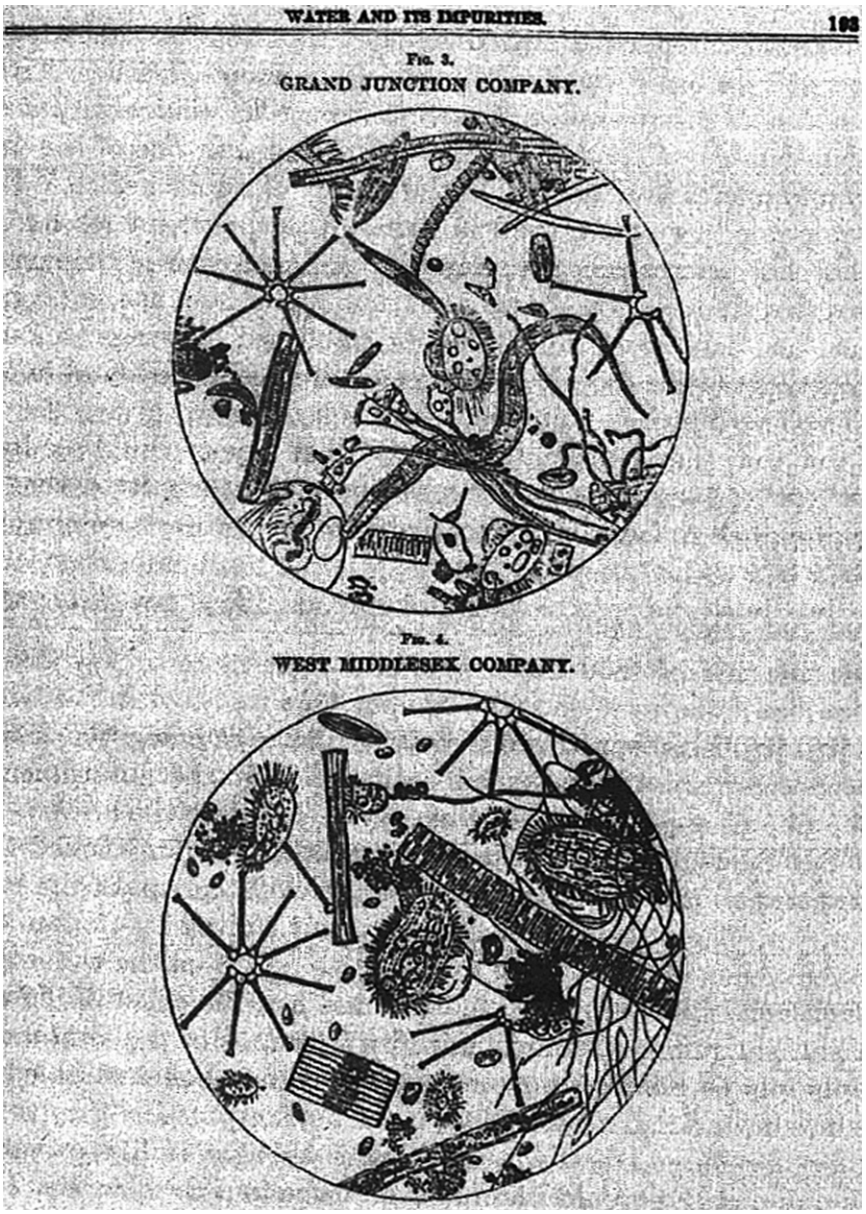


Figure 4 Water impurity in London's commercial supplies, illustrations from *The Lancet*, 1851. Caption below Figure 3: This engraving represents the chief animal and vegetable productions contained the water as supplied by the Grand Junction Company. Caption below Figure 4: The above engraving exhibits the principal animal and vegetable productions contained in the water supplied by the West Middlesex company. Drawn with the camera lucida and magnified 220 diameters. The microscope enabled water analysts to make precise drawings such as these depicting the organic contents of the drinking water supplied by London's increasingly notorious private companies. Some of these companies' defective systems were clearly implicated by pioneering epidemiological research into the major cholera epidemics of the period. Reproduced from Szreter, S. (2003). The population health approach in historical perspective. *American Journal of Public Health* **93**(3), 421–431.

support for decisions. The National Institute for Clinical Excellence in England and Wales has formalized public involvement in technology appraisal and developing interventional procedure guidance, but not without criticism.

Voting Models

Many public health interventions are targeted at conditions that predominately affect disadvantaged groups of the population. Indeed, some even have the primary objective of reducing inequalities in health. This in itself may imply a need to depart from pure cost-effectiveness criteria, which imply an objective of maximizing aggregate health outcomes. The median voter model may explain why such public health interventions often receive less political backing than others that benefit a wider spectrum of the population, or why investments into healthcare are favored over investments in public health. The model focuses on the politician as a maximizer of votes (Hotelling, 1929; Anderson, 1999). The 'median voter' theorem shows

that in a representative democracy, political parties tend to move toward the political position of the median voter in order to secure election.

The median voter model highlights the importance to the government of obtaining the support of crucial electoral constituencies. In public health, this may explain why policy makers seek to direct resources toward key population groups at the expense of others, notwithstanding the apparently reasonable claims of the latter on resources from an efficiency or equity perspective. For example, median voters are likely to perceive that they or their family benefit from screening services for common conditions such as cancers. Therefore, the provision of such services is likely to receive widespread support, even if evidence of cost-effectiveness is weak, or indeed they might do more harm than good, as has been suggested for routine mammography. However, policies directed at poor lifestyle choices (e.g., smoking, alcoholism, and risky sexual behavior) may receive less popular support because the median voter does not perceive any personal or family need for such services. Even if the latter services are very cost effective, politicians seeking reelection may find it difficult to attach high priority to them. Similarly, many common healthcare interventions, such as treatment for acute myocardial infarction, hip and knee replacements, or cataract removals, are likely to be demanded by the median voter at a certain point in life. Following that line of argument, the median voter model cannot explain the success of HIV/AIDS interest groups in lobbying the governments, and it is interesting to note that for HIV/AIDS it stands in disagreement with the interest group model.

More generally, economic models of voting in health services have received little attention in the literature (Tuohy and Glied, 2011). In industrial democracies, an ageing population suggests that older people are becoming an increasingly important electoral force. Although an ageing population is itself partly the result of effective public health interventions, perversely, the preoccupation of older people is likely to be with curative rather than preventative interventions, compared with young people, reinforcing the tendency for politicians to favor health services over public health.

Bureaucratic Decision Making

The behavior of interest groups and voters can be understood only in terms of the institutional context in which they occur (Tuohy and Glied, 2011). Tullock's (1965) and Niskanen's (1971) institutional theories focus on the interests of 'bureaucrats' in maximizing their influence and the effect of their behavior on the level and nature of government output. Here the concept of the bureaucrat is interpreted broadly to embrace all public sector actors with significant influence over the allocation of resources. The essence of this approach is the belief that such bureaucrats receive power and remuneration in proportion to the size of their enterprise, with the implication that bloated, and inefficient public services emerge if there is a lack of effective control on the growth of government. Under the bureaucratic model, government agencies will seek to implement policies that maximize the size of their own enterprises and to undermine activities that are outside their direct control. They are able to do so because they have an informational advantage over their political counterparts. 'Bureaucrats' may therefore influence the pattern of healthcare expenditures in ways that do not accord with efficiency and equity considerations. If this model applies, it would also suggest substantial inertia in spending, making it difficult for politicians to change entrenched patterns of services.

It is difficult to find hard evidence, but the tendency of bureaucrats to maximize their own budgets and sphere of influence at the expense of others can be observed across many government sectors. For example, bureaucrats in health ministries often find it difficult to persuade bureaucrats in other ministries, such as education, to adopt policies designed to improve health, because of the reluctance of each sector to relinquish control. Public health, perhaps more than other government activities, requires collaboration across sectors and cross-departmental actions for which responsibilities cannot be clearly delineated.

Multiple levels of governance add further complexities that affect variations in spending on public health, although the direction and magnitude of effects is likely to depend on specific funding arrangements for such policies (Tuohy and Glied, 2011). A system under which subnational governments make policy decisions, but a significant share of the associated costs is covered by the national government is likely to lead to higher investments in public health than a system under which national governments provide a fixed payment to subnational governments, which then bear the full marginal costs of the interventions. In a decentralized system of governance different levels of government are likely to free ride on interventions with public goods characteristics that are provided by other levels. More than 160 years ago implementation of basic sanitation measures was frustrated by tensions between central government and councilors of Britain's large cities (**Figure 5**).

In the public sector services, it has also been argued that 'street-level bureaucracy' plays a powerful role in the way in which policy is implemented. The considerable degree of discretion accorded to healthcare workers ('street-level bureaucrats') in determining the nature, amount, and quality of services provided by their agencies has a powerful impact on the rationing of resources, and the factors governing their decisions might not be based on cost-effectiveness or similar principles.

Conclusion

The authors have considered three economic perspectives on public choice that help to explain why it has often proven difficult to obtain political backing for apparently commonsense public health interventions. The health and societal implications of many public health interventions can never be assessed in their entirety in advance of implementation. Even if they could, however, public health advocates have often found themselves in conflict with powerful interests groups, and politicians or bureaucrats who pursue their own objectives. This introduces an important and complex set of constraints into the priority-setting process, implying that available funds might be spent in particular areas or on

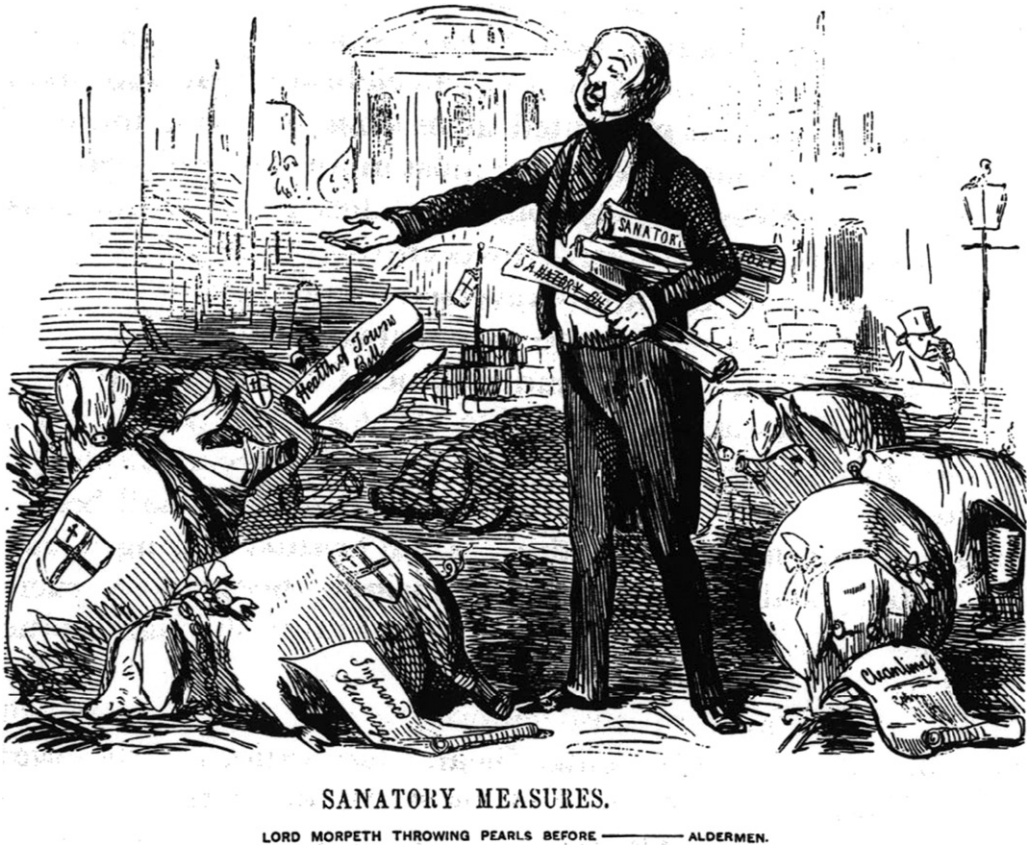


Figure 5 Cartoon about implementation of Chadwick's Public Health Act in British cities, Punch, June 1848. A cartoon of Lord Morpeth, the central government's representative, promoting the bill for Chadwick's Public Health Act. The legislation is depicted as 'sanatory' pearls being thrown in vain by the enlightened national statesman to the unappreciative 'swine,' the councilors of the nation's cities. Reproduced from Szreter, S. (2003). The population health approach in historical perspective. *American Journal of Public Health* **93**(3), 421–431.

specific programs determined by informational advantages or power structures within society. In general, such constraints will result in departures from conventional criteria such as cost-effectiveness rules.

Is it possible to counter those powerful influences? History has plenty of success stories. From the second half of the nineteenth century, Britain's major cities embarked on a reform of the municipal social health amenities and social services that resulted in significant improvements in health. Historians believe that a class-bridging coalition between grass root organizations of the growing urban population, a new generation of civic leaders with social conscience that originated from the well-off urban elite, and a strong cadre of public service professionals, notably Medical Officers of Health, secured these reforms. Similar developments happened in other European countries and the USA.

There are modern day examples of spirited initiatives by governments, government departments, or community organizations that instigated radical improvements in public health. For example, the outbreak of the plague in the Indian city of Surat in 1994 led to a decisive reorganization and introduction of stringent performance management of the civic waste department, named 'transformation from AC to DC' – making bureaucrats leave their 'air conditioned' offices to the 'daily chore' of direct supervision of waste management on site. The participatory budgeting model introduced in 1989 in the Brazilian city of Porto Alegre was an innovative reform program that successfully overcame severe inequality in living standards among city residents. Part of the program was introduction of participatory budgeting; community members now decide how to allocate part of the public budget. The recent decisive ruling of the Australian government on plain packaging laws for cigarettes is celebrated as a great success in the fight against cigarette smoking. For decades, the tobacco industry has successfully prevented such laws, to protect the value of their brands' image, which they used more or less openly to link the brand with popular causes or films – such as the placement of Phillip Morris brands in Superman I and II, see Figure 6.

The authors have merely scratched the surface of this underresearched area of study. There is great scope for a much better understanding of decision-making behavior – both in low- and high-income countries – and a range of hypotheses can be tested by examining the priority-setting process itself and the resulting patterns of healthcare and public health expenditure. They would not challenge the desirability of seeking to maximize the health gains of the health system through the use of CEA. However, health economists have not yet taken advantage of the full range of economic approaches at their disposal. Only by securing a better understanding of

Figure 6 Product placement of cigarette brand by Phillip Morris in the youth films Superman I & II. Reproduced from http://www.fitmedia.org/whatsanembeddedad.html

the decision-making process can the impact of CEA be enhanced. They would argue that this can be achieved by augmenting the understanding of the political context of priority setting, using a variety of well-established models of political economy.

See also: Advertising as a Determinant of Health in the USA. Economic Evaluation of Public Health Interventions: Methodological Challenges. Infectious Disease Externalities. Pay for Prevention. Priority Setting in Public Health. Public Health in Resource Poor Settings. Public Health: Overview. Public Health Profession. Smoking, Economics of

References

Anderson, G. M. (1999). Electoral limits. In Racheter, D. P. and Wagner, R. E. (eds.) *Limiting Leviathan*. Cheltenham, UK and Northampton, MA: Edward Elgar.

Brownell, K. D. and Warner, K. E. (2009). The perils of ignoring history: Big tobacco played dirty and millions died. How similar is big food? *Milbank Quarterly* **87**(1), 259–294.

Glied, S. (2008). Public health and economics: Externalities, rivalries, excludability, and politics. In Colgrove, J., Markowitz, G. and Rosner, D. (eds.) *The contested boundaries of American Public Health*, pp. 15–31. New Brunswick: Rutgers University Press.

Hotelling, H. (1929). Stability in competition. *Economic Journal* **39**(153), 41–57.

Kar, D. and Freitas, S. (2011). *Illicit financial flows from developing countries over the decade ending 2009. G. F. Integrity*. Washington, DC: Global Financial Integrity. Available at: http://www.gfintegrity.org/ (accessed 11.06.13).

Niskanen, W. A. (1971). *Bureaucracy and representative government*. Chicago: Aldine-Atherton.

Sharma, L. L., Teret, S. P. and Brownell, K. D. (2010). The food industry and self-regulation: Standards to promote success and to avoid public health failures. *American Journal of Public Health* **100**(2), 240–246.

Stigler, G. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science* **3**, 3–18.

Szreter, S. (2003). The population health approach in historical perspective. *American Journal of Public Health* **93**(3), 421–431.

Tullock, G. (1965). *The politics of bureaucracy*. Washington, DC.

Tuohy, C. H. and Glied, S. (2011). The political economy of health care. In Glied, S. and Smith, P. (eds.) *The Oxford handbook of health economics*, pp. 58–77. Oxford: Oxford University Press.

Further Reading

Mueller, D. (2003). *Public choice III*. Cambridge: Cambridge University Press.

Relevant Websites

http://www.youtube.com/watchv=XT35E7PM6Zo
A Frank Statement to Cigarette Smokers.

http://www.ameribev.org/nutrition–science/school-beverage-guidelines/news-release/
American Beverage Association.

http://www.angloamerican.com/media/releases/2008pr/2008-12-01/
Anglo-American.

http://regnet.anu.edu.au/sites/default/files/CTSI-WorkingPaper76-full.pdf)
Australian National University.

http://news.bbc.co.uk/1/hi/business/2180930.stm
BBC News.

http://news.bbc.co.uk/1/hi/health/5063352.stm
BBC News.

http://www.bbc.co.uk/news/business-19264245
BBC News.

http://tobaccodocuments.org/youth/AmToMUL19640000.Co.html
Cigarette Advertiser Code.

http://www.clintonfoundation.org/main/our-work/by-initiative/alliance-for-a-healthier-generation/programs/industry-initiatives/school-beverage-agreement.html
Clinton Foundation.

http://summaries.cochrane.org/CD001877/screening-for-breast-cancer-with-mammography
Cochrane Library.

http://www.fitmedia.org/whatsanembeddedad.html
FIT Media Coalition on Embedded Advertising.

http://iffdec2011.gfintegrity.org
Global Financial Integrity.

http://www.hptn.org
HIV Prevention Trials.

http://www.dh.gov.uk/en/Publicationsandstatistics/Legislation/DH_083348
House of Commons Health Select Committee on NICE.

http://ideas.repec.org/p/zur/iewwpx/287.html
IDEAS Federal Reserve Bank of St. Louis.

http://legacy.library.ucsf.edu/
Legacy Tobacco Documents Library.

http://timeline.lshtm.ac.uk/
London School of Hygiene and Tropical Medicine.

http://www.mcdonalds.co.uk/ukhome/nutrition-ingredients/nutrients/sugar.htm
McDonalds.

http://www.aidsmap.com/The-road-to-PrEP-trials-regulation-and-roll-out/page/2403757/
Nam aidsmap.

http://www.nytimes.com/2010/02/07/business/economy/07view.html
New York Times.

http://www.opensecrets.org/industries/indus.phpInd=H
OpenSecrets.

http://www.opensecrets.org/lobby/indusclient.phpid=A09&year=2009
OpenSecrets Websites on the Food Industry.

http://www.opensecrets.org/industries/totals.phpcycle=2012&ind=A02
OpenSecrets Websites on the Tobacco Industry.

http://jnci.oxfordjournals.org/content/82/9/730.extract
Oxford journals.

http://en.wikipedia.org/wiki/Participatory_budgeting
Participatory Budgeting.

http://www.pharmatimes.com/Article/10-06-25/Council_of_Europe_Assembly_slams_WHO_pharma_over_pandemic.aspx
Pharmatimes.
http://archive.org/details/tobacco_zsc72i00
Public Messages and Advertising.
http://www.ifad.org/rpr2011/
The International Fund for Agricultural Development.
http://www.nice.org.uk/getinvolved/patientandpublicinvolvement/patient_and_public_involvement.jsp
The National Institute for Clinical Excellence.
http://articles.timesofindia.indiatimes.com/2003-07-27/lucknow/27201050_1_plague-city-surat-public-money
Times of India.
http://www.who.int/mediacentre/factsheets/fs301/en/index.html
World Health Organisation.

Public Health in Resource Poor Settings

A Mills, London School of Hygiene and Tropical Medicine, London, UK

Glossary

Allocative and technical efficiency A resource allocation is efficient if it is not possible to reallocate resources so as to increase one person's utility (or health, or output) without decreasing another person's utility (or health, or output). In health economics the entity maximized is generally assumed to be utility, health, or welfare. Technical efficiency is a part of cost-effectiveness: not using more resources than are necessary to produce a given set of outcomes. There may be many allocations of resources that meet this condition. The least costly of them is the cost-effective allocation.

Cost-effectiveness analysis A method of comparing the opportunity costs of various alternative health or social care interventions having the same benefit or in terms of a common unit of output, outcome, or other measure of accomplishment.

Disability-adjusted life years (DALYs) A measure of the burden of disability-causing disease and injury. Age–specific expected life-years are adjusted for expected loss of healthy life during those years, yielding measures of states of health or, when two streams of DALYs are compared, potential health gain or loss by changing from one health care or social intervention to another.

Global burden of disease The DALYs lost through illness and premature death in all countries of the world.

Gross domestic product or gross national product (GDP or GNP) The total expenditure by residents and foreigners on domestically produced goods and services in a year is GDP. It is the main indicator used to measure the size or output of an economy. GNP is GDP plus income earned abroad by residents less income earned in the economy by foreigners, i.e., GDP plus net property income from abroad.

National health accounts A record of the resource flows in a country's health system by the main elements of health care financing: resource mobilization and allocation, pooling and insurance, purchasing of care, and the distribution of benefits. National health expenditures include in principle expenditures on activities 'whose primary purpose is to restore, improve, and maintain health regardless of the type of the institution or entity providing or paying for the health activity.'

Overseas development assistance The Organisation for Economic Cooperation and Development's term for foreign aid. Also known as overseas development assistance.

Randomized controlled trial A scientific experiment conducted to test the effect of an intervention by randomly assigning participants to a treatment and control group. Differences between the treatment and control group participants are interpreted as the causal effect of the intervention.

Real terms The use of the adjective 'real' in economics is to distinguish monetary changes in value that are merely inflationary from others corresponding to changes in the flow of goods and services. Index numbers are used to deflate nominal values and thereby generate real values.

Introduction

This article addresses the distinctive challenges of planning, financing, implementing, and evaluating public health policies in low- and middle-income countries. By public health is meant the 'science and art of promoting and protecting health and well-being, preventing ill-health and prolonging life through the organized efforts of society' (http://www.fph.org.uk/what_is_public_health).

A key feature of low- and middle-income countries is a health discourse that diverges in some key ways from that in high-income countries. In particular, 'public health' in high-income countries is a recognized and accepted subject area that is generally formally planned and structured, always with agencies with specific public health roles and sometimes even with a government minister for public health. Elsewhere, and especially in low-income countries, it tends to be assessed and planned in ways that are more fragmented, and less coherent, for reasons explored in this article.

What is Distinctive about Low- and Middle-Income Countries?

Burden of Disease

The global burden of disease (GBD) across the world has been studied for some time. In 2012, a comprehensive analysis was published for 2010, including comparison with the original 1990 analysis. It uses the metric of the Disability-Adjusted Life Year (DALY), which combines years of life lost due to premature death with years of healthy life lost due to illness and disability, for specific diseases and conditions. The regional breakdown is done by 21 geographical regions rather than broad country income grouping, although the identification of subregions does largely distinguish lower and higher income groupings within broader regions (e.g., Europe is disaggregated to Western, Central, and Eastern Europe).

Table 1 shows total DALYs and DALYs per thousand population by the 21 regions and the change between 1990

 doi:10.1016/B978-0-12-375678-7.00403-X

Table 1 Disability-adjusted life years for 291 causes by region for 1990 and 2010, and the percentage change from 1990 to 2010

	Total DALYs (thousands)			*DALYs (per thousand)*		
	1990	*2010*	*%Δ*	*1990*	*2010*	*%Δ*
High-income Asia Pacific	38 934 (35 997–42 301)	42 486 (38 842–46 586)	9.1	231 (213–250)	239 (218–262)	3.5
Western Europe	115 151 (106 794–124 174)	113 364 (103 991–123 930)	−1.6	302 (280–326)	272 (250–298)	−9.8
Australasia	5382 (4966–5853)	6101 (5538–6733)	13.3	264 (243–287)	235 (214–260)	−10.7
High-income North America	79 582 (74 150–85 639)	91 073 (84 342–98 239)	14.4	287 (267–309)	268 (248–289)	−6.6
Central Europe	43 442 (40 918–46 341)	38 978 (36 355–41 960)	−10.3	355 (335–379)	327 (305–353)	−7.9
Southern Latin America	14 626 (13 755–15 688)	15 562 (14 458–16 917)	6.4	299 (281–321)	259 (240–281)	−13.5
Eastern Europe	88 654 (84 173–93 891)	93 104 (88 367–98 267)	5.0	400 (380–424)	449 (427–474)	12.3
East Asia	379 565 (355 627–405 991)	332 437 (306 978–358 541)	−12.4	319 (299–342)	238 (220–257	−25.5
Tropical Latin America	53 824 (50 633–57 102)	56 781 (52 636–61 338)	5.5	349 (329–371)	281 (261–304)	−19.5
Central Latin America	53 375 (50 672–56 555)	57 706 (53 753–61 997)	8.1	321 (305–340)	250 (233–268)	−22.2
Southeast Asia	192 296 (180 655–204 699)	188 512 (175 435–202 574)	−2.0	418 (392–444)	309 (287–332)	−26.0
Central Asia	30 298 (28 853–31 889)	28 539 (26 801–30 395)	−5.8	441 (420–464)	356 (334–379)	−19.3
Andean Latin America	16 513 (15 558–17 564)	14 164 (13 074–15 304)	−14.2	427 (402–454)	265 (244–286)	−38.0
North Africa and Middle East	123 183 (116 867–130 540)	124 617 (115 374–134 555)	1.2	408 (387–432)	279 (259–302)	−31.5
Caribbean	15 582 (14 757–16 483)	26 698 (21 182–39 812)	71.3	437 (414–462)	614 (487–915)	40.6
South Asia	747 529 (705 906–798 664)	680 859 (633 905–727 982)	−8.9	665 (628–710)	422 (393–452)	−36.6
Oceania	4015 (3527–4618)	4779 (3907–5825)	19.0	621 (546–714)	481 (393–586)	−22.6
Southern sub-Saharan Africa	23 794 (22 429–25 299)	44 027 (41 666–46 474)	85.0	452 (426–481)	625 (591–659)	38.1
Eastern sub-Saharan Africa	207 130 (196 459–219 636)	204 526 (193 904–216 317)	−1.3	994 (943–1054)	575 (546–609)	−42.1
Central sub-Saharan Africa	60 702 (56 022–66 082)	77 391 (71 187–83 385)	27.5	1132 (1044–1232)	802 (738–864)	−29.1
Western sub-Saharan Africa	209 023 (196 925–221 795)	248 683 (232 208–266 906)	19.0	1040 (980–1103)	740 (691–794)	−28.8

Data are DALYs (95% uncertainty intervals) or % change. DALY, disability-adjusted life years; %Δ, percentage change.

Source: Reprinted from Murray, C. J. L., Vos, T., Lozano, R., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2197–2223.

and 2010. A full discussion of the data is in the source article. Key points to note are that in 2010 as in 1990, poorer regions have a much greater burden of disease per 1000 population, with the regions with the greatest burden per 1000 being in sub-Saharan Africa. Note, though that these regions have seen some of the sharpest reductions in burden per 1000 (although a marked increase in total burden due largely to population growth).

Figure 1 shows how the total burden of disease is broken down by broad cause in these 21 regions, comparing again 1990 and 2010. In 2010, high-income countries (high-income Asia-Pacific, Western Europe, Australasia, high-income North America, and central Europe) had only 7% of DALYs due to communicable, maternal, neonatal, and nutritional disorders. Cancer and cardiovascular disease accounted for 36% of DALYs. In contrast, in east, west, and central sub-Saharan Africa, the former disease groups accounted for 67–71% of DALYS. Nonetheless, comparing 1990 and 2010, the great reduction in common infectious diseases in poorer regions is evident (the reduction in the light yellow bars), although the rise in HIV is very visible (in dark yellow) in southern and eastern sub-Saharan Africa. Even in the poorer regions there is a significant burden due to noncommunicable diseases and injuries; hence, the common label of the 'double burden' in low- and middle-income countries – the unfinished agenda of communicable diseases and the new burden of noncommunicable diseases and injuries. Analysis by age groups and region can be explored at the GBD website of the Institute for Health Metrics and Evaluation (http://www.healthmetricsandevaluation.org/). For example, the analysis by region of DALYS in children under 5 shows vividly that disease in children is virtually absent in high-income regions and is concentrated in low-income regions (see http://www.healthmetricsandevaluation.org/gbd/visualizations/gbd-2010-patterns-broad-cause-groupunit=pc&sex=B&metric=daly&stackBy=region&year=5).

From a public health point of view, it is important to understand the risk factors that underlie the burden of disease. Figure 2 shows the relative importance for the 21 regions of the risk factors defined in the GBD study. For central, eastern, and western sub-Saharan Africa, for example, the three leading risk factors are underweight, household air pollution from solid fuels, and suboptimal breastfeeding. In South Asia, household air pollution ranks top, although the changing pattern of disease is shown by tobacco smoking and high blood pressure ranking 2 and 3. In high-income regions, top ranking factors are high blood pressure, smoking, alcohol use, and high body mass index. This demonstrates visibly how the pattern of risk factors changes as countries grow richer.

Economic Structure

Economic structures not only help explain the patterns of burden of disease but also have a strong influence on the capacity of countries to respond to public health needs.

Key risk factors have their origins in the economic structure of low-income countries. The overall low level of income, and high proportions of the population living in absolute poverty, help explain under nutrition and the unsafe living environments. 'Structural' factors, such as low levels of education and lack of access to employment and informal employment, also

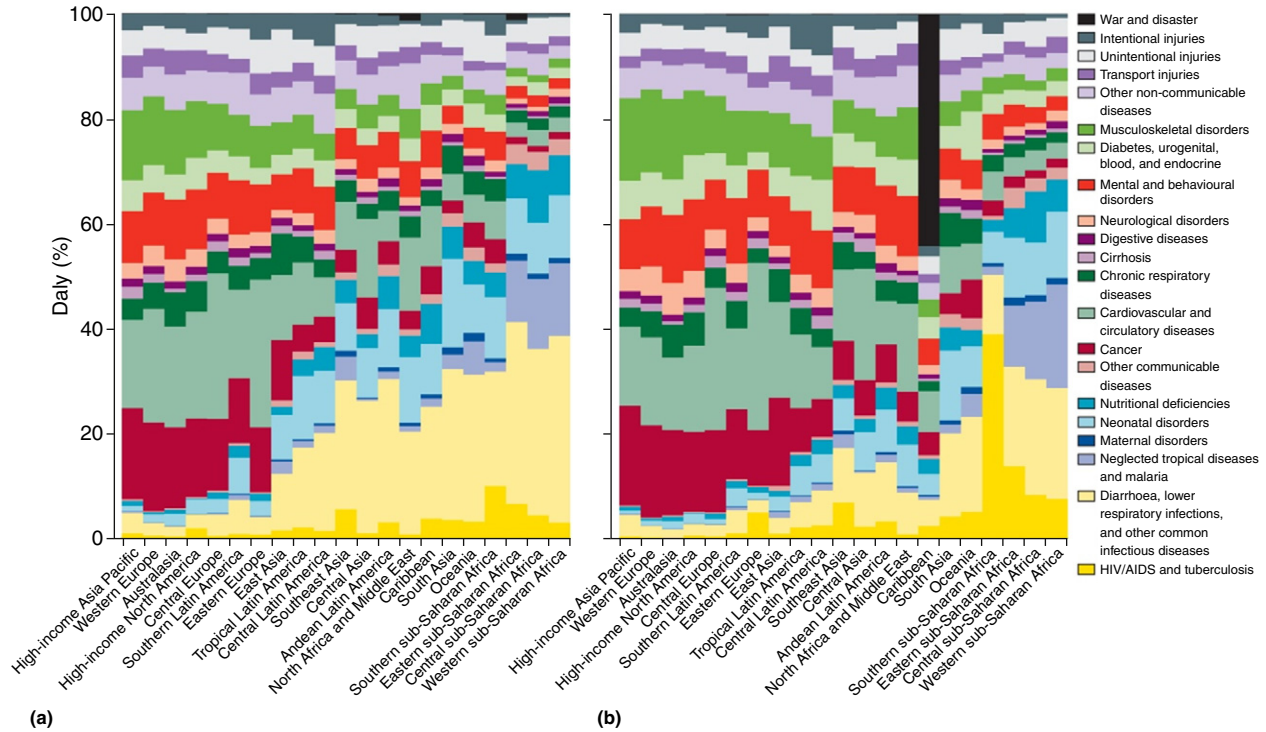


Figure 1 Percentage of disability-adjusted life years by 21 main cause groupings and region, 1990 (a) and 2010 (b). Reprinted from Murray C. J. L., Vos, T., Lozano, R., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2197–2223. An interactive version of this figure is available online at: http://healthmetricsandevaluation.org/gbd/visualizations/regional

help explain the prevalence of unsafe sex which encourages the spread of sexually transmitted diseases (note that lack of data led to unsafe sex being omitted from the GBD risk factor list in Figure 2). In middle-income countries, growing affluence and changing behaviors associated with higher incomes, global spread of information and the influence of the multinational industry on diet and smoking help explain the very different pattern of risk factors.

The poverty of a country severely affects its ability to finance an adequate response to public health needs. Figure 3 (taken from the WHO National Health Accounts website) maps levels of per capita government expenditure on health across the world in 2010, and Table 2 shows shares of total, government and private health expenditure by country income group in 2009.

Many low-income governments in Africa and Asia spend less than $34 per capita on all health care. Mean total health expenditure in 2009 in low-income countries was approximately 5% of GDP and less than half of this was channeled through government. In 2001, the WHO Commission on Macroeconomics and Health estimated that a set of 49 priority health interventions, largely for personal preventive and primary care, would cost 6% of the GNP of low-income countries. Hence, current levels of government expenditure are inadequate to fund these priority interventions and in addition are funding many other interventions, which might be seen as lower priority, notably high level hospital care, which is costly but benefits relatively few people. Although private out-of-pocket spending is often at least as high as government spending, the bulk of this is spent on over-the-counter drug purchases rather than the high-priority preventive activities needed to improve the health of the public.

Poverty has another consequence, that of low salaries for heath workers in the public sector. Professionals have always been internationally mobile, and improved communications have made it easier for professionals to move to high-income countries where salaries are higher. Moreover, in a context where government physicians often supplement their low government salaries by private clinical work, public health practice may be less attractive because it offers less scope to increase reputation through clinical practice in the public sector, and postings to the more rural and remote areas limit the opportunity for profitable private practice.

Low levels of government expenditure reflect another consequence of the economic structure, the difficulty of raising adequate taxation to finance public services. Low-income economies have a large share of their populations in the informal sector, making it difficult to levy direct taxes on a significant proportion of the total population.

Political and Social Institutions

A corollary of underdevelopment is that the political and social institutions needed to underpin an effective government tend to be weak. These include the institutions of democracy

Ranking legend
1–5 6–10 11–15
16–20 21–25 26–30
31–35 36–40 >40

Risk factor	Global	High-income Asia Pacific	Western Europe	Australasia	High-income Noth America	Central Europe	Southern Latin America	Eastern Europe	East Asia	Tropical Latin America	Central Latin America	Southeast Asia	Central Asia	Andean Latin America	North Africa and Middle East	Caribbean	South Asia	Oceania	Southern sub-Saharan Africa	Eastern sub-Saharan Africa	Central sub-Saharan Africa	Western sub-Saharan Africa
High blood pressure	1	1	2	3	4	1	2	2	1	2	4	1	1	2	1	1	3	6	2	6	5	6
Tobacco smoking, including second-hand smoke	2	2	1	2	1	3	3	3	2	4	5	2	3	5	3	3	2	3	5	7	12	10
Alcohol use	3	3	4	4	3	2	4	1	6	1	1	6	2	1	11	5	8	5	1	5	6	5
Household air pollution from solid fuels	3	42	..	..	..	14	23	20	5	18	11	3	12	7	13	9	1	4	7	2	2	2
Diet low in fruits	5	5	7	7	7	5	6	5	3	6	7	4	5	10	6	8	5	9	8	8	11	13
High body-mass index	6	8	3	1	2	4	1	4	9	3	2	9	4	3	2	2	17	2	3	14	18	15
High fasting plasma glucose	7	7	6	6	5	7	5	10	8	5	3	5	7	6	4	4	7	1	6	10	13	11
Childhood underweight	8	39	38	37	39	38	38	38	38	32	23	13	25	18	21	14	4	8	9	1	1	1
Ambient particulate matter pollution	9	9	11	26	14	12	24	14	4	27	19	11	10	24	7	19	6	32	25	16	14	7
Physical inactivity and low physical activity	10	4	5	5	6	6	7	7	10	8	6	8	9	8	5	7	11	7	11	15	15	16
Diet high in sodium	11	6	10	11	11	9	11	9	7	9	13	7	6	13	8	15	14	16	13	21	17	18
Diet low in nuts and seeds	12	11	9	8	8	8	8	8	12	10	8	15	8	12	9	10	13	13	16	22	16	21
Iron deficiency	13	20	32	21	35	22	17	21	19	14	12	12	17	4	12	6	9	11	10	4	4	4
Suboptimal breastfeeding	14	..	..	..	..	..	27	..	24	22	15	14	16	9	15	13	10	10	4	3	3	3
High total cholesterol	15	12	8	9	9	10	9	6	13	11	10	16	14	16	10	16	20	14	19	28	27	30
Diet low in whole grains	16	10	16	16	17	11	12	11	11	12	14	26	13	17	14	12	15	15	32	24	19	24
Diet low in vegetables	17	14	13	12	13	13	10	12	15	16	20	10	11	14	18	11	16	12	15	23	23	20
Diet low in seafood omega-3 fatty acids	18	17	15	13	16	16	14	13	17	17	18	19	15	23	16	17	18	20	23	27	25	25
Drug use	19	13	14	10	10	20	13	17	18	13	16	18	20	11	19	18	22	19	12	19	24	22
Occupational risk factors for injuries	20	24	24	20	25	26	16	25	20	19	22	23	21	21	23	31	12	22	22	20	22	17
Occupational low back pain	21	15	17	15	23	18	20	24	14	15	24	17	24	22	20	26	23	17	24	17	21	19
Diet high in processed meat	22	22	12	14	12	15	18	15	29	7	9	27	19	15	27	24	25	27	28	31	28	28
Intimate partner violence	23	18	22	23	22	25	21	22	21	23	26	22	27	19	25	23	21	25	14	18	20	23
Diet low in fibre	24	16	18	18	18	19	15	16	16	25	28	20	18	28	22	22	23	21	33	36	34	36
Unimproved sanitation	25	38	39	39	41	42	40	40	40	40	38	30	37	31	32	28	19	18	18	9	8	9
Lead exposure	26	23	21	19	24	17	19	23	22	20	25	24	23	20	26	21	24	30	20	25	26	26
Diet low in polyunsaturated fatty acids	27	19	19	17	20	21	22	18	26	24	27	21	22	29	24	25	32	23	30	33	30	29
Diet high in trans fatty acids	28	29	23	24	15	23	28	19	28	21	21	33	26	27	17	38	28	34	35	37	36	37
Vitamin A deficiency	29	40	40	38	40	41	41	42	43	41	37	32	34	34	37	33	30	31	17	11	7	8
Occupational particulate matter, gases, and fumes	30	34	33	32	28	32	33	31	23	29	32	28	29	33	31	34	26	33	29	29	29	31
Zinc deficiency	31	37	37	36	37	39	39	39	39	39	29	29	28	25	35	27	31	28	21	13	10	14
Diet high in sugar-sweetened beverages	32	28	31	31	19	33	29	27	37	26	17	25	32	30	28	20	27	26	26	32	32	34
Childhood sexual abuse	33	26	25	22	21	30	25	26	30	28	30	37	30	26	29	30	29	35	31	26	31	27
Unimproved water source	34	41	41	40	38	40	42	41	42	42	40	31	36	35	30	29	34	24	27	12	9	12
Low bone mineral density	35	21	20	25	26	24	30	28	25	30	33	35	35	36	34	32	36	37	38	35	37	33
Occupational noise	36	33	35	34	36	35	35	35	33	33	31	34	31	32	36	35	37	36	34	30	33	32
Occupational carcinogens	37	31	26	29	31	34	32	34	27	38	35	38	33	40	38	40	39	41	37	41	42	42
Diet low in calcium	38	25	28	27	29	27	29	30	31	34	39	39	39	39	40	37	40	39	39	38	39	38
Ambient ozone pollution	39	36	36	41	33	36	43	37	34	43	43	43	43	43	43	43	35	43	43	42	38	41
Residential radon	40	32	27	35	27	28	36	33	32	36	41	41	38	42	41	42	41	42	42	43	43	43
Diet low in milk	41	27	29	30	30	29	34	32	35	37	42	40	41	41	42	39	42	40	41	39	41	39
Occupational asthmagens	42	35	34	33	34	37	37	36	41	35	36	36	42	37	39	36	38	29	36	34	35	35
Diet high in red meat	43	30	30	28	32	31	31	29	36	31	34	42	40	38	33	41	43	38	40	40	40	40

Figure 2 Risk factors ranked by attributable burden of disease, 2010. Regions are ordered by mean life expectancy. No data = attributable disability-adjusted life-years were not quantified. Reprinted from Lim, S. S., Vos, T., Flaxman, A. D., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study. *The Lan*cet **380**, 2224–2260.

and representation, of civil society, and of professional groupings. One consequence for public health is that the ability to regulate health-damaging products may be far more limited than in the rich world. For example, low- and middle-income countries have come under pressure from the global tobacco industry, as well as self-interested governments who benefit indirectly from overseas tobacco sales through domestic tax revenues and employment, not to impose restrictions on the sale of tobacco or on cigarette advertising. In the late 1980s, the US used bilateral trade relations to exert pressure on countries such as Thailand and South Korea to open up their domestic markets to cigarette imports. Regulation of pharmaceuticals is

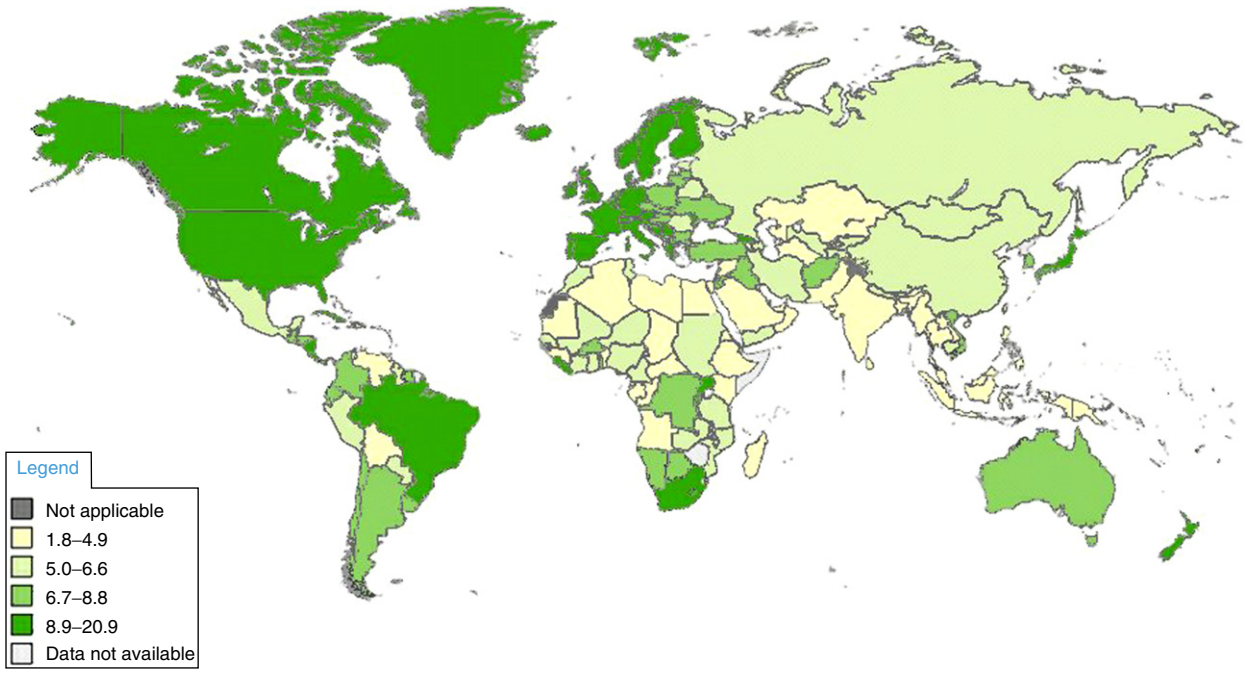


Figure 3 Per capita government expenditure on health at average exchange rate (US$), 2010. Reproduced from WHO National Health Accounts website http://apps.who.int/nha/database/StandardReportList.aspx

Table 2 Government and private health expenditure 2009

	THE as % of GDP	*GHE as % of THE*	*Private expenditure on health as % of THE*	*GHE as % of total government expenditure*
Low-income	4.9	38.9	61.1	8.5
Lower-middle income	4.4	39.0	61.0	5.5
Upper-middle income	6.1	54.8	45.1	10.5
High-income	12.5	61.9	38.0	17.1
Global	9.4	59.1	40.8	14.3

THE, total health expenditure; GDP, gross domestic product; GHE, Government health expenditure.
Source: Reproduced from WHO (2012). World Health Statistics Tables. *Global Health Expenditure Atlas*. Geneva: WHO. Available at: http://www.who.int/nha/atlasfinal.pdf (accessed 03.10.13).

also often very weak, leading to widespread and uncontrolled usage of antibiotics and antimalarials, for example, which can encourage the development of resistance. Artemisinin, a relatively new antimalarial, is a case in point. WHO has urged countries to adopt regulatory measures to stop the marketing of oral artemisinin-based monotherapies and to promote access to artemisinin-based combination therapies: the effect of the combination is to protect the individual drug elements from the development of parasite resistance. But artemisinin monotherapy remains widely available in many countries due to both inability to enforce regulation and the lack of separation of government and commercial interests.

Management Capacity in the Public Sector

Low- and middle-income countries often struggle to translate policies into action. For example, their plans usually do give priority to diseases and conditions that give rise to the greatest burden of disease, yet in practice coverage rates of interventions such as immunization, vitamin A, and emergency obstetric care, remain well below universal. Although to some extent this reflects a lack of prioritization in resource allocation decisions, it also reflects the limitations of management capacity. This encompasses not only trained managers but also management and information systems more broadly and ability to implement programs effectively across countries with poor communications infrastructure and often remote and hard to reach populations.

Influence of Agencies External to the Country

On average, external financing makes up 27% of the total health expenditure of low-income countries (but far less in middle-income countries). In the last decade or so, official development assistance for health has increased substantially in real terms: it amounted to $7331 m in 2003, and $17 856 m in 2010 (constant 2010 US$). Along with this increase has come a proliferation of agencies at the global level concerned to

Table 3 International initiatives relevant to public health in low- and middle-income countries (1978–2011)

1978:	International Conference on Primary Health Care, Alma Ata
1979:	'Health for all' goal
	Eradication of smallpox
1982:	Child Survival Revolution
1985:	Universal Program on Immunization
1986:	Global Program on AIDS
1987:	Bamako Initiative
	Safe Motherhood Initiative
1988:	Global Polio Eradication Initiative
1991:	World Health Assembly resolution to eliminate leprosy
1992:	Integrated Management of Childhood Illness
1995:	International Commission for the Certification of Dracunculiasis (guinea-worm disease) Eradication
	DOTS strategy for TB control
1996:	UNAIDS
1997:	Lymphatic filariasis elimination
1998:	Roll Back Malaria Partnership
1999:	GAVI Alliance
2000:	Stop TB Partnership
	Commission on Macroeconomics and Health
	Millennium Declaration and Millennium Development Goals
	Bill and Melinda Gates Foundation
2001:	Measles Initiative
2002:	Global Fund to Fight AIDS, Tuberculosis and Malaria
	Clinton Health Access Foundation
2003:	U.S. President's Emergency Plan for AIDS Relief (PEPFAR)
	3×5 Initiative (3 m people on antiretroviral treatment by the end of 2005)
	World Bank Report 'Investing in Health'
	Framework Convention on Tobacco Control
2005:	Commission on Social Determinants of Health
	Partnership for Maternal, Newborn and Child Health
2006:	UNITAID
2007:	H8 (WHO, UNICEF, UNFPA, UNAIDS, Global Fund, GAVI Alliance, Gates Foundation, World Bank)
	International Health Partnership
2008:	High Level Taskforce on Innovative International Financing for Health Systems
2010:	Global Strategy for Women's and Children's Health
2011:	WHO Commission on Information and Accountability for Women's and Children's Health

address specific health issues. Table 3 lists the key ones, starting with the landmark event of the Alma Ata conference in 1978, which launched the primary care movement. Since then, the tendency has been for more and more of the international initiatives to be focused on specific diseases – known as 'vertical programs' when they are organized and delivered in silos, separate from general health services. Most prominent have been HIV, TB, and malaria, but other diseases also feature, notably vaccine-preventable diseases such as polio and measles and so-called 'neglected' tropical diseases such as filariasis. In partial response to the attention given to these diseases, and to the resources they have attracted, other responses have developed particularly to seek remedy of the relative lack of emphasis on maternal, neonatal, and child health.

These international initiatives have very important consequences for low-income countries. Although they may provide the means to address effectively the needs of certain population groups, such as AIDs patients, they fragment the policy and planning environment at country level. Funds often flow vertically within a particular disease area, from international donors directly down to regional and local levels in a disease-specific chain of command and rarely in response to a systematic and comprehensive cross-disease planning process at country level. Low-income countries are on the receiving end of a multiplicity of funder requirements and procedures, which stretch already limited management capacity. Funders may bypass core government structures, such as central procurement agencies or indeed the Ministry of Health itself, rather than strengthening them.

Despite the growing importance of noncommunicable diseases, there are very few initiatives to address these, the framework convention on tobacco control being the only one in Table 3. An advocacy process, which resulted in a UN high level meeting on noncommunicable diseases in 2011, is generally considered to have produced very little in the way of concrete action.

Addressing Public Health in Low- and Middle-Income Countries

Concerns about the effective functioning of the health systems of low- and middle-income countries have led over many years to processes of what in the 1980s and 1990s were called 'health sector reform' and now are more commonly referred to as health systems 'development' or 'strengthening'. Such efforts can be criticized for focusing primarily on the health care system, rather than on broader public health. For example, from the 1980s onward debates have focused on the choice of policies for health financing. Advocates in international agencies, such as the World Bank, and the UK, French and German bilateral aid agencies, have taken up various positions over time on the appropriate mix of user fees for government health services, community-based health insurance schemes, social health insurance, and increased use of private financing arrangements including private voluntary health insurance. However, notably absent from the policy proposals until very recently have been statements on the critical importance of devoting increased government tax revenue to health. Given that public health above all concerns services that have the characteristics of public goods and externalities, it is apparent that planning for public health will not feature prominently when the main focus of debate is on how to stimulate sources of financing beyond government revenue.

The recent attention given by WHO and other international agencies to the 'social determinants of health' can be seen as an attempt to redress the balance in favor of a comprehensive and multisector approach to health. Social determinants are economic and social conditions, and their distribution among the population, that influence individual and group differences in health status. A WHO-appointed Commission on the Social Determinants of Health, set up in 2005 and reporting in 2008, examined the relationship between a variety of economic, political, legal, social and physical factors, and health. Although the report has been followed by a world conference in Rio and a WHO Executive Board resolution, it is not clear whether any low- or middle-income country has really embedded the thinking of social

Table 4 Key messages of the World Development Report 1993

Foster an environment that enables households to improve health
- Pursue economic growth policies that will benefit the poor (including, where necessary, adjustment policies that preserve cost-effective health expenditures)
- Expand investment in schooling, particularly for girls
- Promote the rights and status of women through political and economic empowerment and legal protection against abuse

Improve government investments in health
- Reduce government expenditures on tertiary facilities, specialist training, and interventions that provide little health gain for the money spent
- Finance and implement a package of public health interventions to deal with the substantial externalities surrounding infectious disease control, prevention of AIDS, environmental pollution, and behaviors (such as drunk driving) that put others at risk
- Finance and ensure delivery of a package of essential clinical services. The comprehensiveness and composition of such a package can only be defined by each country, taking into account epidemiological conditions, local preferences, and income
- Improve management of government health services through such measures as decentralization of administrative and budgetary authority and contracting out of services

Facilitate involvement by the private sector
- Encourage social or private insurance (with regulatory incentives for equitable access and cost containment) for clinical services outside the essential package
- Encourage suppliers (both public and private) to compete both to deliver clinical services and to provide inputs, such as drugs, to publicly and privately financed health services
- Generate and disseminate information on provider performance, essential equipment and drugs, the costs and effectiveness of interventions, and the accreditation status of institutions and providers

Source: Adapted from World Development Report (1993). *Investing in health*. Washington, DC: Oxford University Press.

determinants into its government-wide public health planning. Indeed, cross-sectoral action is hard to achieve across government departments in rich and poor countries alike.

Thailand offers an innovative approach to financing public health efforts with a focus on health promotion. Borrowing from the model of VicHealth in Australia, which was originally funded by a tobacco levy, The Thai Health Promotion Foundation is funded by a surcharge on tobacco and alcohol excise taxes. It spends approximately USD$100 million a year on health promoting activities and works with a wide range of multisectoral partners. Its status as an independent public agency makes it somewhat less bureaucratic and more flexible than is the case for government departments, and its earmarked tax funding gives it important financial muscle. Although it is difficult to identify its specific impact, Thailand experienced a decline in smoking among more than 15-year olds from 25.47% in 2001 to 20.7% in 2009; a reduction in the proportion of heavy drinkers from 9.1% in 2004 to 7.3% in 2009 (as well as a reduction in household expenditure on alcohol); and a fall in death rates from vehicle accidents from 22.9 per 100 000 in 2003 to 16.82 per 100 000 in 2010.

The Thai experience is very much an exception in its focus on broader public health issues. Most of the discussion on public health in low- and middle-income countries has focused on interventions which address directly the health needs of individuals. It is interesting to track the evolution of this approach. In the early 1990s, there was growing concern about what was seen as the inefficiency of health expenditure patterns in low- and middle-income countries. Although technical inefficiency was a concern, the debate especially focused on allocative inefficiency, with criticism of the high share of government health expenditure going on higher level hospital care when service coverage at the primary care level remained very poor. The landmark World Development Report of 1993 (WDR 93) introduced the DALY metric and analysis of burden of disease in terms of DALYs and also the notion of a 'package' which would respond to the largest elements of the burden. Looking back at the WDR 93, it is notable that not only there was examination of the broader context of disease which led to recommendations on fostering an environment that enables households to improve health (see **Table 4**), but also two packages were recommended, one of public health interventions and the other of essential clinical services. The former included public health action beyond that focused on individuals, including environmental pollution, and addressing drunk driving.

Although the approach of prioritizing a set of interventions has persisted in global health policy, the importance of encompassing both broader public health and clinical interventions seems to have been forgotten. The subsequent two developments of this approach, the WHO Commission on Macroeconomics and Health (2001) and the WHO High Level Taskforce on Innovative International Financing for Health Systems (2009), both defined sets of interventions consisting almost exclusively of those focused on individuals and households (**Tables 5** and **6**). The importance for health of the broader environment appears to have been forgotten. Even water and sanitation did not feature prominently, on the grounds that environmental health interventions of this kind were not considered as cost-effective for health improvement as health interventions targeting individuals.

It is interesting to reflect on the role that evaluation may have played in encouraging this development. The WDR 93 developed its dual package through analyzing both the burden of disease and the cost-effectiveness of interventions. At that time, the evidence base was weak across all areas, so much of the analysis was tentative. Since then, there has been a great increase in the volume of cost-effectiveness analysis but with a focus on specific interventions. Most commonly, the economic evaluation studies have been linked to epidemiological trials of new drugs, diagnostic methods, or more recently health service improvements. Broader public health measures lend themselves less readily to what has been perceived as the gold standard methodology of the randomized control trial.

Table 5 Commission on macroeconomics and health: list of interventions

Disease area	*Nature of intervention*
1. Maternity-related interventions	Antenatal care
	Treatment of complications during pregnancy
	Skilled birth attendance
	Emergency obstetric care
	Postpartum care (including family planning)
2. Childhood disease-related interventions (immunization)	Vaccinations (BCG, OPV, DPT, Measles, Hepatitis B, and HiB)
3. Childhood disease-related interventions (treatment of childhood illnesses)	Treatment of various conditions (acute respiratory infections, diarrhea, causes of fever, malnutrition, and anemia)
4. Malaria prevention	Insecticide-treated nets
	Residual indoor spraying
5. Malaria treatment	Treatment for malaria
6. Tuberculosis treatment	Directly observed short course treatment for smear positive patients
	Directly observed short course treatment for smear negative patients
7. HIV/AIDS Prevention	Youth focused interventions
	Interventions working with sex workers and clients
	Condom social marketing and distribution
	Workplace interventions
	Strengthening of blood transfusion systems
	Voluntary counseling and testing
	Prevention of mother-to-child transmission
	Mass media campaigns
	Treatment for sexually transmitted diseases
8. HIV/AIDS Care	Palliative care
	Clinical management of opportunistic illnesses
	Prevention of opportunistic illnesses
	Home-based care
9. HIV/AIDS HAART	Provision of HAART

HAART: highly active antiretroviral treatment.

Source: Reproduced from Table A5.1 in Annex 5 of the Report of Working Group 5 of the Commission on Macroeconomics and Health (2002). Geneva: World Health Organisation.

Table 6 High-level taskforce on innovative international financing for health systems: list of interventions

Groupings of services	*Include the following interventions*
Maternal and newborn services	Antenatal care (four visits)
	Quality facility births (maternal care during labor, delivery aid immediate postpartum)
	Newborn care (care of the newborn at birth and immediate postnatal care, including exclusive breastfeeding)
	Postnatal care (care provided to the mother up to six weeks after birth and visits at home for the newborn)
	Emergency obstetric and neonatal care (specialized care including treatment of complications during pregnancy, childbirth, and the postnatal period)
	Safe abortion (where legal) and postabortion care
	Family planning
Child services	Oral rehydration therapy
	Case management of pneumonia
	Vitamin A supplementation and vitamin A fortification
	Zinc supplementation, zinc fortification
	Access to processed food, provision of supplementary food and counseling on nutrition
	Full and permanent coverage of immunization programs
	Exclusive breastfeeding for children under six month
HIV	Prevention, treatment, and care programs for HIV
	Prevention of mother-to-child transmission
Malaria	Preventive and curative interventions for malaria
Tuberculosis	Diagnosis and treatment for tuberculosis
Noncommunicable diseases	Heath promotion and early detection of noncommunicable diseases
Presenting conditions	Diagnosis, information, referral, and relief of symptoms for any presenting conditions

Source: Reproduced from High Level Taskforce on Innovative International Financing for Health Systems (2009). *Report of Working Group 1*. Geneva: World Health Organisation.

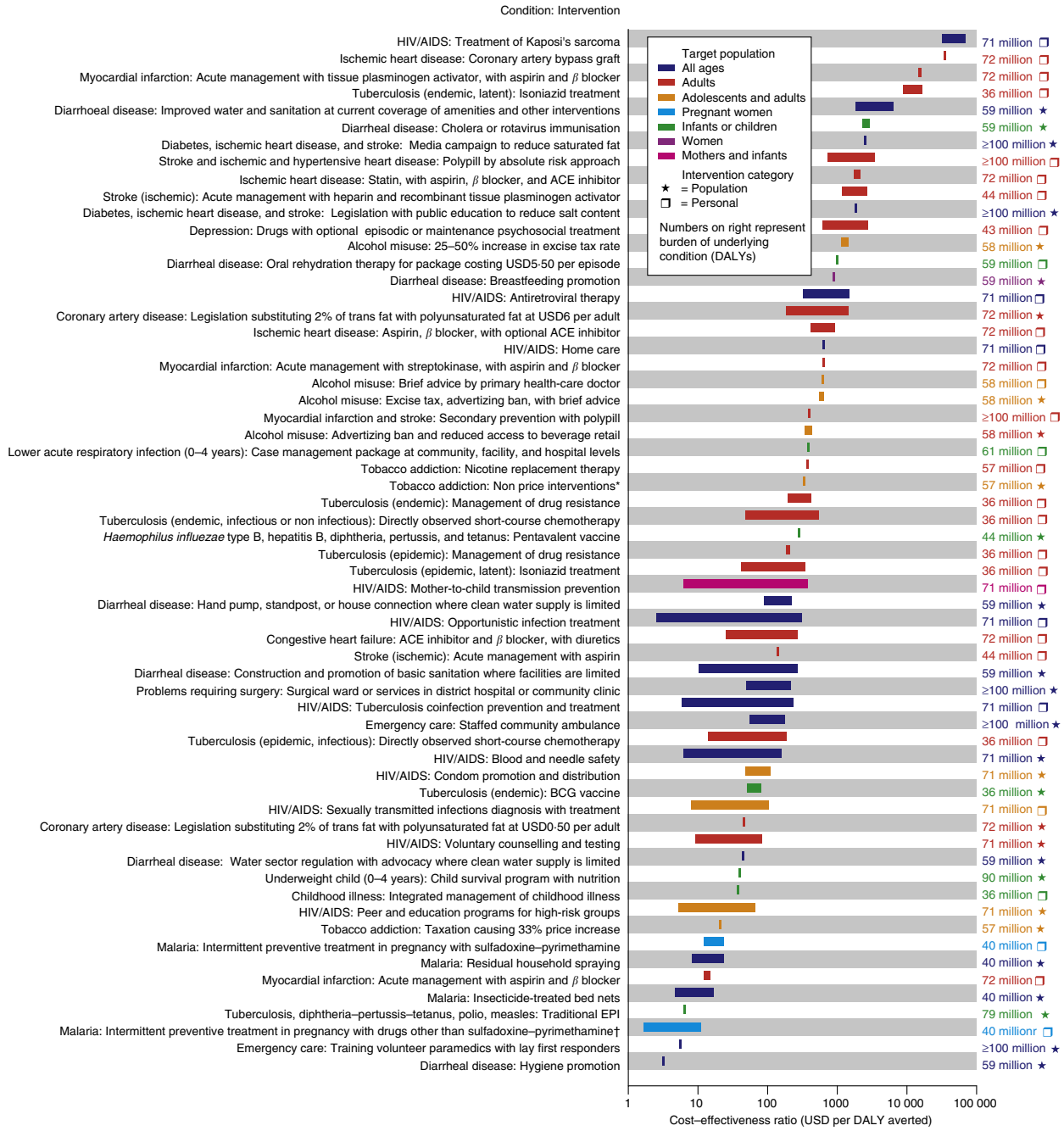


Figure 4 Cost-effectiveness of interventions related to high-burden diseases (>35 million DALYs) in low-income and middle-income countries. Bars = range in point estimates of cost-effectiveness ratios for specific interventions included in each intervention cluster and do not represent variation across regions or statistical confidence intervals. Point estimates obtained from DCP2, calculated as midpoint of range estimates reported, or calculated from a population-weighted average of region-specific estimates reported. Only interventions with cost-effectiveness reported in terms of DALYs are included in figure. *Advertising bans, smoking restrictions, supply reduction, and information dissemination. †Chloroquine = first line drug; artemisinin-based combination therapy = second-line drug; and sulfadoxine-pyrimethamine = first-line or second-line drug. Reprinted from Laxminarayan, R., Mills, A., Measham, A., et al. (2006). Advancement of global health: Key messages from the Disease Control Priorities Project. *The Lancet* **367**, 1193–1208.

Moreover, because such measures tend to have multidimensional outcomes (both different types of health benefit as well as benefits beyond health), but only single health outcomes are usually measured, they appear less cost-effective relative to interventions focusing on specific groups of individuals.

Figure 4 illustrates the problem that broader public health measures can appear less cost-effective relative to specific clinical interventions, using analyses from the Disease Control Priorities Project. This project produced a book, which provided evidence on disease burden and intervention

cost-effectiveness in 73 chapters covering specific diseases and health conditions, risk factors, consequences of disease and injuries, and clusters of services including public health, primary care, hospital care, and surgery. **Figure 4** summarizes the cost-effectiveness of interventions intended to address conditions causing a high disease burden. It is striking how few broader public health measures are included in this list, the only ones being mass media aimed at reducing saturated fat intake, and legislation to influence salt, fat, alcohol, and tobacco consumption.

The analysis on communicable disease control undertaken for the first Copenhagen Consensus (www.copenhagenconsensus.com/Home-1.aspx) also illustrates the difficulties of the evidence base. The Copenhagen Consensus examines the best way to spend aid using the metric of economic evaluation. In terms of benefits relative to costs, and based on the existing evidence, control of HIV/AIDS and control of malaria had far higher benefit cost ratios than the strengthening of basic health services. But it was clear that the nature of the evidence was highly problematic. First, the available evidence for costs and effects of specific disease control measures came mainly from trials, whereas the evidence on strengthening basic health services came from cross-country analyses, which explored the relationship between public expenditure on health care and infant and child mortality rates. One would expect trials to show higher levels of effectiveness that would not be replicated in real life. Moreover, the cross-country studies captured only the health care benefits to children rather than the population as a whole (partly because the benefits of multipurpose health services are not easily captured in a single metric, whereas metrics are readily available to evaluate disease-specific trials). Thus, it is likely that the health benefits of basic health care are underestimated.

Another example of the weaknesses of current evaluation methods, and how they can bias against broader public health action, comes from the HIV literature. Evidence is now strong that HIV incidence is declining in Uganda and Zimbabwe, for example. Yet, given what is known about the effectiveness and coverage of specific interventions, this is hard to explain. It may be that there are synergies among interventions that existing research does not capture; or that there are broader social influences at work (e.g., much greater individual awareness of the risk of HIV as a result of knowing close friends and relatives dying from HIV and behavior change as a result). These again are difficult to capture using traditional evaluation methods.

Awareness is growing of the ways in which evaluation methodologies risk distorting policy choices, especially with the increasing emphasis on requiring policy makers to base their decision making on high-quality evidence. Methods for economic evaluation need to develop beyond their current focus on single intervention approaches and single measures of outcome.

Conclusions

This article has sought to examine some of the specific issues concerned with improving public health in resource poor settings. Its prime conclusion is that a number of influences have led to the broader determinants of health, those that can be addressed by public health action beyond personal health services, being grossly neglected. It is critical that planning and evaluation move beyond their current health care and disease-specific focus, not least given the looming threat of non-communicable disease and the need to address its root causes. This will require innovations in research and evaluation methodology, as well as remedying the current global bias to control of specific diseases rather than broader health improvement.

A further challenge will be the changing global dynamics of relations between richer and poorer countries. Already China and India have graduated from low-income country status. In Africa, economic growth prospects now look brighter, especially given Africa's mineral resources. Traditional financial development assistance is likely in the longer term to decline as a source of influence on health policy. Expertise and knowledge will increasingly be of more use to the developing world than cash. Public health institutions for generating knowledge and acting on it will need to develop and evolve both globally and nationally if past health gains are to be sustained and new health risks tackled.

See also: Economic Evaluation of Public Health Interventions: Methodological Challenges. Health Status in the Developing World, Determinants of. Infectious Disease Externalities. Public Choice Analysis of Public Health Priority Setting

Further Reading

Commission on Macroeconomics and Health (2001). *Macroeconomics and health: Investing in health for economic development*. Geneva: World Health Organisation.

Dasgupta, P. (2007). *Macroeconomic history. Economics: A very short introduction*, ch. 1, pp. 1–29. Oxford: Oxford University Press.

Kelly, M. P. and Doohan, E. (2012). The social determinants of health. In Merson, M., Black, R. and Mills, A. (eds.) *Global health. Diseases, programmes, systems and policies*. Burlington, MA: Jones and Bartlett.

Laxminarayan, R., Mills, A., Measham, A., et al. (2006). Advancement of global health: Key messages from the disease control priorities project. *The Lancet* **367**, 1193–1208.

Macfarlane, S., Racelis, M. and Muli-Musiime, F. (2000). Public health in developing countries. *The Lancet* **356**, L841–L846.

Mills, A. (2011). Health Systems in low- and middle-income countries. In Glied, S. and Smith, P. C. (eds.) *Oxford handbook of health economics*. Oxford: Oxford University Press.

The Lancet (2012) The Global Burden of Disease Study 2010. *The Lancet* **380**, December 15/22/19, 2012.

Walt, G., Buse, K. and Harmer, A. (2012). Cooperation in global health. In Merson, M., Black, R. and Mills, A. (eds.) *Global health. Diseases, programmes, systems and policies*. Burlington, MA: Jones and Bartlett.

World Development Report (1993). *Investing in health*. Washington, DC: Oxford University Press.

Relevant Websites

http://www.healthmetricsandevaluation.org/gbd
Global Burden of Disease Home.

http://www.who.int/nha/en/
WHO National Health Accounts Home.

Public Health Profession

G Scally, University of the West of England, Bristol, UK

Introduction

The origins of public health can be found in ancient Greek and Roman civilizations. Many of the prominent themes in the writings of that era, such as Airs, Waters, and Places from the Hippocratic corpus, have echoes in today's major concerns about how one can have health amid both climate change and an increasing burden of noncommunicable diseases. The Greeks also developed the concept of city physicians. Their role, paid for by the city, was to look after the health of the citizens and to advise on the overall health of the city.

It is a constant difficulty for those working in public health as to how their specialty should be described. Every public health professional will be asked repeatedly during his or her career to explain exactly the meaning of public health. One useful conceptualization describes it as five different, but commonly encountered, images:

1. The system and social enterprise.
2. The profession.
3. The methods (knowledge and techniques).
4. Governmental services (especially medical, and for the poor).
5. The health of the public.

Over the years, many definitions of public health have been used. These often seem to morphose at times of organizational crisis or reorganization. At these times, a definition is needed that fits with the prevailing or future circumstances. But in modern times, whatever definition is used for public health, there remains a publicly accountable system, which is staffed by professionals who identify with the task of improving the health of human populations. The question is, however, sometimes asked as to whether public health is really a system or a profession. In times past, it was sometimes, perhaps accurately, referred to as an 'endeavor.'

The Origins of the Public Health Profession

The history of public health tells us that the major improvements in the health of populations have resulted not through the efforts of medical systems orientated toward the care of individuals with specific diseases but through the improvement of general social conditions such as housing, food supply and quality, water, and sanitation (see Figure 1). Although this is a historical perspective, being mainly associated with the nineteenth century sanitary revolution that started in England in the 1830s and 1840s, the rise in the importance of noncommunicable diseases globally, including obesity, diabetes, and alcohol-/tobacco-related diseases, has underlined the importance of primary prevention. The modern construction, the equivalent of the sanitary movement, is centered around the social determinants of health.

In the UK, the history of professional engagement with public health in a structured way dates back to the mid-nineteenth century when the post of Medical Officer of Health (MOH) was created among the English local authorities. The first MOHs were mostly part-time, who combined the local authority post with clinical practice. The first formal qualification in public health was the Diploma in State Medicine that was instituted in 1871 by Trinity College in Dublin. The breath of public health concern was illustrated by the inclusion in the syllabus of subjects such as statistics, meteorology, and engineering. It was not until the early-twentieth century that the possession of a professional qualification in public health became compulsory in Britain for those holding the MOH post. Other related qualifications, such as those awarded to sanitary inspectors, developed separately but simultaneously with the medical world.

In the USA, the first public health structures came in to being in the second half of the nineteenth century in the port cities on the East coast. By the 1870s and 1880s, most States had established their own public health structures. It was industrialization and rapid population growth that spurred the development of public health in the big cities, just as it had happened in England.

It was the inception of the National Health Service (NHS) across the UK in 1948 that created different strands of medical engagement with population health issues. The main professional public health staffing, and a wide range of public health services, had remained within the remit of local authorities. The new structures of the NHS, however, required population health skills, particularly in healthcare planning, and medical officers were appointed at a senior level within these new organizations. The skill set however was different from that required in the traditional public health role, and the existing professional organizations were not well fitted to service the future requirements of this new mixture of professional roles.

The opportunity to reconstruct the profession engaged in population medicine, in whatever role, came with the 1974 (1973 in Northern Ireland) reorganization of the NHS. It finally brought together the three key components of hospital services, primary healthcare, and public health in the NHS. The transfer of public health from local government into the NHS was not without its problems. Many of the MOHs opposed the transfer and did not appreciate the move of focus away from issues such as infectious disease, housing conditions, educational medicine, and child health. Instead, they found themselves deeply engaged in issues of healthcare management, and were frequently relegated to a purely advisory role with limited command over staff and resources.

This major change in the nature of the profession was the greatest for more than a century, and it was necessary to reconstruct the organs of the profession to match the new challenges. In particular, in order to leave behind the historical baggage of sanitarianism that attached to the title 'public

 doi:10.1016/B978-0-12-375678-7.00402-8

Figure 1 A court for King Cholera.

health,' it was felt that the branch of the medical profession dealing with population health required a new name of 'community medicine' – although this attempted change of name was short-lived as seen below forthwith. The transfer of public health functions from the local authority world into the NHS was not complete however, as environmental health responsibilities still remained with councils.

Academic Public Health

By the first half of the twentieth century, the development of the academic endeavor surrounding public health had moved substantially from its origins in the sanitary revolution. The decline of infectious diseases in Britain had resulted in a change in perspective amongst doctors who were interested in the academic questions surrounding disease prevention and control. From the 1930s onwards, a clinical perspective, which was more closely rooted in the practice of bedside medicine than in sanitarianism, had developed to become the predominant ethos of the academic public health world. The individual being most closely associated with this trend, for leading it in many ways, is John Ryle. Ryle was a political progressive who moved from his post as Professor of Physics in Cambridge University to lead the newly created Institute of Social Medicine at the University of Oxford. He believed that the new paradigm of social medicine, as it thus became, should be based on the study of disease causation in populations of patients. This became the predominant academic approach that was closely associated with the development of epidemiological methods in studying noncommunicable diseases. Although academic departments continue to teach courses leading to the Diploma in Public Health, their research has actually shifted substantially toward a social medicine focus.

The academic bedrock in the USA was established following the publication of the Welch-Rose Report in 1915. Substantial funding from the Rockefeller Foundation in 1916 has enabled the founding of what is now known as the Johns Hopkins Bloomberg School of Public Health. The subsequent development of a network of schools of public health across the USA has set the basis for the system of public health training that continues even today.

The Creation of Community Medicine

The transfer of public health responsibilities and staff to the NHS in 1974 was an opportunity to create a unified group within the medical profession of those whose activities were orientated toward improving the health of the population. Thus three strands were brought together; the former Medical Officers of Health and their staff, the medical administrators in the hospital service, and the social medicine and epidemiology academics. The new title chosen for this unified specialty was 'community medicine.' The creation of such a unified professional grouping had already been recommended by the 1968 report of the Royal Commission on Medical Education, although it took the restructuring of the NHS to give it the momentum for producing the necessary organizational changes.

The Faculty of Community Medicine became the professional organization that was created to be responsible for the training and professional development of the specialty. It was an unusual creation in that it was a faculty of not one but three medical Royal Colleges: the Royal College of Physicians

of London, the Royal College of Physicians of Edinburgh, and the Royal College of Physicians and Surgeons of Glasgow. Membership of the new Faculty was restricted to those who held a medical qualification and, following the period of its formation, to those who passed its two-part examination. It was therefore cast in the traditional mold of a medical college where the training programs followed the well-established pattern for the training of hospital consultants.

Multidisciplinary Public Health

As the great majority of national public health associations across the world have a multidisciplinary membership, the World Federation of Public Health Associations will not admit into membership an association that draws its membership from only one professional background. The American Public Health Association, founded in 1872, has a long tradition of multidisciplinary public health working and, in its case, it has been seen to add to the general strength of the professional group. In the USA, public health training has always been multidisciplinary. By 1938, federally funded training had been provided to more than 4000 people in schools of public health, of whom only approximately 1000 were medically qualified.

In the UK, although the move toward multidisciplinary public health was difficult for many of the more traditionally minded members of the specialty, the way forward was eventually greatly helped by the example of the Royal College of Pathologists that had for some time admitted both medical and nonmedical members. It was the creation of the new specialty grouping of community medicine in the 1970s that created the tensions. The very fact that training for senior positions in the new system was so closely modeled on the medical training scheme, and that the newly formed Faculty of Community Medicine only opened its doors to registered medical practitioners, was regarded as little short of an insult by many of the distinguished academics from disciplines other than medicine, who had been making such a substantial contribution to the various academic departments of social and preventive medicine across the country. The attempt to persuade academic departments to adopt the uniform title of 'Department of Community Medicine' was a failure. Many departments continued to operate with their traditional titles, whereas very few adopted the new title.

As the service component of community medicine found its footing in the new NHS structures during the 1980s, the skill mix began to develop in the building of departments. The multidisciplinary trend was particularly prominent in the growth of health education units as well as their development into the new and more progressive approach of health promotion. Similarly, the requirements in the new organizations for advanced skills in the handling and analysis of large datasets, accompanied by the development of small and usable computers, led to the development of groups of staff with significant epidemiological and statistical skills. The gradual opening up of postgraduate courses in public health to students from disciplines other than medicine meant that there was the beginning of a professional development pathway for nonmedical graduates that would to some extent mirror that of the doctors.

The steady growth of multidisciplinary working in both academic and service settings had gradually increased the demands for a proper career structure for nonmedics working in public health as well as for access to established and recognized training routes. At one point, there was a danger that the specialty would split into two or more professional groupings. However, with a great deal of diplomatic activity, it had eventually become possible to bring together the different factions. In 1997, the Tripartite Group (the Faculty of Public Health Medicine, the Royal Institute of Public Health and Hygiene, and the Multidisciplinary Public Health Forum) signed the Tripartite Agreement taking forward the development of multidisciplinary public health. This formed the basis for admission into the Faculty, on an equal basis, of public health professionals from different professional backgrounds. It also paved the way for equal access to official training posts in the specialty.

Training in Public Health

Across the world there are various routes of entry into specialized public health work. The most common by far is through studying for a Masters level degree in public health at a University or School of Public Health. In many countries, such training may be supplemented by further study to obtain a Doctorate level qualification, which may be obtained through a taught route or by research. This route is usually open to graduates from a wide range of disciplines and to those from vocational backgrounds such as nursing.

The US in particular has a very substantial number of Masters in Public Health (MPH) courses, and approximately 15 000 students study for a MPH every year. The Council on Education for Public Health accredits courses in public health in the USA and has recently started to operate internationally with accreditation taking place in Canada, Mexico, and France. Canada in particular has seen what is described by some as an 'explosion' in MPH courses.

The growth of academic qualification in public health is rightly seen as a bedrock for good public health practice in society. In some countries, however, a longer training period, that usually includes a MPH component, is regarded as the norm for those wanting to become specialists in public health. This is a system that is often based on the British approach to training, which parallels the system for training of doctors in clinical specialties. The approaches adopted in Ireland, Australia, and New Zealand – all involve a significant period of work-based training attachment as well as, in some cases, qualifying examinations.

System Failure

Although community medicine had become embedded within the NHS systems in the UK during the 1970s and 1980s, this had meant a substantial move away from the origins of public health, which were based on the concept of environmental concerns and infectious disease. The connection with NHS

management and the involvement in the functions of NHS administration had meant that public health practitioners had moved ever further away from their origins, and this was not without consequence. There were a series of serious failures of the public health system, resulting in a significant number of deaths. Notably, these included the major salmonella outbreak in 1984 at the Stanley Royd Hospital in Wakefield and the 1985 Legionnaires' disease outbreak at Stafford General Hospital. The then Chief Medical Officer (CMO) of England, Sir Donald Acheson, chaired a review of the public health system and published a report in 1988 entitled 'Public Health in England.' The major thrust of the report was that the specialty had drifted too far from its roots, neglecting some of the major risks to the population's health. Two important outcomes were that the title 'public health' should be restored to the specialty and that the senior post holder at a local level should be designated as Director of Public Health (DPH). He or she was also mandated to produce an annual report on the health of the respective population in much the same way as the predecessor, the MOH, had done.

Doctors specializing in the control of infectious disease became separated from the generalist public health professionals in due course, a move that was reinforced by the incorporation of doctors into a new national body known as the Health Protection Agency, which itself disappeared in 2013. This separation of communicable disease control from general public health is contentious, and is regarded as creating a fault line in the specialty.

The Chief Medical Officer

It is very common for national health systems to have an individual operating at national level with the key responsibility for the population health aspects of the country's health. Inevitably, a range of titles are used to designate such a role, but internationally, the generic title of CMO is often used – in the European Union, for example – even though in some instances, the incumbent may not be medically qualified. A study of all the counties in the European Union has shown that CMOs operate in a wide range of roles. This might be within the central Government Department of health concerns or within a separate agency in charge of undertaking national public health responsibilities. The role of the CMO also ranges from being purely advisory to having substantial executive powers and numerous staff. Very few European countries do not have any identifiable CMO-type posts.

In England, the local post of MOH preceded the post of CMO. The first MOH known to be appointed for a full-time post in the UK was Dr. William Henry Duncan of Liverpool. He was appointed in 1847 as a result of a private Act of Parliament that preceded the 1848 Public Health Acts. He became a well-known figure in the city to the extent of sharing with his contemporary, the famous London physician Dr. John Snow, the accolade of having a public house named in his honor. But although MOHs became prominent local figures, it was indeed the creation of the public health post currently known as CMO at the heart of Government that had become the most enduring one.

The first holder of the post of CMO was Dr. John Simon. He had been an active and outspoken MOH for London, who became the CMO of the General Board of Health in 1855. Never one to shy away from controversy, Simon continued to be a passionate advocate for the health of the population. He demanded that the post of CMO should be of prominence in the structure and functioning of the General Board of Health, and subsequently in the Privy Council. Although he resigned eventually because of the downgrading of the post – particularly on the issue of allowing direct access to ministers, yet Simon had firmly established the principle of a chief public health advisor to the government and the post continues to this day. Although it is nowhere specified that the CMO has to come from a public health background, this has in effect been the position until relatively recently. Only two of the CMOs in England have come from outside the public health system. The requirement to publish an annual report on the health of the population has been a key task of the CMO, which has been seen as analogous to the duty that fell to the MOH, and subsequently, to the DPH, at a local level. The CMO post is, however, under threat in the English system, as the current and 16th incumbent is not only from a nonpublic health background but is also on a short-term contract, and the post itself has been merged with the most senior research and development post in the Department of Health.

The USA has a similarly long-lived tradition of having a doctor close to the center of government. The first Surgeon General of the USA was appointed in 1871 and, unlike the UK position, is a political appointee. The incumbent holds office at the pleasure of the President, and although there is a tradition and public expectation that the Surgeon General will speak out on controversial issues, this at times has led to the President's intervention to dismiss him or her. This has been seen most recently in the dismissal of a Surgeon General by President Clinton because of her statements on sexual health. Former Surgeon Generals have complained publicly regarding political interference in their erstwhile official roles (**Figure 2**). As in the UK, there is a presumption that the post of Surgeon General will be appointed from within the existing public health medical workforce.

A very significant difference between the public health workforce in the USA and the UK is with regard to their official status. In the USA, the Public Health Service Commissioned Corps is one of the uniformed services of government and its uniformed staff is therefore subject to a degree of military style discipline, with the Surgeon General holding the rank of Vice-Admiral in the service.

Future Directions

The development of international cooperation between organizations representing public health professionals appears to be on a steady upward trajectory. The African Federation of Public Health Associations was launched in April 2012, since then representing the latest step in the creation of an effective global, regional, and national network of public health bodies. The basic priority internationally is the development of a global approach to the public health workforce, which would recognize that strengthening the training and the role of

Figure 2 Three former surgeon generals testifying before congress regarding political interference in their role.

public health professionals are key elements for improving global health.

In England, the most recent changes to the NHS have moved public health in a very different direction from the situation in other parts of the UK. The return of a substantial proportion of public health functions to local authorities is a major reversal of the 1974 reorganization. Similarly, the creation of Public Health England, an executive agency of the Department of Health, is remarkably a substantial centralization of power and authority. This centralization is to a greater extent than anything seen hitherto in public health in the UK. Meanwhile, public health in Scotland, Wales, and Northern Ireland continues to be closely associated with the NHS. The effect on professional practice of the changes in England is not yet discernible. The Coalition Government in England has agreed to implement statutory regulation of the specialist cadre of the profession, and this may provide some protection in respect of the *laisser faire* approach that is likely to accompany the local control, which will rest with individual local authorities.

There is a real opportunity arising from the move to local government in England. Many of what are now known as the social determinants of health lie within the remit of local authorities. The ability to influence planning, housing, leisure and recreation, education, economic development, etc. is a prize worth griping. The ability to function effectively within what is a radically different environment is however likely to require a different set of skills from those most recently deployed in the NHS. In particular, the ability to deal efficiently and effectively with, and win the respect of, elected politicians will be at a premium.

The move of a substantial proportion of the public health workforce in England into local authorities will give a new opportunity to rethink approaches to ensuring the quality of public health practice. The current model that is based largely on processes such as audit and revalidation, which are drawn from the clinical world, may well prove to be inadequate in a world where professional hierarchy is dissolved. Instead, new approaches that aim to provide assurance regarding the quality of local public health departments may evolve. This may well be based on recent experience from the USA, where they have been trying to cope with a devolved system that displays significant variation in the quality of public health practice. The development of USA style accreditation systems for local public health departments is one way in which professional standards and development can be assured at a time of increased devolution of authority.

One of the important changes in recent decades has been the way in which doctors in clinical practice have moved away from engagement with preventative medicine and the major public health issues of the day. This is in stark contrast to the successes of the broader medical profession during the later half of the twentieth century in relation to issues such as tobacco, seat belts, crash helmets, and car windscreens. There have however been stark warnings that health services in developed countries will become unaffordable unless there is a wholehearted and wholesale engagement with primary prevention. If this is to materialize, stronger links will need to be forged between clinical medicine in health services and the operation of local authorities and others with control over the determinants of health. This has the potential to usher in a new era of preventative medicine in which the barriers between clinical medicine and public health that have been painstakingly erected over the past hundred years can start to be demolished.

See also: Ethics and Social Value Judgments in Public Health. Priority Setting in Public Health

Further Reading

Acheson, R. M. (1980). Community medicine: Discipline or topic? Profession or endeavour? *Journal of Public Health* **2**(1), 2–6.

Detels, R., Beaglehole, R., Lansang, M. A. and Gulliford, M. (2009). *Oxford textbook of public health*, 5th ed. (Online Oxford medicine edition 2011, doi: 10.1093/med/9780199218707.001.0001). Oxford University Press.

Donaldson, L. J. and Scally, G. (2009). *Donaldsons' essential public health.* Oxford: Radcliffe.

Fee, E. and Acheson, R. M. (1991). *A history of education in public health.* Oxford: Oxford University Press.

Jakubowski, E., Martin-Moreno, J. M. and McKee, M. (2010). The governments' doctors: The roles and responsibilities of chief medical officers in the European Union. *Clinical Medicine* **10**(6), 560–562.

Pencheon, D., Guest, C., Melzer, D. and Muir Gray, J. A. (2006). *Oxford handbook of public health practice*. Oxford University Press. (Oxford medicine online edition 2010, doi: 10.1093/med/9780198566557.001.0001).
Porter, D. (1999). *Health, civilization and the state*. London: Routledge.
Sheard, S. and Donaldson, L. (2005). *The nation's doctor: The role of the chief medical officer, 1855–1998*. Oxford: Radcliffe Medical Press.
Turnock, B. J. (2004). *Public health: What it is and how it works*, 3rd ed. Gaithersburg, MD: Aspen Publishers.

Relevant Websites

http://jech.bmj.com/content/57/3/164.full
Berridge and Loughlin's Glossary of Public Health History.
http://www.enotes.com/history-public-health-reference/history-public-health-173217
eNote History of Public Health.
http://www.academicearth.org/lectures/the-sanitary-movement-and-the-filth-theory-of-disease
Frank Snowden Yale Lecture on the Sanitary Movement.
http://www.healthknowledge.org.uk/public-health-textbook
HealthKnowledge Free On-Line Public Health Textbook.
http://www.eupha.org/site/history.php
History of the European Public Health Association.
http://www.cdc.gov/about/history/ourstory.htm
History of the US Centers for Disease Control and Prevention.
http://www.who.int/about/history/en/index.html
History of WHO.
http://new.paho.org/
Pan American Health Organization.
http://www.sagepub.com/upm-data/3989_Chapter_1.pdf
Population-Based Public Health Practice.
http://www.fph.org.uk/about_us
The UK's Faculty of Public Health.
http://www.fph.org.uk/what_is_public_health
UK Faculty of Public Health Definition of Public Health.
http://www.whatispublichealth.org/what/index.html
US Association of Schools of Public Health 'What is Public Health?'.
http://publichealth.jbpub.com/turnock/3e/sample_chapters.cfm
What is Public Health? (Chapter 1 of Public Health: What It Is and How It Works by Bernard Turnock).
http://www.who.int/social_determinants/en/
WHO Social Determinants of Health.
http://en.wikipedia.org/wiki/Public_health
Wikipedia Entry on Public Health.
http://www.wfpha.org/
World Federation of Public Health Associations.

Public Health: Overview

R Cookson, Centre for Health Economics, University of York, York, UK
M Suhrcke, Norwich Medical School, University of East Anglia, Norwich, UK, and UKCRC Centre for Diet and Activity Research (CEDAR), Cambridge, UK

Glossary

Bounded rationality The idea that people may be content with decisions that are merely satisfactory rather than ideal, that they operate by rules of thumb, take short-cuts, etc.
Cost-benefit analysis A form of economic evaluation by comparing the costs and the (money-valued) benefits of alternative courses of action.
Cost-effectiveness A measure of the cost per desired outcome or effect of an intervention or course of action. Whether a given intervention is considered cost-effective typically depends on how it compares to other relevant alternatives with similar outcomes; the intervention with the lowest cost per desired outcome is the most cost-effective intervention.
Epidemiology The study of the relationship between risk factors and disease in human populations, including factors that can change the relationship and the application of such analysis to the design and management of health care systems.
Marginal The additional benefit, health, cost, etc. attributable to a small increase in a factor bringing it about (other things equal).
Observational data Data from studies that observe 'what is' without observer intervention, say, in the form of creating controls or blinding or randomizing.
Public choice theory Public choice study includes collective decision-making and political behavior. Analysts model voters, politicians, and bureaucrats as mainly self-interested. It also includes the study of constitutions and constitutional change.
Quasi experimental or natural experiment Comparative research in which the assignment of subjects to comparator groups is not random or a control group is not used.
Randomised control trial A scientific experiment conducted to test the effect of an intervention by randomly assigning participants to a treatment and control group. Differences between the treatment and control group participants are interpreted as the causal effect of the intervention.
Rent seeking The processes through which individuals and corporations seek to use government to promote their own interests and, in particular, to acquire streams of money (rents). An example is members of a regulated industry manipulating the regulatory agency.

Introduction

The phrase 'public health' can be used to mean (1) population health, or (2) public policy intervention to prevent ill health. This article focuses on public health in the latter sense, as used by the public health profession and the broader public health community. However, the discipline of economics also has much to contribute to understanding public health in the former sense. So by way of background, this introductory section lists a few of the many contributions that economists have made to measuring population health and analyzing its determinants. One can read about these and many other important economic analyses of population health in other entries in the encyclopedia.

Contributions by economists to measuring population health include work by:

- Alan Williams and George Torrance in helping to develop the quality-adjusted life-year measure of overall health.
- Christopher Murray in helping to develop the Disability Adjusted Life Year measure of overall disease burden and the Global Burden of Disease reports, together with epidemiologist Alan Lopez.

What is distinctively 'economic' about these contributions, compared with contributions by clinicians, epidemiologists, psychologists, and others, is the development of overall summary measures of health that allow diverse mortality and morbidity outcomes from diverse health conditions to be compared with one another in terms of a common generic unit of health.

Contributions by economists to analyzing the determinants of population health include work by:

- Samuel Preston on distinguishing the contributions of income growth and new technology to improvements in population health in the twentieth century.
- Victor Fuchs on distinguishing the total contribution of healthcare to population health from the much smaller marginal contribution of additional health care expenditure at the current level of medical technology.
- David Cutler and Mark McClellan on the substantial health benefits of medical innovation in the latter half of the twentieth century, building on work by anesthesiologist John Bunker.
- Angus Deaton on disentangling the relationships between income, health, and wellbeing, including work addressing the hypothesis of social epidemiologist Richard Wilkinson that income inequality is a health hazard.
- David Grossman on the concept of health capital and the contribution of health capital investments over the life-course to the production of health.
- James Heckman, Janet Currie, and Robert Fogel on the contribution of *in utero* and early childhood circumstances to health and human capital formation, building on work by epidemiologist David Barker.

 doi:10.1016/B978-0-12-375678-7.00401-6

- Garry Becker on the theory of rational addiction and subsequent empirical work by Frank Chaloupka and others, confirming some (though not all) of its testable predictions in relation to smoking and other unhealthy addictive behaviors.
- Tomas Philipson on economic epidemiology and the role of prevalence-elastic prevention behavior in determining the prevalence of infectious disease.
- Don Kenkel on the role of antismoking public sentiment as a cause of both antismoking public policy and declining smoking rates, an example of the general issue of 'endogenous policy'.
- Harold Holder on general equilibrium modeling of alcohol consumption ('SimCom'), including feedback loops between alcohol consumption and policy formation.

What is distinctively 'economic' about these contributions includes the focus on marginal analysis (since marginal effects on health are more relevant to decision makers than average or total effects) and the focus on understanding how the health-related behavior of individuals and organizations changes in response to changes in their incentives and constraints.

Other distinctive characteristics of these contributions include the recognition that individuals and governments have important objectives other than health improvement, the explicit modeling of complex causal pathways, and the focus on seeking robust estimates of effect using experimental and quasi-experimental methods. However, it is less clear that these are distinctively 'economic' characteristics as opposed to distinctive characteristics of high quality public health and social science research, more generally.

The Nature and Scope of Public Health Intervention

Public health intervention is an important topic, for two reasons. First, preventing ill health is an important objective. Bad health is not only intrinsically bad but also instrumentally bad, as it makes it harder for people to lead flourishing lives and contribute to society by undertaking productive work, family, and social activities. Second, history suggests that public health intervention can succeed in preventing ill health. Careful analysis of historical mortality and fertility records by historian Simon Szreter and others has shown that the nineteenth century 'sanitary movement' and other historical public health interventions did contribute to the steady improvements life expectancy seen in the past 200 years, despite earlier findings to the contrary by physician Thomas McKeown.

Public health intervention is also a broad topic. In a 1920 article in *Science*, entitled 'the untilled fields of public health', the renowned US bacteriologist and professor of public health at Yale, Charles-Edward Amory Winslow, defined public health as: "the science and art of preventing disease, prolonging life, and promoting physical health and efficiency through organized community efforts for the sanitation of the environment, the control of community infections, the education of the individual in principles of personal hygiene, the organization of medical and nursing service for the early diagnosis and preventive treatment of disease, and the development of the social machinery which will ensure to every individual in the community a standard of living adequate for the maintenance of health." In 1988, the US Institute of Medicine put it more generally, and more succinctly: "Public health is what we, as a society, do collectively to assure the conditions for people to be healthy."

The scope of public health intervention thus potentially encompasses any kind of population level policy instrument (see **Box 1**) implemented by any kind of government or nongovernment organization or group in any sector of social or economic policy (see **Box 2**), which is undertaken with the (not necessarily exclusive) aim of preventing any kind of disease, illness, disability or injury, whether physical or mental, fatal or nonfatal, mild or severe.

Box 1 Public health policy instruments

Eliminate choice through regulation	For example, compulsory isolation of patients with highly infectious disease For example, prohibition of narcotics and prostitution
Restrict choice through regulation	For example, restrict the location and timing of alcohol sales For example, ban smoking in public places
Guide choice through disincentives	For example, cigarette sales taxes, alcohol minimum prices For example, parking and congestion charges
Guide choice through incentives	For example, tax breaks for work-related bicycle purchase For example, payments for stopping smoking in pregnancy
Guide choice through 'nudges'	For example, regulations requiring 'nonneutral' food labeling For example, changing the default option
Enable choice through public funding	For example, public funding of public goods, such as sanitation infrastructure, green spaces, cycle lanes For example, public funding of private goods, such as primary care, sports facilities, free fruit in schools
Provide information through public funding and regulation	For example, publicly funded research on health risks For example, public information campaigns to inform people about health risks and encourage healthy behavior For example, advertising standards, 'neutral' food labeling

Note: The types of policy instrument are listed in ascending order of restriction on individual freedom, based on the Nuffield Council on Bioethics 'interventions ladder.'

Box 2 Public health policy sectors

- Preventive healthcare policy – for example, screening, vaccination, oral health, mental health, child and maternal health; includes secondary prevention such as medication for heart disease as well as primary prevention, such as dietary advice for people at risk of developing heart disease.
- Health education policy – for example, communicating information about disease prevention and healthy lifestyles via clinics, schools, workplaces, and social media.
- Health protection policy – for example, disease surveillance, responses to major incidents, emergency planning, and managing outbreaks of infectious disease; may involve nonhealthcare professionals such as fire services, the police, and army.
- Safety policy – for example, workplace safety, transport safety, and domestic safety.
- Crime policy – for example, policies to prevent homicide and domestic violence.
- Sanitation policy – for example, sewage, waste disposal, and water quality.
- Food policy – for example, food hygiene and safety, nutritional labeling, and food provision.
- Local government policy – for example, policy on housing, green spaces, congestion, air quality, and social services.
- Family policy – for example, preschool education and child welfare services.
- Education policy – for example, general primary and secondary education.
- Employment policy – for example, employment rights, job centers, and regional subsidies.
- Trade policy – for example, international safety standards and intellectual property.
- Social protection policy – for example, unemployment benefits, disability benefits, childcare benefits, and pensions.
- Taxation policy – for example, income taxes, property and inheritance taxes.

Public health differs from healthcare insofar as it involves public policy intervention to reduce the risk of future ill health, rather than to treat current ill health. The risk reductions caused by particular public health interventions are often small and imperceptible at individual level, but can add up to large and tangible benefits at population level. Indeed, public health interventions that deliver small reductions in individual health risk to a large population of relatively healthy people can offer greater total health benefits than healthcare interventions that deliver large individual benefits to a small population of relatively unhealthy people. This is known as the 'prevention paradox', a term coined by the epidemiologist Geoffrey Rose.

Two great pioneers of public health in the nineteenth century were Edwin Chadwick and John Snow. In 1843, Chadwick's Report on the Sanitary Condition of the Labouring Population of Great Britain helped catalyze the sanitary movement that substantially contributed to increases in life expectancy across the globe. Snow is widely considered to be the father of modern epidemiology, following his classic 1855 treatise, On the Mode of Communication of Cholera. Among other things, this treatise reports his famous 1848 study that convincingly traces the cause of an outbreak of cholera in London to the Broad Street water pump by collecting data to test rival hypotheses.

In the nineteenth century, the central task of public health intervention has been to prevent communicable or infectious diseases, to which young children are particularly vulnerable. These 'infectious diseases of childhood' are now reasonably well controlled in most parts of the world, but there is still a substantial burden of disease among young children in much of Africa and Asia from cholera and other diarrheal diseases, lower respiratory diseases, meningitis, tetanus, measles, tuberculosis, malaria, HIV/AIDs, leishmaniasis, hepatitis, leprosy, and other infectious diseases (Figure 1).

Infectious agents such as viruses, bacteria, and parasites or fungi can spread directly from person to person, and people can also shed them into the air or water or onto food or other surfaces where other people may come into contact with them. Infectious diseases therefore generate 'technological externalities': one person can change another person's risk of infection through their actions, without bearing any of the costs or gaining any of the benefits of that change. Externalities provide a standard economic rationale for government intervention to prevent infectious disease, for example though investment in sanitation infrastructure or quarantine regulations. Public infrastructure investments to prevent infectious disease – such as building sewers, or draining malaria-infested swamps – can be seen as 'public goods' in the technical economic sense of being nonexcludable and nonrival: no one can be excluded from use, and one person's use does not reduce the good's availability to others. Governments have a role in providing such public goods, because markets have difficulty providing goods that customers can easily consume within paying anything.

Since the nineteenth century, much of the world has undergone a 'demographic transition' from high to low rates of birth and death. This has important implications for public health in the twenty-first century, which is increasingly focusing on the prevention of noncommunicable chronic diseases and disorders to which older people are particularly vulnerable, such as circulatory diseases, cancers, diabetes, neurological disorders, and musculoskeletal disorders. The nature of this prevention task is different, as one of the main ways of preventing (or, at least, delaying) these 'chronic diseases of old age' is to encourage people to adopt healthier lifestyle behaviors in relation to diet, physical activity, smoking, drinking, substance abuse, and musculoskeletal load. Technological externalities are largely irrelevant to lifestyle behavior, insofar as an unhealthy lifestyle only harms the individual's own health – though of course, there are exceptions such as passive smoking and drink driving. So public health interventions to promote healthy lifestyles are not 'public goods' in the technical economic sense. However, there may be other economic justifications for such interventions, as discussed below.

In most countries, public health interventions in each of the policy sectors listed in Box 2 are planned and implemented by at least one and usually many different organizations. It therefore stretches credulity somewhat to talk about a public health 'system,' as if the organizations in these diverse areas of social and economic policy were all exclusively designed for the purpose of working together to improve population health. Nevertheless, most countries do attempt a degree of coordination between public health interventions in different policy sectors, in at least two ways. First, though an officially recognized 'public health profession,' such as the US

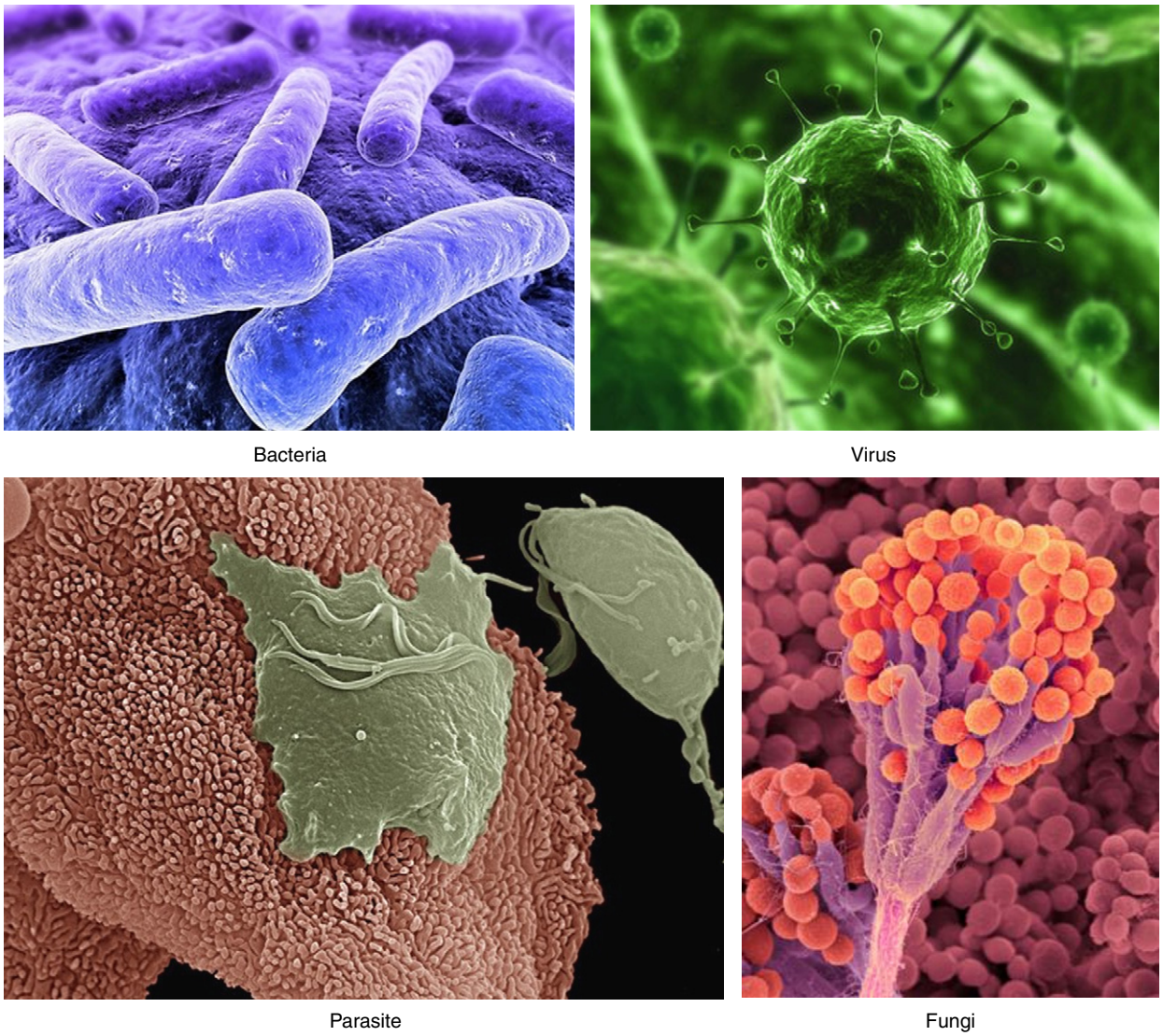


Figure 1 Some microorganisms. These four images are taken from the web from the following four different sites: Bacteria http://www.universityobserver.ie/2012/02/22/bacterial-arms-race. Virus: http://science.howstuffworks.com/environmental/life/cellular-microscopic/light-virus.htm. Parasite: http://sciencehubb.co.uk/my-enemys-enemy-is-my-enemy. Fungi: http://sharon-taxonomy2009-p2.wikispaces.com/Fungi.

Public Health Service Commissioned Corps and the UK Faculty of Public Health, whose members implement many different public health functions and lead some of the relevant public policymaking agencies. Second, through the appointment of senior policymakers and policymaking agencies with responsibility for cross-government coordination of public health policy, such as the Surgeon General and the Centers for Disease Control and Prevention in the US, and the Chief Medical Officer and Public Health England in the UK.

Economic Arguments for Government Intervention in Public Health

One can distinguish five types of normative economic argument for government intervention in public health:

1. Asymmetric information
2. Technological externality and public goods
3. Pecuniary externality
4. Paternalism and bounded rationality
5. Equity

The first two are classic 'market failure' arguments, which show how fully rational and self-interested market participants can fail to achieve a Pareto efficient outcome due to the presence of a single distortion or imperfection in an otherwise perfect market setting. When markets fail in this sense, it may be possible for government intervention to deliver a Pareto improvement that makes at least one person better off without making anyone else worse off – though this possibility may be constrained by sources of government failure, such as asymmetries of information between government officials and market participants, and self-interested behavior by government officials. The third type of argument relies on the welfare economic 'theory of the 2nd best' in the presence of more than one market distortion. The fourth type of argument relies on

individuals being less than fully rational and self-interested. The final type of argument goes beyond market failure in terms of Pareto inefficiency and analyses distributional concerns for equity or justice.

Asymmetric information refers to a situation in which one party to a market transaction has better information than another party. Healthcare markets are pervaded by asymmetries of information – between doctors and patients, insurers and insurees, buyers and sellers of new medical technology, and so on. These asymmetries may help to justify government intervention in both curative and preventive healthcare. Information asymmetries are also relevant outside the healthcare market. For example, sellers may have better information than buyers about the health risks associated with consuming their goods and services; and employers may have better information than employees about the health risks associated with their working conditions. This asymmetry can provide a market failure argument for ex ante safety regulation (e.g., health and safety requirements enforced by licensing and inspection processes, advertising standards, requirements for provision of safety information) and ex post tort law compensation claims for health damages caused by transactions made on the basis of hidden information. Note, however, that it is the asymmetry of information between market participants about health risk that distorts market behavior and generates market failure, rather than the mere presence of health risk or the mere lack of perfect information about the true nature of health risk. For example, asymmetry of information about the health risks of smoking in the 1950s between tobacco company executives and consumers may have generated market failure. By contrast, continuing uncertainty and imperfect information among all market participants about how far smoking will damage any particular individual's health do not generate 'market failure' in the classic economic sense.

In their renowned 1988 textbook on the theory of environmental policy, William Baumol and Wallace Oates define a technological externality as follows: ''An externality is present whenever some individual's (say A's) utility or production relationships include real (that is, nonmonetary) variables, whose values are chosen by others (persons, corporations and governments) without particular attention to the effects on A's welfare.'' This definition clearly applies to infectious disease externality, because individual A's utility depends on the number and type of infective microorganisms present in their living and working environments, which in turn depends on choices made by other persons, corporations, and governments. It also applies to passive smoking, drunk driving, and other cases, in which individual A's risk factors for noncommunicable disease or injury are directly influenced by other people's choices. It is less clear whether it applies to cases in which individual A's health risk factors are indirectly influenced by other people's choices through their influence on individual A's own choices – for example, choices generating congestion and crime in the local area, which influence individual A's choices about physical activity.

As described earlier, an important class of technological externalities in public health are nonexcludable and nonrival public goods such as investment in sanitation infrastructure to prevent the spread of infectious disease. Public investment in basic universal healthcare and education systems in low and middle income countries also has public good characteristics. Almost everyone is better off living in a high income country with a healthy population and a growing economy (even the super-rich). Yet, the market alone may fail to coordinate this large and sustained infrastructure investment, precisely because almost everyone benefits, whether they pay or not. Another important example of a public good is the creation of new information about health risk through research and development (R&D). R&D is a nonrival and nonexcludable good insofar as the new information it generates can subsequently be acquired by potential beneficiaries at very low cost and is hard to keep secret. These 'information externalities' may help to justify R&D subsidy and government regulation of intellectual property rights, such as patent protection and copyright legislation. It is less clear that information externalities help to justify public health information campaigns, however, because the transmission of existing information (as opposed to the generation of new information) often has the characteristics of a private good – for example, leaflets (as opposed to the information they contain) are excludable goods.

A quite different form of externality arises in the case of external costs imposed upon taxpayers due to public expenditure on health and social care. Here, the externality is monetary or 'pecuniary' in nature, and not a real variable entering into utility or production relationships. According to the welfare economic 'theory of the 2nd best', pecuniary externalities can nevertheless cause market failure to achieve a constrained Pareto efficient outcome in economies with multiple market imperfections (such as information asymmetries, taxes, and so on). However, this argument needs to be used with caution, as policy prescriptions from '2nd best' welfare economic analyses are context-dependent and sometimes counter-intuitive, and it is hard to construct realistic models of actual economies with multiple imperfections.

Another note of caution is that pecuniary externalities are ubiquitous, and can be used as a spurious justification by all sorts of interest groups seeking special favors. For example, there is a pecuniary externality argument for subsidizing private schools, on the grounds that sending a child to private school may reduce the cost of operating public schools. A further note of caution is that pecuniary externality arguments for public health intervention can be a double edged sword. For example, preventing smoking may reduce taxpayer expenditure on healthcare for lung cancer but may increase taxpayer expenditure on pensions and long-term care for those who survive longer. Hence, whether the pecuniary externality associated with a particular form of unhealthy behavior is positive or negative is an open empirical question, and will depend on the context. A final note of caution is that the root cause of pecuniary externalities on taxpayers is government intervention – in this case, public programs offering free health and social care. Some economists argue that limiting entitlements to free health and social care may be a more attractive way of reducing pecuniary externalities than introducing new taxes on unhealthy behavior. Whatever the pros and cons of the latter argument, policymakers do need to bear in mind that taxes can have high administrative costs and unintended behavioral effects. For example, the Danish

government introduced a tax on saturated fats in 2011 but rescinded it a year later. According to a newspaper report in The Economist in November 2012, "in practice, the world's first fat tax proved to be a cumbersome chore with undesirable side effects. The tax's advocates wanted to hit things like potato crisps and hot dogs, but it was applied also to high-end fare like speciality cheeses... Besides the bother and cost of installing new systems to calculate the extra tax, retailers were also hit by a surge in cross-border shopping."

The fourth argument – paternalism and bounded rationality – rests on the view that individuals sometimes fail to act in their own best interests, for example, due to weakness of will or limited information-processing ability. There is by now plenty of evidence from behavioral economics and psychology that individual rationality is imperfect in various ways. For example, people's choices are often strongly influenced by nonrational 'cues' in their decision-making environment. In their book, 'Nudge,' Cass Sunstein and Richard Thaler use this kind of evidence as an argument for 'soft' paternalism, which involves altering the nonrational cues in order to gently 'nudge' people toward choices in line with their own best interests as perceived by the paternalistic decisionmaker. However, one can also use evidence of bounded rationality to make a case for 'hard' paternalism involving traditional public health instruments, such as taxes, subsidies, and regulations, which alter the incentives and constraints that people face. In the phrase of Adam Oliver from the London School of Economics, there may be a role for interventions that firmly 'budge' people toward rational ill health prevention behaviors as well as interventions that gently 'nudge' them.

Finally, the fifth argument – equity – relates to concerns about distributional fairness rather than Pareto efficiency. Some economists take the view that distributional concerns are not a proper subject for economic analysis (e.g., the great early twentieth-century economist, Lionel Robbins). However, other economists (e.g., Tony Atkinson and Amartya Sen) adopt a more inclusive 'social choice' approach to normative economics based on explicit analysis of social objectives, which may or may not include Pareto efficiency. Markets may give rise to substantial social inequalities in health and in ill health prevention activities, and social decision-makers may regard the reduction of such inequalities as a policy objective. This is a 'specific egalitarian' objective, rather than the 'general egalitarian' objective of redistributing income. According to classical '1st best' welfare economic theory, redistribution of income is the most efficient way to reduce inequality in the distribution of welfare between individuals, rather than government intervention in specific markets such as the market for ill health prevention services. However, this theoretical result does not carry over into '2nd best' economies with multiple imperfections, and so there is no general economic case against 'specific egalitarian' policy objectives. Nevertheless, there are dangers with making egalitarian objectives overly specific. For example, one would not want to focus exclusively on reducing social inequality in the uptake of bowel cancer screening services without setting this in the context of more important and more general objectives such as reducing social inequality in bowel cancer mortality and life expectancy.

Economic Evaluation in Public Health

The previous section reviewed potential normative economic justifications for government to 'do something' in public health, rather than leave things to the market. However, these theoretical arguments only go part of the way toward justifying particular public health interventions. Government intervention in public health often imposes costs on public budgets, taxpayers and/or businesses, thus requiring an investment of scarce resources, which could be used for other potentially beneficial purposes. To justify this investment, evidence and analysis are needed to show that the particular government intervention under consideration represents good 'value for money' compared with alternative uses of scarce resources. This is what the economic evaluation of public health interventions seeks to establish. Undertaking such economic evaluations – while being time- and resource-intensive activities in themselves – is useful for at least three reasons:

- Improving public policy outcomes: Economic evaluation can help improve outcomes by helping policymakers identify potentially worthwhile and potentially wasteful public health interventions based on the best available international research evidence.
- Improving clarity of thought: Economic evaluation can help public policymakers think through systematically the pros and cons of alternative ways of designing and implementing public health interventions in their own decision-making context.
- Improving public accountability: Economic evaluation can help hold public policymakers to account by identifying and publishing the factual assumptions and social value judgments underpinning their decisions.

The economic way of thinking about the costs and benefits of public health interventions can be contrasted with two commonly held but misguided alternative ways of thinking. In his classic health economic monograph, 'Who Shall Live?,' Victor Fuchs memorably dubbed these the 'romantic' and 'monotechnic' points of view, respectively. The 'romantic' point of view denies that resources are scarce and that resource allocation decisions have opportunity costs in terms of alternative beneficial uses of scarce resources. The 'romantic' believes that resources can be found for their own favoured cause without impinging on other people's favored causes – for example, by making 'efficiency savings,' by diverting resources from disfavored causes (such as defense spending) or by clamping down on the high pay and tax avoidance behavior of the super rich. Fuchs criticizes this viewpoint, writing that: "Because some of the barriers to greater output and want satisfaction are clearly man-made, the romantic is misled into confusing the real world with the Garden of Eden." He goes on: "Confronted with an obvious imbalance between people's desires and the available resources, the romantic-authoritarian response may be to categorize some desires as 'unnecessary' or 'inappropriate', thus protecting the illusion that no scarcity exists." By contrast, the 'monotechnic' point of view fails to recognize the legitimate plurality of individual and social objectives. The 'monotechnic' fixates on a single objective and is unconcerned if allocating additional resources to this objective imposes opportunity costs in terms of other objectives.

According to Fuchs, the 'monotechnic' view is "frequently found among physicians, engineers, and others trained in the application of a particular technology." He goes on to write: "The desire of the engineer to build the best bridge or the physician to practice in the best-equipped hospital is understandable. But to extent that the monotechnic person fails to recognize the claims of competing wants or the divergence of his priorities from those of other people, his advice is likely to be a poor guide to social policy."

Various organizations have adopted a systematic cost-effectiveness approach to evaluating public health interventions, in line with standard health technology assessment methods being used to evaluate clinical healthcare sector interventions. Publicly accessible repositories of this kind of evidence, each of which assesses the cost-effectiveness of a fairly wide range of public health interventions using a common set of methods, include the WHO-CHOICE database, the US Preventive Services Task Force, the UK National Institute for Health and Clinical Excellence public health guidance, and the ACE-Prevention project in Australia.

However, standard cost-effectiveness analyses of this kind are somewhat less useful in public health than in the healthcare sector, for at least four reasons. First, randomized control trial (RCT) data are scarce, so it is hard to attribute effects to interventions, and the exploitation of 'natural experiments' using large observational datasets is still in its infancy in public health. This means that existing repositories of cost-effectiveness evidence in public health are forced to chart a difficult course between the Scylla of 'RCT fetishism' (i.e., focusing unduly on clinically-oriented types of intervention for which RCT data exist) and the Charybdis of 'practitioner bias' (i.e., using overly favorable effect size estimates based on the opinions of a small coterie of policy enthusiasts rather than robust evidence). Second, important costs and benefits often fall outside the healthcare sector – including costs on taxpayers, business and government agencies, and including a variety of nonhealth benefits such as improvements in education, employment, and crime outcomes. This means that cost-effectiveness analyses focusing on health benefits and healthcare sector costs only may not be relevant to the most important decision makers and stakeholders. Third, public health interventions often have explicit policy objectives relating to inequality reduction. Standard cost-effectiveness analysis does not examine the distribution of costs and benefits, and hence cannot offer policymakers any guidance on the existence and nature of potential trade-offs between concerns for efficiency and equality. Finally, some public health interventions have long-term benefits that arise decades in the future – including benefits to future generations. Standard cost-effectiveness analysis does not explicitly distinguish effects on current and future generations, and the standard approach to discounting implies a hefty penalty to health benefits arising many decades in the future – for example, a 5% discount rate implies that a life-year gained in 50 years time is valued at only 7.7% of a life-year gained this year; or 0.6 of 1% in 100 years time.

There are public repositories of cost–benefit analysis evidence of social policies, which take a broader approach and address some (but not all) of these issues – for example, the Washington State Institute of Public Policy. To date, however, cost-benefit analyses of social policies tend to focus on nonhealth benefits; and if health effects are incorporated at all, they tend to be based on mortality and the saving of 'statistical lives' rather than more comprehensive analysis of effects on length of life and health-related quality of life.

As Helen Weatherly and colleagues from the University of York have argued therefore, more research is needed to produce more useful economic evaluations of intersectoral public health policies, including not only the application of existing cost–consequence analysis and cost–benefit analysis approaches, but also methodological research to develop new approaches.

Conclusion

In line with most of the existing economic literature on public health, this article has adopted a standard 'social choice' approach to normative economics, which focuses on providing analysis and evidence that will be useful to a perfectly benevolent social decision-making institution seeking to achieve a set of socially desirable objectives in the face of market failure. However, it is also possible to adopt a 'public choice' approach that treats government institutions as economic agents with nonbenevolent or at least imperfectly benevolent objectives. Economic models of interest group lobbying, rent seeking, and bureaucratic incentives, can all help to understand government behavior in relation to public health, why some public health interventions are more likely to be adopted than others, and why actual decision-making in public health so often departs from policy prescriptions based on standard 'social choice' analyses of the kind described in this overview. For an excellent overview of government failure in public health, and the potential for future research in this hitherto neglected area, see the article in this encyclopedia by Hauck and Smith on 'public choice analysis of public health priority setting'.

Another important frontier in public health research is the role of behavioral economic evidence and insights in helping to design more effective public health interventions. As described earlier, global economic growth means that the task of public health is increasingly shifting away from preventing the 'infectious diseases of childhood' toward preventing the 'chronic diseases of adulthood.' This implies a shift in the nature of the economic problem away from market failures due to infectious disease externality, toward market failures due to bounded rationality that generates unhealthy lifestyle behavior. There is thus an important new role for behavioral economic research into the nature of bounded rationality and the potential role of interventions to improve lifestyle behavior through appropriate 'nudges' and 'budges'.

Finally, a third important frontier for economic research is the economic evaluation of cross-sectoral public health interventions. Compared to the cost-effectiveness analysis of healthcare technologies, the evaluation of public health interventions poses additional – or, rather, more severe – challenges that require the development of new methods. These methodological challenges include (1) estimating health effects when RCT evidence is scarce, (2) measuring and valuing nonhealth benefits alongside health benefits, (3) analyzing costs falling outside the government healthcare budget, (4) analyzing

distributional concerns when reducing inequality is an explicit policy objective, and (5) valuing long-term health and non-health benefits including benefits to future generations.

See also: Economic Evaluation of Public Health Interventions: Methodological Challenges. Ethics and Social Value Judgments in Public Health. Fetal Origins of Lifetime Health. Global Public Goods and Health. Health and Its Value: Overview. Health Econometrics: Overview. Infectious Disease Externalities. Pay for Prevention. Preschool Education Programs. Priority Setting in Public Health. Public Choice Analysis of Public Health Priority Setting. Public Health in Resource Poor Settings. Public Health Profession. Unfair Health Inequality

Further Reading

Akinson, A. B. (2011). The restoration of welfare economics. *American Economic Review* **101**(3), 157–161.

Browning, E. K. (1999). The myth of fiscal externalities. *Public Finance Review* **27**, 3–18.

Carande-Kulis, V. G., Getzen, T. E. and Thacker, S. B. (2007). Public goods and externalities: A research agenda for public health economics. *Journal of Public Health Management Practice* **13**(2), 32–227.

Cawley, J. and Ruhm, C. J. (2011). Chapter three – The economics of risky health behaviors. *Handbook of health economics* **2**, 95–199. Available at: http://www.nber.org/papers/w17081 (accessed 03.10.13).

Colgrove, J. (2002). The McKeown thesis: A historical controversy and its enduring influence. *American Journal of Public Health* **92**(5), 725–729.

Greenwald, B. and Stiglitz, J. E. (1986). Externalities in economies with imperfect information and incomplete markets. *Quarterly Journal of Economics* **101**, 229–264.

Kenkel, D. and Suhrcke, M. (2011). *Economic evaluation of the social determinants of health – A conceptual and practical overview*. World Health Organization. Regional Office for Europe. Available at: http://www.euro.who.int/__data/assets/pdf_file/0005/155579/e96075.pdf (accessed 03.10.13).

Loewenstein, G., Asch, D. A., Friedman, J. Y., Melichar, L. A. and Volpp, K. G. (2012). Can behavioural economics make us healthier? *British Medical Journal* 344. doi: ?http://dx.doi.org/10.1136/bmj.e3482.

Marteau, T. M., Ogilvie, D., Roland, M., Suhrcke, M. and Kelly, M. P. (2011). Judging nudging: Can nudging improve population health? *British Medical Journal* **342**, d228, doi:10.1136/bmj.d228.

Philipson, T. J. (2008). Economic epidemiology. In Durlauf, Steven N. and Blume, Lawrence E. (eds.) *The New Palgrave dictionary of economics online*, 2nd ed. Palgrave Macmillan.

Porter, D. (1998). *Health, civilization, and the state: A history of public health from ancient to modern times*. New York: Routledge.

Rose, G. (1992). *The strategy of preventive medicine*. New York: Oxford University Press.

Sretzer, S. (1988). The importance of social interventions in Britain's mortality decline c. 1850–1914: A re-interpretation of the role of public health. *Society for the Social History of Medicine* **1**, 1–37. Available at: http://shm.oxfordjournals.org/content/1/1/1.full.pdf (accessed 03.10.13).

Vining, A. and Weimer, D. L. (2010). An assessment of important issues concerning the application of benefit-cost analysis to social policy. *Journal of Benefit-Cost Analysis* **1**(1), 1–40. Available at: http://www.bepress.com/jbca/vol1/iss1/6 (accessed 03.10.13).

Weatherly, H., Drummond, M., Claxton, K., et al. (2009). Methods for assessing the cost-effectiveness of public health interventions: Key challenges and recommendations. *Health Policy* **93**, 85–92.

Relevant Websites

Definition of public health

http://publichealth.jbpub.com/turnock/3e/sample_chapters.cfm

http://www.geaugacountyhealth.org/whatisph.html

http://www.healthknowledge.org.uk/public-health-textbook
HealthKnowledge Free On-Line Public Health Textbook.

http://www.sagepub.com/upm-data/3989_Chapter_1.pdf
Population-Based Public Health Practice.

http://www.fph.org.uk/what_is_public_health
UK Faculty of Public Health Definition of Public Health.

http://www.whatispublichealth.org/what/index.html
US Association of Schools of Public Health.

http://en.wikipedia.org/wiki/Public_health
Wikipedia Entry on Public Health.

http://en.wikipedia.org/wiki/Public_health_law
Wikipedia Entry on Public Health Law.

History of public health

http://www.academicearth.org/lectures/the-sanitary-movement-and-the-filth-theory-of-disease
Frank Snowden Yale Lecture on The Sanitary Movement.

http://www.who.int/about/history/en/index.html
History of WHO.

http://www.who.int/healthpromotion/conferences/previous/ottawa/en/
WHO 1986 Ottowa Charter for Health Promotion.

http://www.who.int/healthpromotion/conferences/previous/adelaide/en/index1.html
WHO 1988 Adelaide Recommendations on Healthy Public Policy.

Public health data and evidence

http://www.sph.uq.edu.au/bodce-ace-prevention
Australian ACE-Prevention Project.

http://www.gapminder.org/
Gapminder – The Wealth and Health of Nations.

http://www.healthmetricsandevaluation.org/gbd/visualizations/country
Global Burden of Disease (GBD) Visualizations.

http://guidance.nice.org.uk/PHG
NICE public health guidance in England.

http://www.oecd.org/els/health-systems/theeconomicsofprevention.htm
OECD Economics of Prevention.

http://benefitcostanalysis.org/resources/benefit-cost-analysis-resources
Society for Benefit Cost Analysis.

http://www.instituteofhealthequity.org/
UCL Institute of Health Equity.

https://www.gov.uk/government/publications/applying-behavioural-insight-to-health-behavioural-insights-team-paper
UK Cabinet Office Behavioural Insights Team. Applying Behavioural Insight to Health.

http://www.thecommunityguide.org/index.html
US Preventive Services Task Force Community Guide – Guide to Community Preventive Services.

http://www.wsipp.wa.gov
Washington State Institute for Public Policy.

http://www.who.int/choice/en/
WHO-CHOICE.

http://www.who.int/social_determinants/en/
WHO Social Determinants of Health.

Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation

I Shemilt, Campbell and Cochrane Economic Methods Group, and University of Cambridge, Cambridge, UK
E Wilson, Campbell and Cochrane Economic Methods Group, and University of East Anglia Norwich Research Park, Norwich, Norfolk, UK
L Vale, Campbell and Cochrane Economic Methods Group, and Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK

Glossary

Checklists A means of assessing the quality of a study, which incorporates the extent to which existing guidelines are followed, whereby a set of criteria used to assess the quality of a study are listed and possibly prioritized.

Cohort models Are used to estimate expected costs and benefits based on the average experience of a cohort of identical patients.

Decision trees Represent a systematic approach to describe a problem faced by decision making under conditions of uncertainty. They can be used to determine the costs and outcomes for a hypothetical patient cohort with the disease of interest when they are treated with different clinical strategies.

External consistency Requires that the design and structure of the model makes sense to experts in the field and that the results make intuitive sense as well. A further aspect is consistency of the results compared with other 'independent' data.

Grading of recommendations assessment, development, and evaluation An informal collaboration of people with an interest in addressing the shortcomings of quality assessment tools of research in health care. It has developed a tool used to grade both the quality of evidence and the strength of recommendations (see www.gradeworkinggroup.org/).

Individual sampling models Allow the behavior of individuals within a cohort to be tracked separately. Such models are computationally more demanding than cohort models but allow greater variability in cost and effects as an individual's journey through the events in a model is described rather than an average journey. Such models facilitate the modeling of situations where an individual's future costs and effects depend on that individual's history.

Internal consistency Requires that the mathematical logic of the model is consistent with the model specification and that data have been incorporated correctly, i.e., there are no errors in programming.

Markov chains Are used within Markov models to describe how the health states of individuals change over time. Within this article the terms have been used synonymously with the term Markov model.

Markov models Patients with a specific clinical problem can exist in a finite set of health states (that is: alive in perfect health, alive with a deficit in HRQOL, or dead) between which they can move over time. Movement between these health states occurs during a discrete time interval, usually yearly (known as a Markov cycle) based on preset transition probabilities. By attaching resource costs and health outcome consequences to each Markov state (which may vary based on use of an intervention), it is possible to test how such an intervention might influence clinical outcomes and resource use (on average) for a group of patients with a clinical problem.

Model structure Describes the underlying disease process and service pathways under consideration (see also structural uncertainty).

Parameter uncertainty Relates to uncertainty surrounding the value that a probability, a cost or a utility may take. It might be caused by the statistical imprecision (stochastic uncertainty) or from the existence of multiple conflicting sources of data, the internal or external validity of the data. It can be handled within a model by manually varying the parameter of interest in a deterministic analysis or by sampling from an *a priori* defined distribution in probabilistic sensitivity analysis.

Patient-level simulations See individual sampling models.

Structural uncertainty (also called model uncertainty) Includes uncertainty regarding what comparators should be included in the model; whose costs and benefits are important and how uncertainty around both the disease mechanism and pathways of care might impinge on the design of the model. It also covers uncertainty over how long a treatment effect might persist and how it might best be statistically modeled.

Transmission dynamic models Are a class of more sophisticated individual sampling models that do not assume independence between the individuals modeled. Such models can be useful to model infectious disease where the higher risk of infection is a function of prevalence of disease in the population surrounding the individual which may vary over time and, for example, where an increase in infection rate leads to an increase in risks of further infection and reduce the number of people susceptible.

 doi:10.1016/B978-0-12-375678-7.01405-X

Introduction

Economic modeling techniques are widely used to provide a quantitative framework for economic evaluations that aim to inform policy decisions. Central to the validity of judgments that are based on the results of economic models is an assessment of the quality of the models themselves. Decision makers should have confidence that the quality of the models they are using is sufficiently robust to justify their decisions, and researchers developing models should demonstrate that their work meets acceptable quality standards.

Although assessment of the quality of models can be argued to be important from several different perspectives (in this article the main arguments are summarized), it is also necessary to consider what factors determine the quality of models and how quality might be assessed. To address these issues several approaches are considered. The article considers structure, data, and consistency as suggested by Philips *et al.* (2004).

In general, the quality of models depends on:

- Its fitness for purpose: relevance to the underlying research question;
- The methods by which data inputs to the model are combined (both of which relate to the quality of the model structure);
- The quality of the source studies from which data are taken; and
- The model's internal and external consistency.

The quality of reporting of methods and results of models are also considered. Finally, key implications for both current practice and further research are highlighted.

Factors Determining Model Quality

Fitness for Purpose

Although there is no 'right' answer and analysts have to exercise judgment, it is important that the choices made by the analyst be explained. This includes the choice of the overall modeling approach. Models developed for economic evaluations can be broadly grouped into cohort models (e.g., decision trees and Markov chains) and individual sampling models (also known variously as patient-level simulations) (Briggs *et al.*, 2006).

The choice of appropriate modeling approach depends on the research question posed. For example, with a treatment for an acute condition where the question is whether the intervention is 'effective' or 'not effective' a decision tree model may be sufficient, whereas chronic diseases, especially those characterized by periods of relapse and remission, can be modeled with Markov chains. When individual risks are contingent on previous events or pathways, individual sampling models may be more appropriate. Similarly, the evaluation of vaccination and screening programs for infectious diseases may be most appropriately modeled with more sophisticated individual sampling models ('transmission dynamic models') that take account of the changing risk of infection as the prevalence of disease changes following the introduction of an intervention as the population begins to gain the advantages of herd immunity. The modeling approach and structure must, therefore, reflect the research question, the properties of the evaluated technologies, the characteristics of the disease, and the treatment/intervention setting (Institute for Quality and Efficiency in Health care (IQWiG), 2009).

Structure

What is a 'good' structure? One answer is plausibility: Whether or not the patient pathways and assumptions represented by a model's structure are plausible. All models by definition simplify reality. The task of the analyst is to design one that is sufficiently complex to reflect the nature and subtleties of the pathway being modeled yet simple enough to be (1) efficiently computed; (2) understood by the intended audience; and (3) capable of generating the necessary information for credible and authoritative guidance.

The degree of simplification and compromise is a matter of judgment, and hence there is no unique model structure that could be considered objectively 'correct.' This introduces a type of uncertainty within decision modeling termed structural uncertainty (also called model uncertainty) (Briggs *et al.*, 2012). This is to be distinguished from parameter uncertainty. Aspects of structural uncertainty include the overall modeling approach; choice of comparator(s); scope; duration of treatment effect; the events handled by the model; and any statistical model used to estimate parameters, clinical uncertainty about the mechanism or pathway of care, and the absence of clinical (and other) evidence (Bojke *et al.*, 2006).

The most common approach to handling structural uncertainties like the duration of treatment effect and the scope of the analysis is through the use of scenarios (Bojke *et al.*, 2006). A 'base case' is commonly recommended, with alternative possible scenarios presented for the decision maker to judge the fit with their setting. Investigating the impact of more fundamental structural choices, such as alternative overall designs of the model or selection of different events such as wider or narrower scopes of costs and benefits may sometimes be possible: Alternative models could be constructed and tested to see if they yielded, or were likely to yield, different results. They could also be formally combined using a variant of Bayesian model averaging (Bojke *et al.*, 2006). As there is a very large number of plausible potential model variants, such attempts will always have to be restricted to a reasonable subset of possible models.

Data Inputs to a Model

The quality of data used in model development can impact directly on the reliability of results. One way in which this can occur is through the impact of data quality on a model's parameters. Each probability, cost, and outcome in an economic model is expressed in terms of a set of measurable, quantifiable characteristics, or parameters. In economic models, common types of parameters are probabilities, costs, relative treatment effects, and utilities (Briggs *et al.*, 2006).

Components of data needed to assign values to these parameters are summarized in Box 1 below.

These data are collected or derived from sources that may include empirical research studies, routine administrative databases, reference sources, and expert opinion. In principle, and often in practice, there is more than one potential source for each data component. For this reason, and because the results of economic models may depend in large part on choices between available sources of data, the processes of data identification, appraisal, selection, and use need to be explained and justified.

However, the use of data in model development is not restricted to the assigning of values to parameters. Data are also used to support every stage of model development, from establishing a conceptual understanding of the decision problem as noted above, through to the choice of sensitivity and uncertainty analysis (Briggs *et al.*, 2006; Paisley, 2010). Thus, economic models have multiple information needs, requiring different types of data, drawn from various sources, and the factors that determine data quality encompass all of these uses, types, and sources.

Philips *et al.* (2004) identify four dimensions of quality regarding data: identification, modeling, incorporation, and assessment of uncertainty. Each dimension has a corresponding set of attributes of good practice, or factors that determine quality, and each attribute refers to the quality of processes used in the identification, appraisal, selection, or use of data at the different stages of the model development process.

Data quality is only one of two criteria to be used in identifying and selecting data in economic models. Before quality assessment, the available data also need to be assessed in terms of applicability or relevance. The initial assessment of applicability may result in a large proportion of potential data sources being rejected (Kaltenthaler *et al.*, 2011).

Consistency

Assessing the consistency, or as some agencies have termed it validity (Canadian Agency for Drugs and Technologies in Health, 2006; Institute for Quality and Efficiency in Health care [IQWiG], 2009), of a model is a subject of debate and there is no unanimous agreement among experts (Philips *et al.*, 2004; Institute for Quality and Efficiency in Health care [IQWiG], 2009 and Canadian Agency for Drugs and Technologies in Health, 2006 all offer different views). Regardless of how it is assessed it is useful to briefly consider what is meant by consistency. Philips *et al.* (2004) identified four aspects: internal consistency; external consistency; between model consistency; and predictive validity. Each of these has been briefly summarized.

Internal consistency

Internal consistency requires that the mathematical logic of the model is consistent with the model specification and that data have been incorporated correctly, for example, that there are no errors in programming. One source of inconsistency in model specification relates to the conditioning of an action on an unobservable event. For example, in a model of cancer surveillance, decisions to change treatment may be made as soon as progression or recurrence of the cancer occurs despite the fact that in reality recurrence or progression would not be observed until some monitoring test has been performed.

Internal consistency may also be affected by asymmetries within the model. Such asymmetries may or may not represent errors but probably serve to highlight areas for further investigation. An example is when the physiological response of a patient to a particular event that occurs at several different points within a treatment pathway is modeled differently in corresponding areas of the model. It goes without saying that internal consistency will also be affected by the accuracy of the model programming and data entry. Errors in the model syntax or administrative errors (e.g., in labeling of data) are, at least initially, inevitable in the design and execution of any model. Proof reading and testing element by element are, therefore, a critical part of the design process. Ideally, these tasks would be completed by a researcher who is familiar with the decision question but who was not involved in the design and execution of the model and hence is less subject to bias and to uncritical acceptance of the often implicit assumptions that the original analysts may have made.

Other checks for internal consistency are to test formulae and equations separately before they are entered into the model to ensure that they are correctly expressed and can give the anticipated results. Sensitivity analyses using extreme or zero values can be used to identify apparently counterintuitive results. Likewise, examining the results of known scenarios, even ones that are unlikely or cannot occur in practice, can also be useful to identify counterintuitive results. The presence or absence of counterintuitive results does not necessarily indicate a problem (or lack of one) with internal consistency. However, when counterintuitive results are identified they should be explained, which will involve unpicking and examining the mechanisms that have led to them. A more elaborate test of internal consistency is to attempt to replicate the model in another software package and then comparing the results of both.

Box 1 Components of parameter data (Kaltenthaler *et al.*, 2011; Coyle *et al.*, 2010)

- clinical effect sizes (i.e., relative treatment effects for beneficial and adverse effects)
- disease natural history or epidemiology
- resource use or service utilization
- unit costs
- health state utilities
- survival
- other time to event data
- compliance or participation patterns
- relationships between intermediate and final endpoints

External consistency

External consistency requires that the design and structure of the model makes sense to other experts in the field (a process that can be aided by presenting the model in pictorial form (Canadian Agency for Drugs and Technologies in Health, 2006)) and also that the results make intuitive sense. This is particularly the case where the results of the model run counter to expectations, and there may be temptation to disbelieve them. Given that the model is internally valid, it should then be structured such that a plausible narrative can be extracted explaining the logic of the results and why, if it departs from prior expectations, it does so.

Such face validity is only one aspect of external consistency. A further aspect is consistency of the results compared with other 'independent' data. However, if such independent data exist it would be more appropriate for them to be included in the model in the first place. This issue is not addressed in methods guides produced by some national HTA bodies (e.g., Institute for Quality and Efficiency in Health Care (IQWiG), 2009 and Canadian Agency for Drugs and Technologies in Health, 2006).

Between model consistency

This type of consistency is sometimes also included as an aspect of external consistency. It relates to a comparison of a model's results with those of other models. The models to be compared should be developed independently but, because models might be developed at different times and in subtly different contexts, interpretation should proceed with caution as results may legitimately differ so that similarity of result does not necessarily confer confidence in the robustness or validity of either, nor does dissimilarity necessarily undermine confidence. Furthermore, even where two models are developed for the same purpose and at the same time, convergence of results might not occur because each model is the product of myriad judgments and assumptions, which might differ between research teams. This occurs in the multiple technology appraisals process conducted for National Institute for Health and Clinical Excellence (NICE) in England. Different stakeholders (usually manufacturers) can submit their own models as evidence to NICE. These models may also be used to inform the development of a further model by an independent academic group. The stakeholders will each have access to different data (as commercially sensitive data will not be shared with other manufacturers) with only the independent academic model having access to all data. Taken as a body these models can all be thought of as providing mutual sensitivity analyses enabling one to identify critical assumptions, important parameters, and the sensitive ranges for those parameters.

Predictive validity

A final form of consistency identified by Philips *et al.* (2004) is predictive validity. They noted that some commentators argue that the predictions of a model should be compared with the results of a predictive study. On the contrary, they argue that it is inappropriate to expect a model to predict the future with such accuracy because it can only use the data and understanding that was available at the time it was constructed.

The reason for constructing a model in the first place is the lack of appropriate data from a single source with which to make a decision, so it inevitably represents a synthesis of current knowledge. Once such an appropriate prospective study has been undertaken, it should surely replace the model. However, such a comprehensive single study is most unlikely to exist in practice. Furthermore, the interventions under evaluation may be implemented under different conditions to those assumed in the model because of factors such as technological change. It is notable that consideration of such an element of consistency is not included in some of the more recent methods guides from Health Technology Assessment Agencies (National Institute for Health and Clinical Excellence, 2008; Institute for Quality and Efficiency in Health care (IQWiG), 2009 and Canadian Agency for Drugs and Technologies in Health, 2006).

Assessing the Quality of Models

Structure

Several tools and checklists have been employed to assess the quality of economic evaluations in general, and of decision models in particular. The majority cover different aspects of the structure, data, and consistency of the model. Philips *et al.* (2004) proposed a framework by which models could be assessed. In common with other quality assessment tools, this framework comprises a series of criteria that capture methodological dimensions which analysts could be expected to have addressed in the conduct and reporting of their evaluations. These can be assessed with specific yes/no responses accompanied by supporting commentary. The nine structural dimensions identified by the Philips checklist are summarized in **Box 2**.

A clear statement of the decision problem is essential to define the entire evaluation: A vague question can either lead to a vague answer or one that fails to address precisely the decision problem at hand. Therefore, the objective of the analysis should be stated, along with a statement of who the primary decision maker is.

Philips *et al.* (2004) suggested that the scope (analytic perspective) of a decision model, which crucially affects which costs and outcomes are included in the analysis, should be stated and justified. Many decision-making organizations have adopted specific perspectives for their reference cases.

Box 2 Structural dimensions in the Philips *et al.* Checklist

Statement of the decision problem/objective
Statement of scope/perspective
Rationale for structure
Structural assumptions
Strategies/comparators
Model type
Time horizon
Disease states and pathways
Cycle length

The structure of the model should reflect both the underlying disease process and service pathways under consideration. As noted above, the model type should be 'appropriate' to the decision question. Similarly, the time horizon of a model should be 'appropriate,' i.e., sufficient to capture all important differences in costs and outcomes between the options.

As again noted above, the model should have a degree of face validity in that its structure should make intuitive sense.

In principle, every 'feasible and practical' alternative treatment strategy should be considered rather than only a subset of comparators. This is because cost-effectiveness is always a relative concept (i.e., a treatment is considered cost-effective relative to another treatment). Therefore, exclusion of relevant comparators may lead to erroneous conclusions being drawn about which intervention(s) to invest in or disinvest from. Current practice should be included among comparators.

Finally, as with the choice of model structure, the disease states/pathways included and cycle length in a decision model with discrete time intervals should ideally reflect the underlying biology of a disease as well as the impacts of interventions.

Data

Most published guidelines on assessment of quality of data in economic models focus on the transparency of reporting of methods and results and on the quality of methods to identify data to populate model parameters. They do not assess methodological quality *per se*, nor do they consider the wider uses of data in the model development process (Kaltenthaler *et al.*, 2011). A key reason for the general lack of consensus on standards for quality assessment with respect to data is that the scope of data relevant to an economic model is not entirely predefined, but emerges in the course of the iterative model development process. As such there is no objectively 'right' or 'wrong' set of data for use to inform the model development process, but rather an interpretation of what data is relevant to the decision problem at hand (Kaltenthaler *et al.*, 2011).

The process of identifying data to inform model development is necessarily an iterative, emergent, and nonlinear process rather than a series of discrete information retrieval activities, such that the searches conducted and the sequence in which they are conducted will legitimately differ between models (Kaltenthaler *et al.*, 2011). Although this process can still be systematic and explicit, it is also highly specific to each individual model. Given these issues, recently published guidelines (Kaltenthaler *et al.*, 2011) have recommended approaches to data identification that focus on:

- Maximizing the rate of return of potentially relevant data;
- Judging when to stop searching because 'sufficient' data have been identified, such that further efforts to identify additional relevant data would be unlikely to improve the analysis; and
- Prioritizing key information needs and paying particular attention to identifying data for those parameters to which the results of a model are particularly sensitive. (There is a potential Catch 22 here in that it is unclear *a priori* which data points the model will be sensitive too. In practice, the analyst relies on experience from other similar models, although a modeling solution is possible in that running a model with dummy data can inform this as well.)

Coyle *et al.* (2010) proposed a hierarchy to rank the quality of sources of the different components of data used to populate model parameters. This hierarchy highlights variation between different components of data in which sources may be regarded as 'high quality,' emphasizing sources that in principle generate causal inferences with high internal validity for clinical effect-sizes, and sources that are in principle highly applicable to the specific decision problem at hand for all other components.

Although such hierarchies can, with refinement, offer a useful tool for the quality assessment of data sources, they do not incorporate assessment of specific dimensions of quality, or risk of bias, within each source (Kaltenthaler *et al.*, 2011). Various instruments and checklists have been developed for assessing risk of bias in both randomized controlled trials and nonrandomized studies of effects that may be used to populate parameters in economic models. Perhaps, the most prominent is the Cochrane Risk-of-Bias tool (Higgins *et al.*, 2011). However, these instruments are not equally applicable to all the diverse potential sources of data for model development. The grading of recommendations assessment, development and evaluation (GRADE) system offers more promise in this respect. It provides a consistent framework and a set of criteria for rating the quality of evidence collected (or derived) from all potential sources for all data for populating model parameters. Sources include both research- and nonresearch-based sources (e.g., national disease registers, claims, prescriptions or hospital activity databases, or standard reference sources such as drug formularies or collected volumes of unit costs) (Brunetti *et al.*, 2013). Consistent with the hierarchy proposed by Coyle *et al.* (2010), the GRADE system allows flexibility in the quality assessment process to include additional considerations alongside internal validity. These include, as part of the 'indirectness' criterion in GRADE, the applicability of data components to the specific decision problem at hand.

Reporting Methods and Results

Lack of transparency in the reporting of methods and results of economic evaluations undermines their credibility and, in turn, threatens the appropriate use of evidence for cost-effectiveness in decision-making. The development and use of increasingly sophisticated economic modeling techniques may increase the complexity of economic models at the expense of transparency. It is, therefore, important for analysts to strike an appropriate balance between scientific rigor, complexity, and transparency, to reduce the 'black box' perception of economic models and provide decision-makers with an ability to understand intuitively 'what goes in' and 'what comes out.' This requires analysts to use and record formal, replicable approaches at each stage of the model development process and report these approaches in a transparent and reproducible way (Cooper *et al.*, 2007).

In the quality assessment checklist based on their 2004 review of published good practice guidelines in economic modeling, Phillips *et al.* found that 19 of 42 identified

attributes of good practice and 33 of 57 questions for critical appraisal related to the transparency of reporting and the explicitness of the justification of methods. These attributes and questions are concerned with transparency and justification of choices made at every stage of the model development process. The key principle underlying checklists for economic models is that decision-makers should be able to reach a judgment easily, based on information in the published report alone, on each of the following:

- Whether analysts have invested sufficient effort to identify an acceptable set of data to inform each stage of the model development process;
- That sources of data have not been identified serendipitously, opportunistically, or preferentially; and
- Whether any choices between alternative sources of data, assumptions (in the absence of data), or adaptations or extrapolations of existing data, are explained and justified with respect to the specific decision problem at hand (Kaltenthaler *et al.*, 2011).

Implications for Practice and Research

As economic models are increasingly used to inform policy and practice, concerns about whether or not their quality is 'good enough' for this purpose come to the fore. Assessing quality is important for the credibility of the whole process but it can be time consuming. Sufficient time must be planned for this work to be undertaken. Given the sometimes pressing time restrictions that modelers face, policy makers need to be aware that a full assessment of all aspects of quality may not be possible. Analysts in such circumstances have to be clear about what they have and have not done. They can help ensure that quality has been maximized within the constraints of the research process by referring to existing checklists described above for assessing the quality of models in general or of those designed for specific circumstances and stakeholders. Because no model can be perfect, any limitations should be explored and if they cannot be rectified or improved then the limitation should be highlighted in the study report along with an analysis of its consequences, the direction, and possible size of any bias and some cautions about external validity.

See also: Economic Evaluation, Uncertainty in. Problem Structuring for Health Economic Model Development. Searching and Reviewing Nonclinical Evidence for Economic Evaluation. Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies. Synthesizing Clinical Evidence for Economic Evaluation

References

Bojke L., Claxton K., Palmer S., Sculpher M. (2000). Defining and characterising structural uncertainty in decision analytic models. *CHE Research Paper 9*. York: Centre for Health Economics, University of York.

Briggs, A., Sculpher, M. and Claxton, K. (2006). Chapter 4: Making decision models probabilistic. In Briggs, A., Sculpher, M. and Claxton, K. (eds.) *Decision modelling for health economic evaluation*, pp. 77–120. Oxford: Oxford University Press.

Briggs, A., Weinstein, M., Fenwick, E., et al. (2012). Model parameter estimation and uncertainty: A report of the ISPOR-SMDM modeling good research practices task force-6. *Value in Health* **15**, 835–842.

Brunetti, M., Shemilt, I., Pregno, S., et al. (2013). GRADE guidelines: 10. Special challenges – quality of evidence for resource use. *Journal of Clinical Epidemiology* **66**(2), 140–150.

Canadian Agency for Drugs and Technologies in Health (2006). *Guidelines for the evaluation of health technologies*, 3rd ed. Ottawa: Canadian Agency for Drugs and Technologies in Health.

Cooper, N. J., Sutton, A. J., Ades, A. E., Paisley, S. and Jones, D. R. (2007). Use of evidence in economic decision models: Practical issues and methodological challenges. *Health Economics* **16**(12), 1277–1286.

Coyle, D., Lee, K. M. and Cooper, N. J. (2010). Chapter 9: Use of evidence in decision models. In Shemilt, I., Mugford, M., Vale, L., Marsh, K. and Donaldson, C. (eds.) *Evidence-based decisions and economics: health care, social welfare, education and criminal justice*, pp. 106–113. Oxford: Wiley-Blackwell.

Higgins, J. P. T., Altman, D. G. and Sterne, J. A. C. (eds.) (2011). Chapter 8: Assessing risk of bias in included studies. In Higgins, J. P. T. and Green, S. (eds.) *Cochrane handbook for systematic reviews of interventions*, pp. 188–242. Version 5.1.0. The Cochrane Collaboration. Available at: www.cochrane-handbook.org (accessed 28.03.13).

Institute for Quality and Efficiency in Health care (IQWiG). (2009). Working paper Modelling. Version 1.0. Cologne: Institute for Quality and Efficiency in Health care.

Kaltenthaler E., Tappenden P., Paisley S., Squires H. (2011). NICE DSU Technical Support Document 13: Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models. Sheffield: NICE Decision Support Unit, School of Health and Related Research. Available at: http://www.nicedsu.org.uk (accessed 28.03.13).

National Institute for Health and Clinical Excellence (2008). *Guide to the methods of technology appraisal*. London: National Institute for Health and Clinical Excellence.

Paisley, S. (2010). Classification of evidence used in decision-analytic models of cost-effectiveness: A content analysis of published reports. *International Journal of Technology Assessment in Health Care* **26**(4), 458–462.

Philips, Z., Ginnelly, L., Sculpher, M., et al. (2004). Review of guidelines for good practice in decision-analytic modeling in health technology assessment. *Health Technology Assessment* **8**(36), 1–172.

Further Reading

Deeks, J. J., Dinnes, J., D'Amico, R., et al. (2003). Evaluating non-randomized intervention studies. *Health Technology Assessment* **7**, 27.

National Institute for Health and Clinical Excellence (2009). *Methods for the development of NICE public health guidance*, 2nd ed. London: National Institute for Health and Clinical Excellence.

Philips, Z., Bojke, L., Sculpher, M., Claxton, K. and Golder, S. (2006). Good practice guidelines for decision-analytic modelling in health technology assessment: A review and consolidation of quality assessment. *PharmacoEconomics* **24**(4), 355–371.

Relevant Websites

http://c-cemg.org/
Campbell and Cochrane Economic Methods Group.

www.cochrane-handbook.org
Cochrane Handbook for Systematic Reviews of Interventions.

www.gradeworkinggroup.org
Grading of Recommendations Assessment, Development and Evaluation (short GRADE) Working Group.

www.nicedsu.org.uk
NICE Decision Support Unit.

www.ispor.org
The International Society for Pharmacoeconomics and Outcomes Research.

Quality Reporting and Demand

JT Kolstad, University of Pennsylvania, Philadelphia, PA, USA, and National Bureau of Economic Research, Cambridge, MA, USA

Introduction

The defining feature of health care markets and the economics of the health care sector is information structure. Kenneth Arrow, in his seminal paper, demonstrates the role that 'missing markets' for information play in explaining the existence of the features of health care that distinguish it from other industries and markets (Arrow, 1963). Information structure can explain not-for-profit firms, widespread insurance coverage, and the role of physician agents who both proffer advice on treatments and sell those same services, amongst other unique features. The importance of these missing markets become clear if one considers a 'simple' market for health care services. At almost every turn a consumer faces a substantial if not (at least privately) insurmountable level of uncertainty in decision making.

Picking between health insurance plans, in principal, requires a clear sense of the value of a particular plan in the event that a person becomes ill with a specific disease. What hospitals and doctors would be available? At what cost? Even with all of this information in hand the choice of plan requires that each eventuality, and the associated care available, be weighed taking into account the probability that a particular disease occurs. Once equipped with coverage, the individual must then choose a primary care physician. The hope is that this physician will provide preventive care to manage the totality of the patients health, skillfully diagnose the universe of possible ailments and, should the patient require more specialized care, recommend and support the choice of a specialist and the assessment of their treatment plan. A specific physician may differ in skill across each of these margins. As with insurance, the choice is made without knowing precisely which health issues even might occur and, therefore, which skills are most valuable. If the consumer then requires more intensive treatment, they must choose a specialist. To accomplish this they must, under the duress of illness, try to assess the quality of a particular specialist as well the efficacy of a given treatment approach.

This stylized depiction makes clear how health care decisions made by the consumer are potentially rife with information problems. In light of these informational challenges, in health care how do we think about the basic building block of economics: the demand curve? Perhaps more importantly, from an economists perspective, how do we think about welfare and market function in a market where demand is determined in this manner? (Congdon *et al.* (2011) discuss issues of decision making in an environment with informational constraints and decision makers with nonstandard preferences. They demonstrate the importance of these issues with respect to the theory of welfare and social choice as well as their potential role in public policy.)

Moving from theory to practice, information plays a key role in public policy. Many of the policy approaches to address the perceived quality and cost issues in the health care market rely on, either explicitly or implicitly, attempting to address the information asymmetries in the market. The best known and most studied of these efforts are the provision of information directly to consumers. This article focuses on the experience of using direct information provision to overcome market failures in the market for health care services. Specifically, the focus will be on the provision of quality information and its impact on demand.

The author begins by developing a simple, stylized model of supply and demand to demonstrate the role of information in market function in health care. Then some extensions are introduced to allow for insurance and uncertainty. With the simple economics of quality reporting and demand as a framework, the role of quality information and quality reporting in provider choice is then discussed. The choice of primary care physician is distinguished from specialists and hospital choice as they are distinct choice environments, each yielding unique market failures and potential for market based or policy solutions. Rather than review the complete literature on quality reporting and demand for specialists, thefocus is on the experience in the market for coronary artery bypass graft (CABG) surgery. This market is the oldest and most studied of the applications of quality reporting. The main issues and empirical conclusions can be drawn from the experience in this market. Then the experience of quality reporting in the CABG market is compared to similar efforts in education. This comparison demonstrates the similarities between the fields in the information structure, and its associated market failures, as well as the impact of quality reporting.

Quality Information and Quality of Care

Baseline Model

To frame the discussion of quality information and demand, the author begins by developing a stylized model of the market for healthcare services. Demand is determined by a set of consumers who choose healthcare providers based on their utility from a specific provider relative to another as well as the gain from getting care at all relative to foregoing care. These consumers care about the price they pay for care and the quality of care provided. Quality in this context can be multidimensional. Patients care about the clinical quality of care they receive (e.g., lower chance of getting an infection from hospital care or lower probability of mortality from bypass surgery) and how satisfied they are with their experience and nonhealth amenities (e.g., the comfort of their bed, the quality of food, or the cleanliness of the waiting room).

The supply of healthcare services consists of healthcare providers who determine the quality of care provided by making costly investments to enhance care delivery. For simplicity, assume these investments are a continuous, convex cost function, though the basic intuition holds in different contexts. In the standard model healthcare providers care only

 doi:10.1016/B978-0-12-375678-7.00809-9

about the profit they gain – the price per patient times the number of patients less the cost of supplying healthcare at a given quality level.

This stylized model of the doctor, patient relationship is the workhorse of health economics and adheres closely to the conventional market model that economists rely on in most markets. In this setup the quality and price of healthcare are determined where the supply of care equals demand for care. Importantly, for the present purpose, this model also provides insight into the level of quality selected by the doctor. How this level is determined has both normative and positive implications for the function of healthcare markets and the role of quality reporting in demand.

To see this, consider a provider choosing a level of quality. **Figure 1** presents this simple example graphically. In the standard model, the profit maximizing provider chooses his optimal quality based on the marginal cost of increasing quality of the good (i.e., time spent with the patient assessing their ailment or additional surgical nurses to support him during surgery) and marginal revenue from the improved quality. This point is q^* where the baseline marginal revenue curve ($P^*(q)$) equals marginal cost. As long as consumers can ascertain each provider's quality, they choose the provider that provides them the greatest gain in utility – quality less the price of care. The doctors, who observe their own and their competitor's quality as well as the response of the market to quality, then face a simple optimization; the doctors choose a level of quality investment such that the marginal cost of quality improvement is just equal to the marginal revenue associated with that quality improvement – the additional patients scaled by the profitability of those patients.

This baseline model provides clear positive and normative predictions for the quality and cost of healthcare. Because the competitive supply and demand conditions are met both the first and second welfare theorems hold and the quality and cost of care cannot be improved on without making some individuals worse off (Arrow, 1963). That is, the effort to maximize profit would yield productive and allocative efficiency in the market for healthcare services. Doctors would deploy resources to minimize the cost of supplying quality given the input prices to produce quality (e.g., the wages of additional surgical nurses). Furthermore, the level of healthcare quality supplied would reflect societies underlying preferences for healthcare quality, relative to other forms of consumption. (A number of papers have estimated these preferences and suggest a relatively high willingness to pay for health improvements (e.g., Cutler, 2003; Murphy and Topel, 2006). With diminishing marginal returns to consumption, this willingness to pay is also increasing in income (Hall and Jones, 2005). These estimates can explain much of the high spending in many developed countries as well the focus on quality of care and technological improvement. The author returns to this issue in considering whether demand with limited information is lower than what the true demand would be expected to be.) That is, the well documented quality issues in healthcare as well as the high cost and reliance on technologically intensive provision of healthcare, particularly in the USA, would not pose a public policy concern.

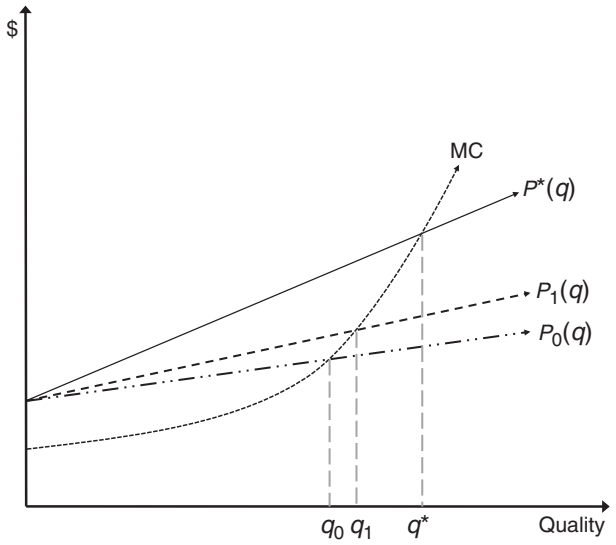


Figure 1 Equilibrium quality with and without quality information.

There are many reasons to doubt the market for healthcare services meets these stringent criteria for market function. One important and pervasive deviation is the presence of adverse selection – another market failure due to information asymmetries. Because consumers differ in profitability (difficulty to treat relative to a fixed payment), even if the respond to quality information providers have an incentive to distort their quality investments to attract relatively more profitable patients. Thus, even with full quality information the equilibrium may not be a first best (Rothschild and Stiglitz, 1976; Glazer and McGuire, 2005). (In practice, enhancing consumer choices can exacerbate adverse selection problems further. Handel (in press) documents this effect empirically in the market for health insurance. Alternatively, appropriately structured information provision can induce first best effort even in the presence of adverse selection (Glazer and McGuire, 2005; Glazer *et al.*, 2008).) Nevertheless, the author starts with this stylized benchmark model because it provides a clear insight into how and why economists and policy makers might expect quality reporting and associated changes in demand to correct market failures in healthcare.

Arguably the single largest violation of the necessary conditions in the benchmark model is the information asymmetry between consumers of healthcare services and producers. In very few cases does an unaided consumer have a good sense of the quality of his or her doctor relative to alternate physicians that might be available or even relative to outside options of a different treatment regime or no treatment at all. Where this is true, a provider who invests in improving the quality of care provided will see few additional patients and, therefore, little additional profit. This is represented by the inverse demand curve ($P_0(q)$) in **Figure 1** that maps quality into a price consumers are willing to pay. $P_0(q)$ is below the full information benchmark curve, $P^*(q)$. Because the provider chooses quality based on trading off the cost of quality improvement compared to the marginal revenue from quality improvement, the equilibrium quality level will be lower than it would otherwise be, q_0 in **Figure 1**. This is not only a positive observation that quality is lower but it also means that the level of quality is suboptimal in a normative

sense. The information problems mean that demand does not reflect the true willingness to pay for quality and therefore the market does not supply the socially optimal level of quality. This can be seen by simply comparing the equilibrium quality under $P^*(q)$ (q^*) to the lower level under $P_0(q)$ (q_0). This basic concern is, either explicitly or more often implicitly, underlying concerns about quality of care provided in most healthcare markets (IOM, 1999).

Quality Reporting to Address the Missing Market for Information

Because asymmetric information is at the heart of the market failure, a natural approach to improving healthcare quality is to try to supply more information to the market actors who are at a relative disadvantage, in this case consumers. Information provision can occur through market-based intermediaries (e.g., consumer reports for consumer products, Yelp for restaurants, or Angies List for skilled professionals including doctors) or as information-based public policy in which government agencies or public–private partnerships gather and disseminate quality information. In both cases, the hope is that consumers will be able to obtain the necessary information to determine the quality of each available healthcare provider and then determine which doctor to choose, given both the cost and the quality of care they will receive. The effect of an intervention of this type can be seen in Figure 1 if the change in marginal revenue from quality after reporting is taken into consideration, represented by $P_1(q)$. Because the information asymmetry has been reduced, it is more profitable to improve quality and the demand curve has rotated upward. As this tracks along the marginal cost of quality improvement a new equilibrium quality level that is higher is reached, q_1. Whether the information is sufficient to overcome the universe of information problems depends on the degree to which $P_1(q)$ moves toward the social optimum, $P^*(q)$. In the example depicted, there remains a large gap both in the incentives ($P_1(q)$ versus $P^*(q)$) and, subsequently, the equilibrium quality, (q_1 vs. q^*).

One appeal of this approach is that changes in quality of care are mediated solely through market incentives, rather than a regulator or payer determining how best to provide high quality healthcare and requiring that physicians practice in a particular manner. Instead, the newly informed consumers will reward the physicians who are best able to provide the quality of care they demand at the minimum cost. (Because demand was relatively unresponsive to quality in the absence of information, such an information intervention is expected to increase the quality of healthcare provided. To the extent that consumers face the true price of care, the improved quality will be unambiguously welfare enhancing.)

Prices and Insurance in Demand

An omnipresent concern in healthcare markets is not merely the information asymmetries in consumer choices but the fact that most consumers seeking care will have access to insurance and, therefore, are unlikely to face the full cost of their care at the margin (Pauly, 1968; Zeckhauser, 1970). The role of moral hazard in demand has important implications for quality reporting for two reasons. First, providing quality information to insured consumers need not enhance welfare even if quality was too low before the information intervention. Because consumers do not face the full cost of their care, they may demand too much quality and this could be exacerbated by the release of quality information. Before quality release the insured consumer was relatively unresponsive to price but could not distinguish high from low quality doctors. Therefore, profit maximizing providers set high prices relative to the cost of their quality, leading to not only inefficiently low quality but excess expenditures given the quality. Once reporting is introduced, the rewards for enhanced quality are greater because higher quality doctors gain more patients. However, because consumers do not face the full price of seeking higher quality care their demand for quality is even greater (they get all of the upside of quality but only pay a fraction of the additional cost). In this case, the introduction of quality information may excessively reward quality improvement relative to the social optimum leading to excessively high investment in quality improvement and, ultimately, quality (Gaynor, 2006; Dranove and Satterthwaite, 2000).

A related issue is the fact that many healthcare providers do not set prices in a standard marketplace. Rather they agree to provide services at an administered set of prices. This is certainly true for doctors and hospitals providing care to patients covered by public programs such as Medicare. Medicare sets prices for hospital services (diagnosis related groups) and physician services (resource based relative value units) based on an estimate for the cost of supplying that particular service. As in any administered price setting, providing appropriate incentives is a challenge (Newhouse, 2002). Because most private payers also base their negotiated price on the Medicare rates these choices not only affect the equilibrium quantity and quality for government paid patients but for the entire market. When prices are more than the marginal cost, the reward for gaining additional patients are large. This leads to the same phenomenon as with moral hazard among consumers. There is a greater reward for additional patients and, therefore, quality improvement. Conversely, when prices are set lower than marginal revenue, administered pricing can lead to inefficiently low quality and reduce incentives for quality improvement associated with shifts in demand. Thus, the introduction of quality reporting in a market with administered prices above cost will yield improve quality of care but may also encourage excess investments beyond the social optimum and vice versa. Clearly there is a relationship here between two policy tools: information provision and the payment rate. If the payer is able to set the appropriate payment such that reimbursement for the marginal patient is just equal to the marginal valuation for quality or care and the marginal cost of quality, the combination of public reporting and payments can provide the first best quality level. (Determining the appropriate marginal patient also poses a challenge. The social planner cares about the average marginal patient but the optimizing provider cares about the marginal patient (Spence, 1980). In this case there can be a further wedge driven between social optimum and competitive equilibrium quality.)

Uncertainty

As with much of healthcare (not to mention other markets), this is easier assumed than done. Distortions to this choice process arise because consumers are typically insured, face uncertain tradeoffs between treatment options and, perhaps most importantly, generally cannot verify the quality of care they received or the quality of different provider options available. Although all of these features contribute to market outcomes in markets for healthcare providers, this article concentrates on the latter: the role of asymmetric quality information and the policy options to address this in determining healthcare quality.

Suppose that consumers (patients) choosing a physician care about three different attributes: price, clinical quality, and amenities. This simplifies the discussion but captures the most salient features of the policy debate regarding cost and quality of healthcare. Dranove and Satterwaite (1992) and Dranove and Satterthwaite (2000) demonstrate the key role that uncertainty plays in determining the level of each attribute and the interaction of consumer preferences and knowledge about each. Although both papers develop technical detail necessary to solve the problem, some simple predictions emerge that motivate this discussion of quality information. If patients are uncertain about two attributes, they will tend to overweigh the observation that is more certain and underweigh the less certain attribute. Translating this into the incentives facing suppliers of care, the relatively observed measure is expected to have excessive investments in improvement and the converse for the relatively harder to observe provider attribute.

This finding is of particular concern in healthcare markets where clinical measure of performance are generally characterized by uncertainty but service amenities are far easier to observe. This is true in the absence of report cards; physicians and hospitals will invest in amenities excessively relatively to efforts to improve clinical quality. For example, hospitals have, for a long time, provided well-appointed entryways and valet parking but their attention to regular hand washing and infection control is a far more recent effort (arguably brought about by Medicare payments not demand changes) (Goldman *et al.*, 2010). It is also relevant in the response to report cards. If quality reporting includes satisfaction or service metrics as well as harder to interpret clinical measures, quality reporting might exacerbate the relative focus on service as opposed to clinical quality.

Evidence on Quality Reporting and Demand

Quality and Demand for Primary Care Physicians

For most patients, their first point of contact with the healthcare system is through their primary care physician. The choice of and role for the primary care doctor distinguishes this choice when quality reporting and demand is considered. Specifically, defining high quality primary care is a challenge and determining what information patients might use to choose a primary care doctor also poses an issue.

Choice of a primary care doctor exhibits the universe of choice problems that characterize demand in healthcare. Quality is highly uncertain and depends on the health of the patient. Jointness in production is also key. A primary care doctor is there to treat conditions but, particularly for patients with chronic conditions, the doctor is likely to suggest complementary behaviors by the patient that are as (if not more) important for improving health. A high quality doctor treating a diabetic patient could screen for the level of sugar in the blood (the HBalC test) but this is not, in and of itself, a treatment for any of the ailments of diabetes. Instead, this measure will allow the physician to adjust medications to manage the disease but also to help council the patients on how they should change their eating habits.

Even when there is a reasonable measure of disease, as in the case of a diabetic patient, defining end points the characterize a high quality primary care doctor is a challenge. Ideally, most patients are healthy and remain so. Attributing the fact that a patient becomes ill to low quality primary care though would not produce the right incentives. Everyone will become ill at one point; death along with taxes are certain in this life. Furthermore, a patient who is identified as becoming ill may have received better care from their primary care doctor because the disease was found.

All of these factors make the introduction of quality report cards for primary care doctors a particular challenge. Instead, most quality reporting efforts in the primary care arena have been focused either on identifying high quality production and documenting such process measures or paying for those measures directly (e.g., pay-for-performance). Some forms of quality information on primary care physicians is becoming widely available through market-based information such as Angie's List. These tools aggregate assessments of physicians quality from customer responses. This form of quality reporting is less studied but presents an important challenge to providing the appropriate social incentives for quality improvement (e.g., whether and how to weigh satisfaction relative to clinical quality depends critically on the form of the production function, consumer tastes and own, and cross quality elasticities with respect to these measures). Recall the results from Dranove and Satterthwaite (2000) discussed above. If people can observe patient satisfaction with relative ease but have a much harder time determining clinical quality or diagnostic ability, these tools may lead to an excess investment in patient satisfaction and an underinvestment in clinical quality. Of course, patients may truly value satisfaction in which case the emphasis on clinical quality could be unfounded. There is, however, ample reason to believe that information problems as well as myriad choice errors documented in behavioral economics are likely to lead to suboptimal choices with respect to clinical quality (Frank, 2004).

Quality Reporting and Specialist Choice

The role of quality information and quality reporting in choice of specialists is the most studied of any of the ways information might impact demand. Specialist choice is also the area in which information-based policy could have the biggest impact on welfare. Specialists perform specific and often measurable tasks. They also encounter patients when

they are sick and receiving the most intensive care; precisely where one might expect quality improvement to enhance outcomes. There is a voluminous empirical literature focused on the response to quality information on specialists. Rather than surveying the literature, a task that has been done in a number of other settings (see, e.g., Kolstad and Chernew, 2008 and Dranove and Jin, 2010 for recent reviews), the focus here is on a specific setting: the introduction of quality report cards in the market for CABG surgery. Though quite specific, the CABG case is the most studied quality reporting initiative. Furthermore, the results, methodologies, and issues are, generally, indicative of the broader findings on the role of quality reporting and demand.

Beginning in 1988, New York State gathered and reported hospital-level risk adjusted mortality rates (RAMR) for CABG surgery. Shortly thereafter, following freedom of information request, surgeon-specific RAMR was reported beginning in 1991. Pennsylvania followed suit shortly thereafter, initially introducing quality report cards in 1993, though report cards were not widely available until 1998 when reports based on 1994–95 data were disseminated. Subsequently, many more states and countries have begun to provide CABG quality report cards. As of 2006, 47 states and the UK all offered CABG report cards (Steinbrook, 2006).

There are two broad veins of literature that provide insight in the CABG surgery experience with quality reporting: survey based and actual choice based. Mukamel and Mushlin (1998) find that both hospitals and doctors with better RAMR saw an increase in market share following the release of quality reporting. Culter *et al.* (2004) also study the New York experience. Their paper extends the basic test for an effect of market share in two important directions. They allow for heterogeneous response to RAMR depending on whether a hospital is above or below expected. They also compare the response of patients who are more likely to be able to switch surgeons, those that are less severe. Their results suggest a significant response to hospitals that are flagged as high mortality (lower quality) than expected. They find a high mortality designation is associated with an average decline of 5 CABG surgeries per month, or approximately 10% of a hospital's volume. Interestingly, they do not find a commensurate increase in demand for hospitals identified as a lower than expect mortality (high quality). The response to quality information is almost entirely driven by changes among patients who are relatively healthy; presumably the patients who have the time and opportunity to choose between hospitals. Dranove and Sfekas (2008) also study the response to the release of quality report cards. Their model makes an important contribution by accounting for prior beliefs of consumers. That is, if consumers or referring physicians already know a hospital is of high quality, one would not expect much effect of releasing information on market share. As in Culter *et al.* (2004), they find that demand responds more strongly to quality information after accounting for prior market-based learning. They also find that there is an asymmetric response with patients more responsive to learning a hospital is of lower than expected quality than to learning a hospital is better than expected. Kolstad (2012) also studies the demand response to quality reporting for CABG surgery, in Pennsylvania in this case. The distinction of this model, with respect to understanding demand, is that it allows for taste heterogeneity among consumers. As with the earlier literature, he finds a significant response to quality after the release of quality report cards. This effect varies significantly in the population with a small share of the population responding very strongly to quality after report cards.

Survey-based evidence also suggests an effect of report cards on demand, though these studies generally find little impact on actual patients with more impact on referring cardiologists. Schneider and Epstein (1996) address this question directly by surveying cardiologists in Pennsylvania. They find that roughly 10% of cardiologists found quality information very important. In New York State, Hannan *et al.* (1996) conducted a similar survey of cardiologists finding a larger response; 38% of cardiologists report that the report card's their referral pattern. A relatively strong responsive of a minority of referring physicians is consistent with the higher-level empirical results that find a similar observed response.

Taken together, these results are indicative of many of the findings in the literature on quality reporting and demand for specialists. First, quality reporting has a small but significant effect on demand. Second, this effect is convex in quality – consumers seem more willing to pay (travel) to avoid low quality specialists than they are to access high quality specialists, even though the change in quality is the same in both cases. Third, there is substantial heterogeneity in the population in whether and how patients respond to the information. A small minority of patients respond strongly by switching hospitals and surgeons but there is relatively little movement by the bulk of the patient population. Fourth, physician agents and/or existing market-based learning means that higher quality providers tend to have greater demand even before the release of report cards. This has the effect both of muting the estimated response to quality information release and suggesting that the institutions of the healthcare market are able to inform consumers somewhat without intervention.

Quality Information and Supply

The focus of this article is on the role of quality information in demand. However, the process of gathering, analyzing, and synthesizing quality information also has the potential to inform suppliers, in this case physicians and hospitals. If these efforts both provide new information and inform suppliers who care about quality beyond the pecuniary rewards, then quality reporting can impact outcomes without shifting demand. How and why this process occurs is a new and relatively unexplored area of research on quality reporting and outcomes. However, Kolstad (2012) demonstrates that, in the market for CABG surgery in Pennsylvania, the impact of information provided to suppliers had an effect on quality improvement that was four times larger than the impact mediated through changes in demand, the type of impact focused on primarily in this discussion. In this model, new information impacts physicians because they care intrinsically about supplying high quality care and quality reporting allows them to better observe their performance relative to their peers. A detailed case study of the impact of New York State's

CABG reporting program also provides anecdotal evidence for a very similar impact. Dziuban *et al.* (2008) document the important role that the release of quality report cards had in motivating efforts to improve the process of care to lower mortality at a large community hospital. Interestingly, they find that the new information led to large changes despite the fact that the hospital had a detailed data capture and outcome review process in place beforehand. This underscores potentially important features that usually characterizes quality reporting: volume and scope of observations and risk adjustment. If quality reporting efforts gather data from across settings (e.g., hospitals or health systems) and develop state-of-the-art risk adjustment models they are likely to provide new information to providers, even if those providers have local monitoring tools in place (e.g., electronic health records or 'morbidity and mortality' conferences to discuss quality and errors). These issues are key to understanding the aggregate role that quality reporting and quality information might play in market outcomes and quality of care. The precise model of beliefs and preferences, however, that leads quality information to affect supplier behavior remains an outstanding question.

Another important supply side response to quality reporting that has been much discussed is the potential for providers to try to select healthier patients to improve their scores. The inclusion of risk-adjustment is intended to address this issue. With sufficiently good risk-adjustment the incentives for selection are eliminated. In practice, however, this is extremely difficult if not impossible. In the absence of perfect risk-adjustment, one must consider the trade-off between selection incentives and the gains in welfare associated with quality improvements due to information. This underscores the fact that the existence of selection against sicker patients does not, in and of itself, eliminate the value of quality reporting efforts. It does, however, raise important welfare trade-offs and require some consideration of the distributional impacts of quality reporting across the spectrum of patients. Despite a great deal of hypothesizing about selection efforts, there are relatively few empirical studies that document selection in response to report cards. The most prominent is work by Dranove *et al.* (2003). They study the impact of New York's and Pennsylvania's introduction of CABG quality reporting in the Medicare population. They find evidence that quality reporting enhanced patient matching to surgeons and hospitals and that there was selection against sicker patients. The aggregate impact of quality reporting was to reduce welfare – the losses from selection outweighed the gains from reduced information asymmetries. This article raises important issues and also presents a useful methodology for evaluating quality reporting efforts.

Quality Reporting and Demand in Other Markets: Comparing Healthcare to Education

Healthcare is one of a number of important fields in which information-based public policy plays a role. It is informative to compare the experience in other markets as a benchmark for understanding the impact of quality reporting on demand. Education provides a useful comparison. There are a number of similarities between choices of healthcare providers and choices of schools that both rationalize the reliance on information-based public policy and make a comparison between the two fruitful. In both cases, information asymmetries are a defining feature of demand. In both cases, consumers are being asked to choose between different suppliers with difficult to verify differences in skills and quality. Outcomes are also characterized by joint production. How much students learn is affected both by teacher effort and student effort. Similarly, the efficacy of many treatments – particularly those for chronic conditions – rely on physician effort as well as patient's willingness to follow advice and change behavior. In both cases, there is also an important component of random noise in outcomes and incentives to teach to the test or select healthy patients. Finally, in both markets the externalities associated with providing the good mean that public provision is preferred and, therefore, most consumers face little, if any, price variation.

So how does the experience with quality reporting in education compare to the CABG market? The main findings appear to be similar, though it is noted this is far from a complete review of the education literature (that would require its own online encyclopedia, let alone section of an article). Hastings and Weinstein (2006) study the response of parents and students to the public provision of information on school quality. They find a significant response to the information on school quality, measured by test scores. After information is released parents are more likely to choose a higher quality school by 5.7 percentage points. They also find that attending a higher quality school improves student test scores. Glazerman (1998) studies school choice and finds that consumers are highly responsive to distance. The effect of quality on choice is diminished substantially as higher quality schools are further away. Hanuscheka *et al.* (2007) study the release of quality information for Texas charter schools. They find that the release of quality information increases the likelihood that low quality schools exit the market significantly.

Although clearly not a comprehensive review, these important papers demonstrate some striking similarities to the role of quality reporting in healthcare markets. First, there appears to be a substantial response to quality reporting, though this effect is relatively small. Consider the main estimate from Hastings and Weinstein (2008). They find that roughly 1 in 20 parents was responsive to quality information. Comparing this to the response in Culter *et al.* (2004) this is a similar magnitude, though smaller, than the response to quality reporting for CABG in New York State. There hospitals saw a decline in volume in the year following a high mortality flag of 10%. Similarly, the findings that distance seems to weigh very strongly in the choice of schools is quite similar to healthcare demand. Both industries are characterized by local markets but it is striking that the healthcare consumers who generally make far fewer trips to a hospital are so responsive to distance. It is less surprising, given the daily trips to school, that distance has an effect. That said, in both cases one would expect the utility for a good outcome (e.g., lower mortality or morbidity and improved lifetime earnings) to be very high so substantial distance effects are surprising and potentially indicative of information problems that limit the response to high quality.

Conclusion

In this article the basic theoretical underpinnings for quality reporting and demand in healthcare markets have been covered. The rationale for quality reporting is based on the potential for asymmetric information on quality that leads to suboptimal outcomes. It has been seen, in a simple framework, how changes in demand due to information provision can change equilibrium quality. Whether these changes improve welfare depends on a number of features of the market. Ultimately, however, the normative standard is the full information demand curve that would exist. The empirical evidence on the impact of quality reporting on demand suggests small average effects with important heterogeneity in the population. Comparing the estimates to the impact of quality reporting in education suggests the impact of quality reporting in cardiac surgery has been larger.

Despite the many simple empirical studies of quality reporting and demand, the application of econometric tools and field experiments to better understand the precise mechanism by which quality reporting improves quality is an important next step for research. For example, incorporating the supply side response into the evaluation of the policy dramatically alters the way in which quality reporting policies should be evaluated as well as the way in which these interventions should be structured (e.g., should information be simplified to target consumers or made more clinically relevant for physicians?). As health reforms in the USA and the inevitable efforts to address cost and quality in all healthcare systems move forward, there is a key role for addressing information failures. Further understanding of the impact of quality information on demand and on market outcomes should be a key item on the research agenda in economics and health services research and an area of focus for policy makers for many years to come.

See also: Advertising Health Care: Causes and Consequences. Comparative Performance Evaluation: Quality. Competition on the Hospital Sector. Demand Cross Elasticities and 'Offset Effects'. Demand for Insurance That Nudges Demand. Heterogeneity of Hospitals. Physician-Induced Demand. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Risk Adjustment as Mechanism Design. Specialists. Switching Costs in Competitive Health Insurance Markets. Value-Based Insurance Design

References

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**, 941–973.

Congdon, W., Kling, J. and Mullainathan, S. (2011). *Policy and choice: Public finance through the lens of behavioral economics*. Washington, DC: Brookings Inst Press.

Cutler, D. (2003). *Your money or your life: Strong medicine for America's health care system*. USA: Oxford University Press.

Culter, D., Huckman, R. and Landrum, M. B. (2004). The role of information in medical markets: An analysis of publicly reported outcomes in cardiac surgery. *American Economic Review (Papers and Proceedings)* **194**(2), 342–346.

Dranove, D. and Jin, G. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature* **48**(4), 935–963.

Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (2003). Is more information better? The effects of 'report cards' on health care providers. *Journal of Political Economy* **III**(3), 555–558.

Dranove, D. and Satterthwaite, M. (1992). Monopolistic competition when price and quality and imperfectly observable. *The RAND Journal of Economics* **23**(4), 518–535.

Dranove, D. and Satterthwaite, M. (2000). *The industrial organization of health care markets*, ch 20. North Holland: Elsevier.

Dranove, D. and Sfekas, A. (2008). Start spreading the news: A structural estimate of the effects of New York hospital report cards. *Journal of Health Economics* **27**(5), 1201–1207.

Dziuban, S., McIlduff, J., Miller, S. and Dal Col, R. (2008). Start spreading the news: A structural estimate of the effects of New York hospital report cards. *Journal of Health Economics* **27**(5), 1201–1207.

Frank, R. (2004). Behavioral economics and health economics. *NBER Working Paper (10881)*. Cambridge, MA: National Bureau of Economic Research.

Gaynor, M. (2006). What do we know about competition in health care markets? *NBER Working Paper (12031)*. Cambridge, MA: National Bureau of Economic Research.

Glazer, J. and McGuire, T. (2005). Optimal quality reporting in markets for health plans. *Journal of Health Economics* **25**, 295–310.

Glazer, J., McGuire, T., Cao, Z. and Zaslavsky, A. (2008). Using global ratings of health plans to improve the quality of health care. *Journal of Health Economics* **27**, 1182–1195.

Glazerman, S. (1998). *School quality and social stratification: The determinants and consequences of parental school choice*. San Deigo, CA: ERIC.

Goldman, D., Vaiana, M. and Romley, J. (2010). The emerging importance of patient amenities in hospital care. *New England Journal of Medicine* **363**, 2185–2187.

Hall, R. and Jones, C. (2005). The value of life and the rise in health spending. *Quarterly Journal of Economics* **122**(I), 39–72.

Handel, B. (in press). *Adverse selection and inertia in health insurance markets: When nudging hurts*. Mimeo, University of California.

Hannan, E., Stone, C., Theodore, B. and DeBuono, B. (1996). Public release of cardiac surgery outcomes in New York: What do New York state cardiologists think of it? *American Heart Journal* **134**(1), 55–61.

Hanuscheka, E., Kainb, J. and Steven, G. (2007). Rivkinb an Gregory Branch. Charter school quality and parental decision making with school choice. *Journal of Public Economics* **91**(5–6), 823–848.

Hastings, J. and Weinstein, J. (2006). Information, school choice and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics* **123**(4), 1378–1414.

Hastings, J. and Weinstein, J. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *Quarterly Journal of Economics* **123**(4), 1373–1414.

IOM (1999). *To err is human: Building a safer health system*. Washington, DC: Institute of Medicine of the National Academics.

Kolstad, J. (2012). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *NBER Working Paper No. 18804*. Cambridge, MA: National Bureau of Economic Research.

Kolstad, J. and Chernew, M. (2008). Consumer decision making in the market for health insurance and health care services. *Medical Care Research and Review* **66**(1), 28S–52S.

Mukamel, D. and Mushlin, A. (1998). Quality of care information makes a difference: An analysis of market share and price changes after publication of the New York state cardiac surgery mortality reports. *Medical Care* **36**(7), 945–954.

Murphy, K. and Topel, R. (2006). The value of health and longevity. *Journal of Political Economy* **114**(5), 871–903.

Newhouse, J. (2002). *Pricing the priceless: A health care conundrum*. Cambridge MA: The MIT Press.

Pauly, M. (1968). The economics of moral hazard: Comment. *American Economic Review* **58**, 531–537.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics* **90**(4), 629–649.

Schneider, E. and Epstein, A. (1996). Influence of cardiac-surgery performance reports on referral practices and access to care – A survey of cardiovascular specialists. *Journal of Health Economics* **335**, 251–256.

Spence, A. M. (1980). Product selection, fixed cost, and monopolistic competition. *Review of Economic Studies* **43**(2), 217–235.

Steinbrook, R. (2006). Public report cards – Cardiac surgery and beyond. *New England Journal of Medicine* **255**, 1847–1849.

Zeckhauser, R. (1970). Medical insurance: A case study in the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**, 10–26.

Quality-Adjusted Life-Years

E Nord, Norwegian Institute of Public Health, Oslo, Norway

Introduction

The quality-adjusted life-year (QALY) is a unit of measurement for valuing health outcomes. The background for it is illustrated in Figure 1.

In the diagram, length of life is expressed on the X-axis, whereas health status is expressed on the Y-axis on a continuum from dead to full health. The lower line shows the health of some hypothetical person over time with standard treatment. The upper line shows the health over time given some alternative better treatment. The total health gain from moving from standard to better treatment – represented by the area between the two lines – consists in gains both in level of health and length of life. The QALY is designed to capture in one single measure of value both these types of benefits so that they are made comparable and also may be added to each other. How this is done technically, is explained in the section Definition, Operationalization, and Meaning.

The value of an outcome measured in QALYs may be related to the cost of achieving the outcome. This is done in a so-called cost-effectiveness ratio (often called a cost-utility ratio). Cost-effectiveness ratios are indicators of value for money. Cost-effectiveness ratios using QALYs allow comparisons of value for money of different interventions in different areas of medicine – in which outcomes of different kinds are achieved – and may thus be an aid in priority setting and resource allocation decisions.

Definition, Operationalization, and Meaning

In the QALY approach, 1 life year in full health for one person is used as a basic, reference outcome. For brevity, this is referred to as '1 well-year.' One gained well-year is assigned a value of one QALY. The idea of the QALY approach is that any health outcome, whatever its nature and size, may be valued relative to the reference outcome, i.e., as equivalent to gaining some fraction of a well-year or some multiple of well-years.

Health outcomes that include gains or losses in quality of life are made comparable to outcomes consisting in gained well-years through the assignment of values to health states (Figure 2). The values on the Y-axis reflect the quality of life associated with the states – often called health-related quality of life. The values are on a scale from zero – corresponding to being dead or in a state as bad as being dead – to unity – corresponding to being in full health. They are used to weight life years in less than full health. For example, in Figure 2, state A is assigned a value of 0.8. Each gained life year in state A then yields 0.8 QALYs. This means that the gain is deemed equivalent to gaining 0.8 of a well-year.

The number of QALYs in an individual's health scenario over time is calculated by determining the value of each year in the scenario and summing these annual values over the whole time horizon. For example, if a person lives 3 years with values 0.8, 0.6, and 0.5 respectively, the value of the whole scenario is $0.8+0.6+0.5=1.9$ QALYs. In Figure 2, 10 years gained in state B yields 6 QALYs (10×0.6), whereas 12 years in state A yields 9.6 QALYs. An improvement from state B to state A lasting 1 year yields 0.2 QALYs ($0.8-0.6$). If the improvement lasts 10 years, it yields 2 QALYs (10×0.2) and is thus equivalent to 2 gained well-years. In Figure 2, the total value of replacing the first scenario with the second one is $10\times0.2+2\times0.8=3.6$ QALYs.

Values and equivalence of health outcomes may be perceived and judged from different points of view. There is, for instance, a difference between pure self interest of individuals

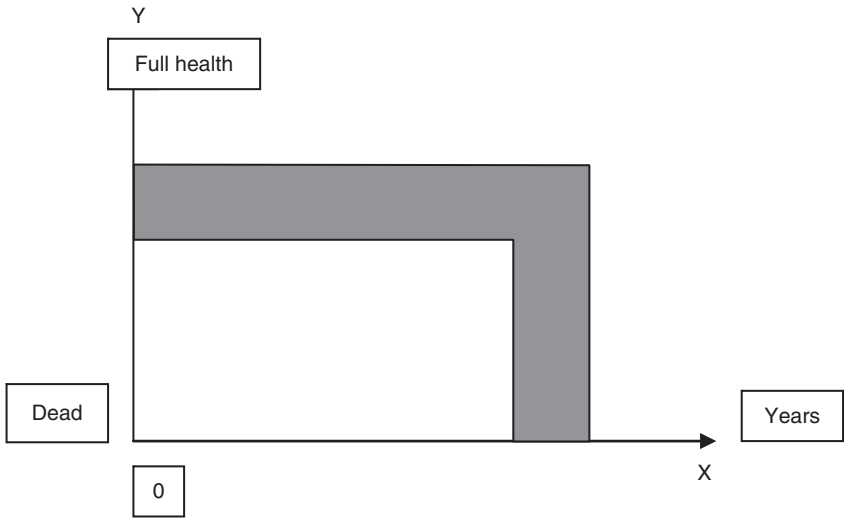


Figure 1 Health scenarios without and with treatment.

 doi:10.1016/B978-0-12-375678-7.00510-1

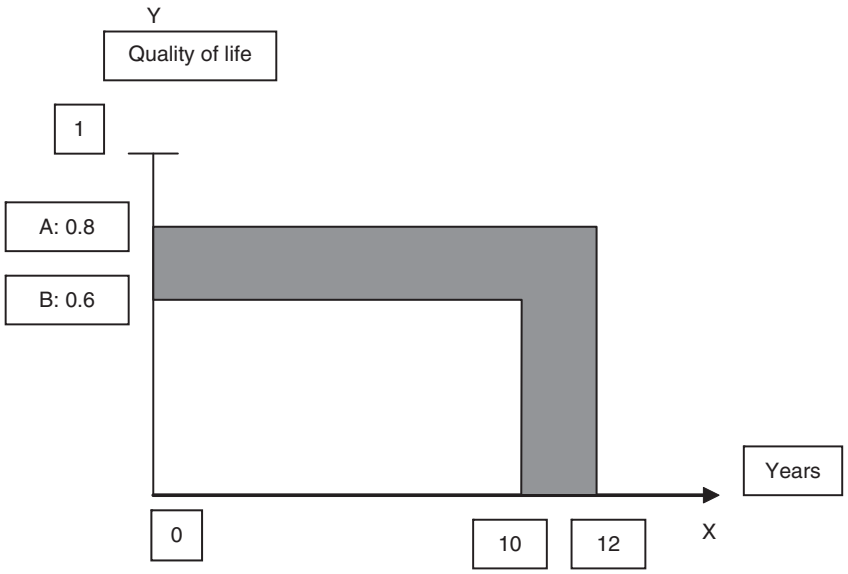


Figure 2 Health scenarios with specification of duration and values for health states.

and judgments of value by societal decision makers when concerns for distributive fairness are taken into account. Estimates of value in terms of QALYs are usually meant to express the personal value of outcomes to the recipients of the outcomes, without regard to distributive issues. Such personal value is commonly referred to as individual utility. An approach to valuing health outcomes from a broader societal perspective was outlined by Anthony Culyer and colleagues as early as in 1971 and later by for instance Erik Nord in 1992 and 1999.

To calculate the area under each curve in Figure 1 requires knowledge of the nature of the health states that are experienced, the sequence of these, the duration of each health state and the value of each health state. The former three kinds of knowledge require medical (clinical and/or epidemiological) data. Valuations of health states, on the other hand, is a psychometric task in which health economists have engaged heavily (together with researchers from other disciplines) with a view to making the QALY approach workable. Valuations are generally elicited from samples of the general population by means of one or more specially designed techniques for preference measurement, thereafter values are assigned based on mean or median responses.

Issues

As noted above, estimates of value in terms of QALYs are usually meant to express personal value. Personal value is commonly referred to as individual utility. There are a number of methodological issues relating to the QALY approach as a way of estimating the individual utility of interventions.

First, health state utilities are usually obtained by asking representative samples of the general population to judge how bad it would be for them to be in different states of illness. This is referred to as decision utility. An alternative is to ask patients and disabled people to value the states they are in themselves. This would yield experience utility. Both approaches have strengths and weaknesses. As noted by Drummond *et al.* (2009), a widely held position in current health economics is that QALYs should be based on decision utilities elicited from people who are well informed about experience utility.

Second, for QALY-calculations to be meaningful and valid, utilities for health states need to have interval scale properties. That is, a change of a given size on the 0–1 value scale must have the same importance wherever it occurs on the scale. For instance, a move from 0.4 to 0.6 must have the same value as a move from 0.7 to 0.9. Techniques for preference elicitation vary with respect to their ability to yield utilities with interval scale properties.

Third, as noted above, the number of QALYs in a health scenario over time is calculated by determining the utility of each year in the scenario and summing these annual utilities over the whole time horizon. This means that the number of QALYs obtained from spending time in some fixed state is directly proportional to the length of that time. In other words, as noted by Bleichrodt *et al.* (1997) in a paper on 'risk neutrality of life years,' utility measured in terms of QALYs is (by definition) a linear function of length of time. In the real world, there is not necessarily a linear relationship between duration and utility. Individuals may, for instance, have diminishing marginal utility of length of life in the same way as they normally have diminishing marginal utility of goods and services. However, it is customary to discount QALYs in future years to take account of individuals' preferences for present consumption over future consumption. The number of discounted QALYs is less than proportional to the length of time in a state.

Fourth, in the QALY approach, each possible health state is associated with a single, fixed value. The value of a state is thus assumed to be constant across all individuals and contexts in which it may occur. Clearly, this single value convention is a simplification. Its rationale is a need to prevent valuation in terms of QALYs from becoming too demanding with respect to data and thus too complicated and time consuming.

The convention is reasonable in many circumstances where QALYs are estimated in groups of people, in which cases individual deviations from standard health state values to a large extent cancel each other out. But the simplification does have some implausible implications that lead to continuous debates about the validity of QALYs. The most salient issues are noted below. Note that they all refer to valuations from a personal perspective. Some similar issues may be raised on grounds of concerns for fairness.

First, the single value convention means that the disutility of a state is considered to be independent of the gender, age, and other characteristics of the person who experiences the state, including the person's attainable level of functioning. For instance, it means that dependence on eyeglasses or a walking stick is counted as equally bad for an elderly person as for an young person. It also means that inability to walk is counted as equally bad when it is occurs in a person with a longstanding incurable disability as when it is due to temporary, curable disease in a person who is normally in full health. These implications have been challenged on the grounds that most people have considerable capacity to adapt to and cope with durable disease and disability. This is particularly true of impairments resulting from normal ageing. So even if a person has health problems that people of medium age and normal health would regard as clearly undesirable, and if the person's functional level is the best that he or she can reasonably hope for, his or her utility will not necessarily be much lower than that normally associated with full health. As noted by Erik Nord, Anja Enge, and Veronika Gundersen in 2012, this dependence of utility on what is to be expected and what one is used to may have important implications for valuations of gains in both length and quality of life in people with disability or chronic disease.

Second, the single value convention implies that the utility of a state to a person is considered to be independent of the person's health in the past. For instance, if two people are dependent on eyeglasses, and one was blind in the preceding years, whereas the other had normal sight, the two will be assigned the same utility in their situation with eyeglasses. Concerns about this assumption led to a proposal by Mehrez and Gafni (1989) of an alternative to the QALY approach in which any scenario over time is valued as a whole instead of as a sum of independent valuations year by year. In the alternative approach, the unit of measurement is called the healthy year equivalent (HYE). The disadvantage of this approach is that valuation must be undertaken on all relevant sequences of health states, which may be numerous. Thus, there has been little use of the HYE approach in economic evaluation hitherto.

Third, the single value convention means that the value assigned to a health state does not depend on the duration over which the state is experienced. For instance, the utility of a state in the first year after the onset of a disability is the same as the utility of that state 5 years later even if the person in various ways may adapt to the state. At a more technical level, independence of duration further means that there are implicit assumptions of so-called mutual utility independence and constant proportional trade-off between quality of life and length of life when preferences for health states are elicited.

Fourth, the single value convention means that the utility of a state is considered to be independent of its cause. For instance, the utility is the same for a congenital problem as for a problem caused by hospital negligence.

The concept of QALYs has been linked explicitly to expected utility theory. In a decision analytic framework, QALYs are used as the unit of account in expected value calculations for decisions under uncertainty. Expected utility theory dictates that in order for these expected value calculations to be consistent with preferences over uncertain streams of health outcomes, QALYs should fulfill certain requirements of 'utility' functions. These requirements take the form of a set of axioms, formalized by von Neumann and Morgenstern in 1944. Although much has been written about empirical violations of the axioms of expected utility theory and various alternatives have been proposed, expected utility remains an important point of reference in many discussions of normative decision theory.

Several techniques are available for eliciting health state utilities from respondents, including the standard gamble, time trade-off, and the rating scale. Empirical studies have found that the different techniques produce values that differ systematically. Debates about the relative merits of the different methods refer both to economic theory and to comparisons of psychometric properties of the different measurement techniques.

It is clear from the points made above that the interpretation of health state values as estimates of individual utility in many ways is questionable. It is furthermore a fact that the second factor in QALY-calculations – the duration of health benefits – is purely a quantitative factor with no personal value judgments related to it. For instance, 10 years is simply counted as twice as much as 5 years, with no value judgment involved. Altogether, some therefore prefer to regard the results of QALY-calculations as indicators of the size of health effects rather than the utility – or personal value – of those effects. This alternative interpretation does not, however, alter the basic purpose of QALYs, which is to yield a quantitative estimate of efficiency in different areas of health care.

The QALY approach has been criticized on ethical grounds for implying priority to those individuals who have the greatest capacity to benefit from health care and for not taking into account concerns for fairness in the distribution of health care resources. Historically, the critique is understandable, given the primacy that cost-effectiveness ratios have been assigned in most of the health economics literature as guidance to priority setting. But a distinction needs to be made between calculations of QALYs and how the calculations are used in decision making. QALY estimates are essentially estimates of the aggregate individual utility of interventions. Although such information may be an important input in decision making about resource allocation, it does not follow that priorities should be set such that QALY gains are maximized as this may run counter to concerns for fairness such as wishes to give priority to the worse off and wishes to secure equal access to people in equal degree of need even if they are of different ages and/or have different potentials for health.

Historical Overview

(Reproduced from the Encyclopedia of Public Health with the permission of the publisher (Elsevier) and the author, Josh Salomon.)

Although the term QALY first appeared in the published literature in 1976, some earlier precedents may be found. Herbert Klarman and colleagues in 1968 compared three options for treating patients with chronic renal disease in terms of life years gained, with and without adjustments for 'differential(s) in the quality of life.' In 1970, Sol Fanshel and James Bush proposed measures of dysfunction-free years in evaluation of a tuberculin skin testing program. In 1971, Anthony Culyer and colleagues proposed a scheme for weighting life years within a social indicator framework. At the same time, George Torrance and colleagues introduced the index day and health day. Finally, the term quality-adjusted life-year was used in two separate publications in 1976, one by Milton Weinstein and William Stason examining policies for control of hypertension and another by Richard Zeckhauser and Donald Shepard in a more general exploration of analytic approaches to evaluating social policies with life-saving or health implications. An article by Weinstein and Stason appearing in the *New England Journal of Medicine* in the following year introduced QALYs to a broad medical and public health audience and is frequently cited as a major milestone in the development of cost-effectiveness analysis in health and medicine. In England, a highly influential early paper applying QALYs to evaluation of coronary artery bypass grafting was reported in an article by Williams (1985) in the *British Medical Journal*.

Researchers have developed a registry of cost-effectiveness studies that report outcomes specifically in terms of costs per QALY. Peter Neumann and colleagues reviewed the literature from 1976 through 2001 and identified 533 original studies meeting their inclusion criteria. Consistent with the earlier reviews of the broader cost-effectiveness literature, a major increase in the volume of studies on costs per QALY was evident, with 228 studies over the two decades from 1976 to 1997, followed by 305 studies over the 4-year period from 1998 through 2001.

Alternatives to QALYs

The disability adjusted life year (DALY) is a summary measure much like the QALY, developed by Christopher Murray and colleagues. The main difference is that it uses a scale of severity of illness from zero (full health) to unity (as bad as being dead) instead of a scale of utility from zero (dead) to unity (full health). The DALY was first developed for the primary purpose of quantifying the global burden of disease. However, the developers of the DALY explicitly intended that the measure could be used also as a metric for health benefits in the denominator of cost-effectiveness ratios. In the present day, the DALY is widely used – in fact much more than QALYs – in economic evaluation of health programs in developing countries.

The QALY procedure focuses on life years and the quality of these. In a paper in the *British Medical Journal* in 1992, Erik Nord argued that the health care system is concerned with providing care for people ('living, breathing, feeling, and thinking individuals'), not with maximizing numbers of abstract time entities. The health care system is also concerned with meeting moral claims on treatment. The concept of claims is related to living subjects. Life years as such are not subjects and therefore do not have moral claims. Nord thus suggested the saved young life equivalent (SAVE) as an alternative to the QALY that focuses on persons rather than years. In the SAVE approach, the reference outcome consists in saving the life of a young person and restoring him or her to full health. The value assigned by society to this reference outcome is called a SAVE. In valuations that have a societal rather than an individual viewpoint, different kinds of health outcomes may all be valued relative to the SAVE, for instance using the so-called person trade-off technique. However, there has been little use of the SAVE in economic evaluation hitherto.

Acknowledgments

In writing this entry, Erik Nord received valuable inputs from Anthony Culyer and Joshua Salomon.

See also: Cost–Value Analysis. Disability-Adjusted Life Years. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Measurement Properties of Valuation Techniques. Time Preference and Discounting. Utilities for Health States: Whom to Ask. Valuing Health States, Techniques for

References

Bleichrodt, H., Wakker, P. and Johannesson, M. (1997). Characterizing QALYs by risk neutrality. *Journal of Risk and Uncertainty* **15**, 107–114.

Drummond, M., Brixner, D., Gold, M., et al. (2009). Toward a consensus on the QALY. *Value in Health* **12**(Supplement), S31–S35.

Mehrez, A. and Gafni, A. (1989). Quality-adjusted life years, utility-theory, and healthy-years equivalents. *Medical Decision Making* **9**, 142–149.

Williams, A. (1985). Economics of coronary artery bypass grafting. *British Medical Journal* **291**, 326–329.

Further Reading

Broome, J. (1993). QALYs. *Journal of Public Economics* **50**, 149–167.

Drummond, M. F., Sculpher, M., O'Brien, B., Stoddart, G. L. and Torrance, G. W. (eds.) (2005). *Methods for the economic evaluation of health care programmes.* Oxford: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (eds.) (1996). *Cost-effectiveness in health and medicine.* Oxford: Oxford University Press.

Hauber, A. B. (2009). Healthy-years equivalent: Wounded but not yet dead. *Expert Review of Pharmacoeconomics Outcomes Research* **9**, 265–269.

Neumann, P. J., Greenberg, D., Olchanski, N. V., Stone, P. W. and Rosen, A. B. (2005). Growth and quality of the cost-utility literature, 1976-2001. *Value in Health* **8**, 3–9.

Nord, E. (1999). *Cost-value analysis in health care: Making sense out of QALYs.* Cambridge: Cambridge University Press.

Nord, E., Daniels, N. and Kamlet, M. (2009). QALYs: Some challenges. *Value in Health* **12**(Supplement), S10–S15.

Richardson, J. (1994). Cost-utility analysis: What should be measured? *Social Science & Medicine* **39**, 7–21.

Weinstein, M., Torrance, G. and McGuire, A. (2009). QALYs: The basics. *Value in Health* **12**(Supplement), S5–S9.

Rationing of Demand

L Siciliani, University of York, York, UK, and Centre for Economic Policy Research, London, UK

Glossary

Asymmetry of information A situation in which the parties in a transaction have different amounts or kinds of information as when, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances, or people seeking insurance have more reliable expectations of their risk exposure than insurance companies.

Copayment An arrangement whereby an insured person pays a particular percentage of any bills received for health services, with the insurer paying the remainder.

Creaming A form of favorable selection in health insurance by which the insurer obtains a higher proportion of good risks (people with a low probability of needing care or who are likely to need only low-cost care or both) in their portfolio of clients than is assumed in the calculation of the insurance premiums. It is also called cherry-picking.

Diagnosis-related group (DRG) Diagnoses grouped according to their clinical similarity and the cost of treatment.

Elasticity of demand A measure of the responsiveness of the consumption of a good or service to a change in its price.

Fee-for-service A method of remunerating professionals (especially medical doctors) according to an agreed fee schedule specifying what is payable for each item of service supplied.

Marginal cost The additional cost incurred if the output rate is increased by a small amount.

Moral hazard A risk that can occur when the insurer has imperfect information on the likely behavior of insured individuals. There are two main types. Ex ante moral hazard refers to the effect that being insured has on safety behavior, generally increasing the probability of the event insured against occurring. Ex post moral refers to the possibility that insured individuals will behave in such a way after an insured event has occurred that will increase the claim cost to insurers, partly because the user-price of care is lower through insurance and demand may therefore rise. It is also often related to insurance fraud.

Oregon experiment In 1989 the US state of Oregon initiated a controversial reform of its Medicaid program by simultaneously increasing the number of people it covered but reducing the number of services that were insured. The services included were to be based on an explicitly prioritized list after extensive consultation with the public. A second experiment occurred in 2008 when 10 000 Oregon residents were randomly drawn from the non-Medicaid population to join the scheme. This presented an opportunity for exposing the benefits for these poor residents of membership. While the results did not show health improvements, they did show an increased use of preventive care, increased outpatient visits and increased utilization of health services. A third Oregon Medicaid experiment is still (2013) under way: A reorganization that integrates community services and clinical health care through 'Coordinated Care Organizations', which seek to address the underlying determinants of ill-health in the community.

Prospective payment A method of reimbursing health service providers (especially hospitals) by establishing rates of payment in advance, which are paid regardless of the costs in actual individual cases.

Skimping It refers to providing less intensive or lower-quality care than that specified in some standard or protocol in order to reduce costs in relation to the reimbursement due to the provider.

Introduction

In the presence of health insurance and limited capacity, an excess demand for services remains a permanent feature of several publicly funded health systems. The demand for health care needs therefore to be rationed in one way or another. This article describes three different common types of demand rationing. It distinguishes between (1) direct rationing, (2) rationing by waiting time and quality, and (3) price rationing.

Direct rationing refers to allocation mechanisms which explicitly rule out the provision of certain types of care to patients within the public sector (either by the rules set by the public insurer and/or because the doctor explicitly tells the patient when they demand care).

Rationing by waiting or by quality refers to allocation mechanisms which are implicit: the patient is not explicitly refused care. It is instead the presence of waiting times or low quality of care (either clinical or nonclinical), which induces some patients not to seek care from the public sector and either to opt for no care or care delivered in the private sector.

Finally, the article discusses price rationing, which in publicly-funded health systems takes the form of copayments or coinsurance rates for specific types of care. The article discusses each of these three types of rationing in turn. Although each is discussed in isolation, in practice these coexist in many health systems. The focus is on publicly funded health systems. Demand rationing within private health insurance markets is not discussed.

Direct Rationing

In the presence of complete coverage under public health insurance, patients could potentially demand treatment up to the point where the marginal benefit is zero. In a system with no capacity constraints, this would induce excessive consumption, the well-known issue of 'ex-post moral hazard.' However, most countries do limit the supply of care to a level below the one required to satisfy all potential demand, which results in an excess demand. The demand for health care has to be rationed in one way or another.

The capacity constraint is set by policymakers who decide on the number of hospital beds and doctors working in the health sector. The supply of health care can vary significantly across countries (and even within countries) and so does the size of the excess demand. Lower supply levels will require more demand rationing. Organization for Economic Co-Operation and Development countries vary significantly in their health expenditure per capita, with National Health Service (NHS)-type systems spending less than public insurance ones. Rationing is therefore more prominent in NHS-type systems.

From an efficiency point of view, in the presence of a capacity constraint, a natural way to manage the excess demand is to allocate care according to the highest benefit–cost ratio. Policymakers could rank all possible health treatments by their benefit–cost ratio and assign care to patients in descending order until the capacity is exhausted. To some extent this is in line with how governments operate. Treatments that are perceived or shown to have low benefit–cost ratio are not available within the public-insurance package. This is the case for some type of dental and ophthalmological care, plastic surgery, physiotherapy, or alternative medicine. However, listing all possible treatments to which patients are entitled and to compare them on the basis of benefit–cost ratios would be very costly. Public agencies (like National Institute for Health and Care Excellence) do increasingly encourage an evidence-based approach to resource allocation (Drummond *et al.*, 2005), but these still cover only a selection of treatments. One attempt to rank all possible treatments on the basis of cost-effectiveness criteria is the Oregon Experiment, which shows how drawing such comprehensive lists may generate surprising and counterintuitive results (Tengs, 1996).

Moreover, even if policymakers can exclude certain treatments (rationing 'across' treatments), it is optimal for governments to pursue rationing also 'within' a given treatment. Consider, for example, all patients who could benefit from hip replacement. For some patients costs will exceed benefits: these patients should therefore be optimally rationed. Governments could provide a detailed description for each treatment of the criteria to be used by the provider to ration care. These could be based on severity, patients' characteristics, pain, and overall health status. Providing such detailed descriptions for each treatment would again be very costly. Although governments do provide guidelines for some treatments, these are unlikely to be comprehensive.

The difficulty for governments in designing detailed rationing rules both 'across' and 'within' treatments suggests that doctors will play a critical role not only in providing health care but also in rationing care to patients. It is they who ultimately decide who should receive treatment. The rationing function is indeed (implicitly or explicitly) delegated to doctors in most countries. Doctors therefore act as agents on behalf of governments in implementing optimal rationing rules (McGuire, 2000). Governments can (and indeed do) outline the basic principles to health care entitlement and let doctors implement them. Below, such basic principles are first discussed and, then, the conditions under which doctors act as (im)perfect agents and how different incentive schemes may affect rationing.

Policymakers in several countries often state the general principle that access to care in publicly funded systems should based on need. The word 'need' can be subject to different interpretations. Need could refer to current health status of the patient, expected benefit from treatment, or patient's severity (Wagstaff and van Dooerslaer, 2000). For some types of care, these criteria go hand in hand if patients with worst health (higher severity) also have high expected benefits. But this may not be case for other types of care where patients with worst health (higher severity) have low expected benefits (as for some types of cancer care). In the first instance, it is high-severity/benefit patients who are likely to get the treatment. In the second one, a tension between equity and efficiency arises (Hauck *et al.*, 2002). On efficiency grounds (health maximization), patients with higher benefit should receive the treatment. On equity grounds, the priority may be reversed and patients with higher severity (and worst health) should receive the treatment. It is also worth emphasizing that the principle that access should be based on need offers little guidance on how costs should be taken into account. If patients with higher need are very costly, it does not necessarily follow that those patients should receive the treatment. In summary, general principles are useful but they are open to different interpretations by providers. A large body of the empirical literature suggests that the amount of care offered can differ to a great extent among doctors (Phelps, 2000).

Even if clear allocation rules could be established by policymakers, doctors would have an incentive to implement them only if they act as perfect agents on behalf of the government. Doctors' behavior may be affected by their own preferences and the way they are paid. For example, salaried hospital doctors may put considerable weight on a patient's benefit as opposed to costs. In terms of hospital payment, fixed budgets give strongest incentive to ration because extra patients do not generate additional revenues. Prospective payment systems of the diagnosis-related groups (DRG) type give weaker incentives to ration because additional patients increase revenues. Providers however may still have an incentive to ration high-cost and unprofitable patients (Ellis, 1998). Generous fee-for-service (FFS) systems or cost-reimbursement rules give weakest incentive to ration patients. The incentive to ration depends also critically on the degree of altruism of the doctors, highly altruistic doctors being more reluctant to ration patients. Altruism, therefore, plays a critical role in determining the design of incentive schemes (Ellis and McGuire, 1986; Chalkley and Malcomson, 1998): higher altruism typically requires lower powered incentive schemes because highly altruistic doctors paid with FFS arrangements are unlikely to exert much rationing.

In summary, direct rationing (explicit refusal of treatment to a patient) is a pervasive feature of many health systems.

Rationing occurs both 'within' and 'across' treatments. Some rationing is implemented through government allocation rules, which define the 'package' covered by the public sector. However, most rationing is exerted by the doctors and will be based on their preferences and their payment schemes.

A key issue in devolving the rationing role to doctors, is that 'turning patients down' can be an unpleasant activity. Without clear rules, doctors may feel reluctant to refuse treatment to patients with positive benefits. This will be exacerbated when the capacity constraint is tighter. Moreover, it is the doctors who will need to explain to patients why they are not offered treatment. They may also be liable for taking an unfair or unjust decision, and they may be at risk of breaking laws, which prohibit discrimination among patients.

If doctors are reluctant to explicitly ration patients, they may instead add the patients to a waiting list so that patients will have to wait before they receive treatment. This will generate a different type of rationing, which is described in Rationing by Waiting Times and Quality.

Rationing by Waiting Times and Quality

If direct rationing is difficult to implement, other forms of rationing are needed to allocate the limited capacity. If doctors simply refer for treatment all patients who could benefit (because they are reluctant to turn down patients with low benefit), a waiting list of patients authorized to receive care is likely to build up. If capacity is well below demand, patients' wait could be long. Demand may exceed capacity in every period, implying that in the absence of some way of limiting waiting lists, the length of the list could grow indefinitely. Waiting times may have a rationing effect if they dissuade some patients to seek treatment. Instead of waiting, the patient may opt for the private sector. Longer wait times will discourage more patients. Another possibility is that while waiting the patient may become unfit for surgery or recover. The empirical evidence from the UK suggests that waiting times do act as a rationing device to equilibrate demand and supply. Most empirical studies find that demand for care is inelastic, and that the elasticity is approximately -0.1: a 10% increase in waiting times reduces demand by only 1% (Martin and Smith, 2003; Iversen and Siciliani, 2011 for a review). The result that demand is inelastic implies that waiting times exert only a moderate rationing effect and that low levels of supply will translate into long waiting times.

Although waiting times eliminate the need for doctors to explicitly refuse treatment to patients, rationing by waiting times can be an inefficient form of rationing compared to direct rationing. Long waiting times impose a cost on patients, which is not necessarily recovered by anyone else (Gravelle and Siciliani, 2008). Some (short) waiting times may generate efficiency savings if they reduce the chance of idle capacity (i.e., the probability that supply is not used; Iversen, 1997). However, these efficiencies are fully exploited with short average wait times (Siciliani *et al.*, 2009). Waiting times are in the order of months for many procedures, well beyond those required for such efficiency savings to arise.

To mitigate the cost for the patients generated by waiting, several health systems prioritize patients on the list so that more severe patients wait less than less severe one. This is often done informally by doctors. Some governments have further reinforced this idea by developing policies (mostly in the form of guidelines), which encourage doctors to prioritize patients through the use of formal scoring systems (patients who score higher points in terms of severity, pain, and need wait less).

Other governments have discouraged the use of waiting time rationing through the development of targets or maximum waiting time guarantees, which introduce penalties for hospitals having many patients waiting for a long time. These policies may encourage increases in productivity but may also induce a switch from waiting time rationing to direct rationing.

Waiting times are not the only factor which induces patients to opt for the private sector. Other factors, like amenities and quality of care, also contribute to the choice between the public and the private sector. Besley and Coate (1991) provide a theory that suggests that a government (which has constraints on distributional tools) may find it optimal to distort quality downward to encourage the richer subset of the population to opt for the private sector. The shift of the rich to the public sector helps to bring the demand in the public sector in equilibrium with the limited supply as well as to redistribute income. This theory seems consistent with casual observation that public hospitals offer lower amenities compared to private one (patients may need to share rooms, have less privacy, and overall less comfort). This is not necessarily the case for clinical quality: whether it is higher or lower in public hospitals is less straightforward. On one hand, improving clinical quality is at the core of policy efforts in many publicly funded systems. On the other hand, public hospitals do face large demands, which induce hospitals to keep length of stay to a minimum, a dimension of lower quality (Barros and Siciliani, 2011).

As for direct (explicit) rationing, also under nonprice (waiting time and quality) rationing, the payment rule for providers may affect the incentive to vary quality and waiting times. For example, a DRG-type hospital payment system may induce creaming or skimping, i.e., the incentive to raise quality for profitable patients and reduce quality for unprofitable ones (Ellis, 1998). DRG systems also induce providers to treat additional patients, which should translate into lower waiting times.

In summary, rationing by waiting and quality is also a pervasive feature of many publicly funded health systems. Compared to direct rationing, these are (to some degree) inefficient because for a given capacity they reduce patients' welfare. However, they release doctors from the responsibility of directly rationing patients.

Price Rationing

Demand for health care can also be rationed through prices. In many publicly funded systems, this takes usually the form of a copayment or a coinsurance rate: the patient is asked to pay a fee or a proportion of the medical expenses. The idea is to make the patient cost conscious, who in turn demand less health care and in this way contain moral hazard (excessive consumption).

Countries vary in the use and design of copayments with some countries making more use than others. With few exceptions copayments remain low in publicly funded systems. Large increase in copayments would also imply a significant reduction in the benefit from being insured against the cost of illness. According to theory, the optimal copayment should be designed such that it efficiently trades off the risk spreading benefits of a lower price against the ex-post efficiency benefits of a higher price (closer to marginal cost; Zeckhauser, 1970). The theory implies that the optimal copayment is positively related to the elasticity of demand: copayments should be higher when the elasticity is higher. Copayments are indeed observed for dental care, ophthalmology care, and drugs where the elasticity is arguably larger. They are more rarely observed for inpatient or surgical care, which is free of charge in most (though not all) countries, due plausibly to its more inelastic demand.

Several empirical studies have estimated the elasticity of demand. Early studies found a wide range of elasticities' estimates, which could vary between -0.1 and -2.1 (Cutler and Zeckhauser, 2000). These estimates are potentially affected by selection bias if sicker individuals choose insurance plans with lower copayments and also demand more care. The Rand experiment eliminates such potential bias by randomizing individuals in different plans (Manning, *et al.*, 1987). It suggests that demand is inelastic with an overall elasticity of approximately -0.1 or -0.2. The article by Sinaiko (2012) updates this literature.

The idea that copayments make patients cost conscious relies on the belief that patients are able to influence the choice on the care. Given the asymmetry of information, which characterizes the patient–doctor relationship, the choice of care is (to say the least) mediated through the doctor. Copayments may have little impact on demand of care if doctors base their recommendation only on medical ground and ignore the financial implications for the patient. Copayments will instead play a role only if patients are well informed (which may be the case for some conditions) or if doctors internalize patients' disutility from higher prices. Behavior of the doctors will also be influenced by their financial incentives.

Both waiting times and copayments ration demand. Compared to rationing by waiting, copayments have the advantage that the cost imposed on the patient is recovered by the provider or the insurer in the form of additional revenues. Furthermore, if consumers accurately appreciate the value of health care, a price rations out low-value uses. However, copayments raise equity issues (if they are not income tested) if poor patients are deterred from demanding care compared to richer patients, a criticism that is often raised against their excessive use within publicly funded systems. Finally, the role played by copayments in rationing demand is mitigated by the asymmetry of information, which characterizes the patient–doctor relationship.

Conclusions

Three different types of rationing have been discussed: (1) direct rationing; (2) rationing by waiting time and quality; and (3) price rationing. Direct rationing can in principle allocate care efficiently by giving care to patients with highest benefit–cost ratio compatibly with the capacity constraint. This could be implemented by listing explicitly the treatments covered by public insurance (as well as those not covered) and/or by asking doctors to ration according to a set of established optimality criteria (potential health gains, heath status, and costs).

In practice, drawing an explicit list of eligible treatments is a complex and costly exercise and there are limits to this approach. Delegating the rationing role to doctors is inevitable. However, doctors may vary in the application of a set of rationing rules due, for example, to different interpretation of such rules, generating variations in clinical practice. Moreover, doctors themselves may be reluctant to exercise to a great extent the rationing role because they may find difficult or unpleasant to turn patients down and they may also be held liable for mistakes. This generates the scope for other forms of rationing.

If all potential patients who demand care are referred for treatment, a waiting list will quickly build up with (implicit) waiting time rationing replacing explicit direct rationing. Waiting times may ration public patients by inducing some of them to seek care in the private sector. Similarly, offering limited amenities in the public sector may shift some patients to the private sector. Compared with direct rationing, these forms of implicit rationing will however come at the cost of lower welfare for those patients who seek care in the public sector due to the waiting time or other costs imposed to them.

An alternative to rationing by waiting is to ration by price through the introduction of copayments or coinsurance rates. Its use can be a useful complement to other forms of rationing but needs to be traded off with the lower insurance coverage against the cost of illness. Moreover, large copayments (if not income tested) raise equity issues because poor patients may be discouraged from utilizing care. This can contrast the principle that access to care should not be based on the ability to pay, which is at the core of many publicly funded health systems.

See also: Demand for and Welfare Implications of Health Insurance, Theory of. Health and Health Care, Need for. Moral Hazard. Physician-Induced Demand. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Specialists. Waiting Times

References

Barros, P. P. and Siciliani, L. (2011). Public-private interface. In Pauly, M., McGuire, T. and Barros, P. P. (eds.) *Handbook of health economics*, vol. 2, ch. 15, pp. 927–1002. Oxford, UK: Elsevier.

Besley, T. and Coate, S. (1991). Public provision of private goods and the redistribution of income. *American Economic Review* **81**, 979–984.

Chalkley, M. and Malcomson, J. (1998). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* **17**, 1–19.

Cutler, D. and Zeckhauser, R. (2000). The anatomy of health insurance. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, ch. 11, pp. 563–643. Amsterdam, The Netherlands: Elsevier.

Drummond, M. F., O'Brien, B. J., Schulper, M., Stoddart, G. L. and Torrance, G. W. (2005). *Methods for the economic evaluation of health care programmes*. Oxford, UK: Oxford University Press.

Ellis, R. P. (1998). Creaming, skimping, and dumping: Provider competition on the intensive and extensive margins. *Journal of Health Economics* **17**(5), 537–555.
Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement. *Journal of Health Economics* **5**, 129–151.
Gravelle, H. and Siciliani, L. (2008). Optimal quality, waits, and charges in health insurance. *Journal of Health Economics* **27**(3), 663–674.
Hauck, K., Shaw, R. and Smith, P. C. (2002). Reducing avoidable inequalities in health: A new criterion for setting health care capitation payments. *Health Economics* **11**(8), 667–677.
Iversen, T. (1997). The effect of private sector on the waiting time in a National Health Service. *Journal of Health Economics* **16**, 381–396.
Iversen, T. and Siciliani, L. (2011). Non-price rationing and waiting times. In Gleid, S. and Smith, P. C. (eds.) *Oxford handbook of health economics*, ch. 27, pp. 649–670. Oxford, UK: Oxford University Press.
Manning, W. G., Newhouse, J., Duan, N., et al. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review* **77**(3), 251–277.
Martin, S. and Smith, P. C. (2003). Using panel methods to model waiting times for National Health Service surgery. *Journal of the Royal Statistical Society* **166**(Part 2), 1–19.
McGuire, T. (2000). Physician agency. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1, ch. 9, pp. 461–536. Amsterdam, The Netherlands: Elsevier
Phelps, C. E. (2000). Information diffusion and best practice adoption. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, ch. 5, pp. 223–264. Amsterdam, The Netherlands: Elsevier.
Siciliani, L., Stanciole, A. and Jacobs, R. (2009). Do waiting times reduce costs? *Journal of Health Economics* **28**(4), 771–780.
Sinaiko, A. (2012). Studies of Demand Response Since the HIE, this *Encyclopedia*.
Tengs, T. (1996). An evaluation of Oregon's Medicaid rationing algorithms. *Health Economics* **5**(3), 171–181.
Wagstaff, A. and van Doorslaer, E. (2000). Equity in health care finance and delivery. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1, ch. 34, pp. 1803–1862. Amsterdam, The Netherlands: Elsevier.
Zeckhauser, R. (1970). Medical insurance: A case study of the trade off between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**, 10–26.

Regulation of Safety, Efficacy, and Quality

MK Olson, Tulane University, New Orleans, LA, USA

Glossary

Asymmetry of information A situation in which the parties to a transaction have different amounts or kinds of information as when, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances.

Consumer surplus The difference between what a consumer pays for a good or service and the maximum they would pay rather than go without it.

Cost–benefit analysis A form of economic evaluation by comparing the costs and the (money-valued) benefits of alternative courses of action.

Incidence In epidemiology, incidence is the number of new cases of a disease identified during a time period. In economics, incidence concerns who pays taxes and who bears various other costs, that is, who bears the ultimate distribution of the burden after all effects in the economy have been worked out.

Marginal benefit The increase in benefit, or sometimes willingness to pay, from a small increase in the rate of consumption or utilization of a service.

Present value The value at a particular point in time of a future flow of income, health, etc., normally discounted using an appropriate interest rate or rates.

Introduction

Pharmaceutical regulation is designed to ensure safety, efficacy, and quality of the drugs available to consumers. This is accomplished through a range of regulatory activities over the course of a drug's life cycle including premarket screening and evaluation of new pharmaceuticals, inspection of manufacturing facilities, regulation of drug labeling and promotional activities, and the postmarketing surveillance of drugs following approval. This regulation extends to drugs and biologics that are intended for use in the diagnosis, treatment, prevention, and cure of diseases.

The rationale for pharmaceutical regulation is imperfect or asymmetric information. Evidence about drug safety and efficacy is difficult to observe and evaluate. As a consequence, physicians and patients lack information about a drug's quality. Information about the benefits, risks, and overall performance of new drugs is critical for their safe and effective use. Without accurate information about drug quality, physicians and patients may make inappropriate drug choices and suffer negative health consequences. Asymmetric information could also lead some patients to avoid taking drugs for fear that they are of low quality. Because firms as the developers of new drugs potentially know more about their quality than physicians and patients, particularly before launch, it is possible that firms may underinvest in developing prelaunch information and/or quality may be lowered or cheapened without consumer's knowledge for economic gain. The adulteration of drug products and detrimental effects on public health were motivating forces for developing early pharmaceutical regulations. Pharmaceutical products and their effects have become increasingly complex over time and evaluating those effects requires special expertise. Expert agencies like the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) regulate the prelaunch data requirements and evaluate the evidence over the life of a drug, to reduce the effects of informational asymmetries and uncertainty in the pharmaceutical market.

Regulators strive to protect and advance the public health by pursuing two goals: ensuring drug safety and effectiveness; and facilitating access to medically useful drugs. The problem facing regulators is that efforts to achieve one goal may in fact reduce the chances of achieving the second goal. Stringent regulations designed to keep dangerous or ineffective drugs off the market may delay patient access to new medicines. Alternatively, efforts to improve drug access by lowering regulatory stringency and accelerating reviews could lead to the approval of some drugs that are either unsafe or ineffective. The trade-off between safety and access is a central one in the regulation of new pharmaceuticals. The challenge for pharmaceutical regulators is balancing an interest for safety and efficacy with an interest for timely access. Finding the right balance, however, requires regulators to weigh the costs of unsafe or ineffective drugs against the costs of delay in the approval of beneficial drugs.

This article begins with an overview of important pharmaceutical regulatory legislation affecting the market access of brand name drugs in the US. It next describes EMA regulation and discusses some similarities and differences between the US and EU systems. It then examines evidence about the cost trade-offs reflected in pharmaceutical policy and concludes with a discussion of recent reforms in FDA regulation.

Overview of Pharmaceutical Regulation

Drug tragedies and safety concerns have motivated increases in pharmaceutical regulations over time. In the US, in the early 1900s, adulterated food and drug products and mislabeled medicines led to the Pure Food and Drugs Act of 1906. This Act banned the sale of misbranded or adulterated drugs and prohibited false labeling. It also required an accurate listing of ingredients on the labels of so-called 'patent medicines,' which were generally unpatented remedies that were heavily advertised and promoted as cures, but did not work as promoted. Although the Act gave government the power to prosecute

 doi:10.1016/B978-0-12-375678-7.01202-5

firms that violated the law, it did not require firms to test products before marketing.

In 1937, a liquid sulfa drug called Elixir Sulfanilamide, which contained a poisonous solvent, was sold to consumers and resulted in more than 100 deaths. This tragedy led to the 1938 Food, Drug, and Cosmetics Act, which required firms to test products for safety before marketing and file that information with the FDA. The burden of proof, however, lay with the agency to reject a drug to prevent it from being marketed. Drugs could be marketed automatically without approval after 60 days (in some cases 180 days) unless the FDA showed them to be unsafe. The Act also gave the FDA the power to regulate drug labels to ensure the accuracy of therapeutic claims and adequacy of instructions for safe use. However, enforcement was challenged by the absence of scientific evidence about drug effects relating to such claims. Because the Act did not require firms to demonstrate that drugs were effective, an important aspect of the informational problem facing consumers and physicians was not addressed.

This changed with the 1962 Drug Amendments to the Food, Drug, and Cosmetics Act. The 1962 Amendments were motivated by the worldwide thalidomide drug tragedy. In 1961, researchers discovered that thalidomide, which was being used as a sedative for pregnant women throughout Europe and in Canada, was responsible for thousands of severe birth defects in which infants were born with missing or truncated limbs. The drug was not being sold in the US because it was held up by an FDA reviewer as reports began to emerge. However, samples of the drug were distributed to US physicians for premarket study. Although Congress was conducting hearings into high prices for drugs of dubious efficacy before the scandal, the near miss of a US tragedy expanded their focus to include drug safety.

The 1962 Amendments required firms to provide 'substantial clinical evidence' of safety and effectiveness for all drugs before they could be marketed. The law eliminated the prospect of automatic approval by removing the time limit within which FDA was required to act to reject a new drug. This shifted the burden of proof to firms to obtain explicit FDA approval before entering the US market. The law also required firms to submit their plans for premarket clinical tests to the FDA and inform the FDA of any serious adverse drug reactions (ADRs) experienced by patients. The FDA was also given oversight of prescription drug advertising and promotion.

Many industrialized countries throughout the world responded to the thalidomide tragedy with increased safety standards and regulation for new drug approval. However, most European countries, with the exception of Norway and Sweden who already required evidence of drug effectiveness, did not establish a mandatory efficacy requirement for new drug approval until several years later. Another way that the US response differed from other countries is the complex set of procedures and requirements governing the premarket clinical development, evidence submission, manufacturing, and approval of new drugs. These procedures continue to be used today by the FDA in drug approval. There has also been a convergence in pharmaceutical regulation relating to a drug's clinical development and good manufacturing processes in recent years as a consequence of efforts toward international harmonization.

The process begins as firms conduct preclinical studies and testing in animals to determine which drug compounds offer promise and are reasonably safe for human testing. In the US, when firms are ready to begin clinical testing in humans, they must submit an Investigational New Drug (IND) application to the FDA. The IND contains the results from toxicology studies, animal studies, and other preclinical tests, plans for clinical testing in humans, details of the manufacturing process, and results from any clinical studies conducted outside the US. The IND allows FDA to provide input into planned clinical tests or reject poorly formulated IND applications. An IND application becomes effective after 30 days unless the FDA places a hold on it.

Once an IND is effective, firms may begin premarket clinical testing with human subjects, which proceeds in three phases. Phase I studies test for drug safety in a small number of healthy human subjects (20–80 subjects). Phase II studies are controlled clinical tests in a small number of human subjects with the target disease to develop preliminary data on effectiveness. If the data from the Phase II trials suggest effectiveness, the firm may begin Phase III studies, which are large-scale, placebo-controlled randomized clinical studies used to confirm a drug's effectiveness (600–3000 subjects on average). In cases where it may be unethical to put clinical trial patients on a placebo, Phase III studies compare the target drug to the current treatment. Clinical studies can take between 2 and 10 years to complete, during which firms collect the data and evidence from these studies. These data, along with samples of the drug, are submitted to the FDA in a New Drug Application (NDA), which includes all the data and results from all clinical and preclinical studies. After accepting an NDA, regulators commence review to analyze the evidence and determine if the drug's benefits exceed its risks. The FDA also negotiates with the manufacturer over the drug's labeling, reviews promotional materials, assesses the need for postmarketing study requirements (i.e., Phase IV studies), and inspects the firm's manufacturing facilities. Phase IV studies expand the safety and efficacy profile of selected drugs in a large population following drug launch and can include formal therapeutic trials or comparisons with existing medicines. Phase IV confirmatory studies are required for drugs approved with limited exposures and/or surrogate endpoints, which are markers (i.e., blood pressure, CD4 cell count, tumor shrinkage) used in clinical trials as an indirect measure of clinically meaningful endpoints (i.e., longer survival or reduced symptoms). Phase IV studies are also sought where questions arise about whether serious adverse events can be attributed to a drug.

To satisfy the requirement for 'substantial evidence' of safety and efficacy, the FDA has generally required two large-scale, controlled clinical studies to demonstrate a drug's effectiveness. The rationale is that the results of any single controlled clinical study must be confirmed by a second such study sufficiently powered to meet the agency's evidentiary standard. This requirement for two Phase III trials, sometimes referred to as the 'gold standard' for approval, has been a subject of debate and controversy over time. Although the requirement has been waived in some cases in recent years, it has greatly added to the cost and length of the FDA approval process. In addition, FDA is the only regulatory authority in

the world that requires companies to submit the raw data collected in clinical trials to assure that results reported by the firm can be replicated from the FDA's own independent analyses of the data. These requirements have increased the size of NDAs as well as the time and effort required to review them.

In the late 1980s and early 1990s, the crisis surrounding Acquired Immunodeficiency Syndrome (AIDS) and growing awareness of delays in the FDA's drug approval process led to new reforms to improve patients' access to new drugs. FDA introduced treatment INDs and a parallel-track program to give patients with life-threatening diseases early access to promising investigational therapies before approval. The FDA also relaxed certain statutory requirements in the testing of AIDS drugs. For instance, Subpart E procedures allowed for the elimination of Phase III studies before approval, whereas accelerated approval permitted sponsors to use surrogate endpoints in clinical studies other than survival or morbidity to demonstrate efficacy. The agency's success in accelerating access to new AIDS drugs led to increased public and industry pressure to reduce regulatory delays for other drugs. FDA argued that they needed more resources to combat delays and Congress responded with new legislation, which introduced industry funding for new drug review.

The 1992 Prescription Drug User Fee Act (PDUFA) initiated the program whereby drug manufacturers ('sponsors') are required to pay fees to the FDA to help finance its review of NDAs. In return, firms receive agency commitments to expedite the review of NDAs and meet a series of performance goals. Two key goals specify review targets for priority-rated drugs, which offer a significant therapeutic advance over existing remedies, and for standard-rated drugs, which offer little to no therapeutic gain over existing remedies. The goals stated that FDA should review and act on 90% of priority drug applications in 6 months and 90% of standard drug applications in 12 months. Another goal directed FDA to eliminate its backlog of NDAs awaiting approval within 24 months of the establishment of the program. Before PDUFA, the average review time was approximately 30 months. To increase accountability, the FDA was required to report annually to Congress on the status of meeting its performance goals. PDUFA was enacted with a 5-year term and its renewal requires new legislation to extend its term. These features allowed stakeholders to assess the agency's performance before renewal and make necessary adjustments in the program. One interesting provision in PDUFA, which became a target for later reform, was that fee revenues could only be used for efforts to accelerate drug reviews and could not be used for other purposes, such as postmarketing drug safety surveillance.

Drug review times began to decline under PDUFA, drug approvals increased, and the program was renewed in the 1997 FDA Modernization Act (PDUFA II) for another 5 years. PDUFA II increased user fees, raised agency revenue targets, and lowered the review deadline for standard-rated drugs from 12 to 10 months. New performance goals were also added to ensure the timeliness of communications between the FDA and sponsors during the clinical development period to further speed-up the process. PDUFA II included timelines for the FDA to schedule sponsor-requested meetings before the submission of INDs or NDAs, resolve disputes with sponsors which arise during clinical development, respond to sponsor questions about study protocols, and develop guidances for industry. To further combat delays in drug development, PDUFA II included a provision allowing the agency to accept a single, large-scale, controlled clinical study as 'substantial evidence' of effectiveness (instead of two), but discretion remained with agency to make that determination. The 2002 Public Health Security and Bioterrorism Preparedness and Response Act (PDUFA III) reauthorized the program, raised fees and revenue targets, and further expanded the FDA's interactions and communication with firms during testing phases and the review cycle.

Drug industry user fees account for 65% of the agency's human drug budget and the time required for the FDA approval phase has been cut by nearly 60% since the enactment of PDUFA. Kaitin and Cairns (2003) report that clinical development times have also declined from a high of 7.2 years in 1993–95 to 5.5 years in 1999–2001 since PDUFA II. However, a series of safety-related drug incidents culminating in the 2004 withdrawal of Vioxx – a Cox-2 inhibitor allegedly linked to thousands of deaths from heart attacks or strokes – raised concerns about the effects of PDUFA on drug safety and the agency's handling of postlaunch drug safety issues. Questions arose about whether the FDA's approval process had become so accelerated that adequate attention was not being given to drug safety issues, especially postlaunch. Questions also arose about the timeliness and effectiveness of risk communications to the public, the transparency of agency decision making, the handling of internal agency conflicts over approval decisions, and the failures by firms to complete required postmarketing safety studies.

Congress responded to these and other safety concerns with the 2007 FDA Amendments Act (PDUFA IV), which renewed the user fee program and provided the agency with new authorities to address drug safety problems that arise after approval. The Act gave the FDA authority to require Phase IV postmarketing studies and clinical trials to address important drug safety questions and power to fine firms that did not complete their studies. It gave the agency new authority to require safety labeling changes (such as new black box warnings) when new serious risks emerged after approval, and the authority to require drug or biologic developers to submit Risk Evaluation and Mitigation Strategies (REMS) as part of a NDA (or for an already approved drug) when deemed necessary by FDA to ensure that the benefits of a drug outweighed its risks. These and other recent FDA reforms will be discussed in detail in the final section.

The European Medicines Agency

By the 1980s, most industrialized countries adopted some form of efficacy regulation in response to increases in the number, complexity, and toxicity of new pharmaceuticals. However, a range of different economic, political, social, and cultural factors in countries resulted in a proliferation of national drug regulatory systems and processes for marketing authorization. Concern about the fragmented regulatory systems in the European Community, uneven drug access among EU countries, and a desire for a single integrated European

market to facilitate the free movement of new medicines led to new policies to standardize the way in which new drugs and biologics were approved and marketed in the EU.

In 1993, the European Commission created the EMA to authorize/approve new drugs and biologics in the EU. The EMA began evaluating new products in 1995 through a centralized procedure in which a drug receives a single marketing authorization valid in all EU countries. The centralized procedure is compulsory for new biotechnology and advanced therapy medicines (i.e., gene therapy) as well as new drugs to treat AIDS, cancer, neurodegenerative disorder, diabetes, autoimmune diseases, viral diseases, and orphan diseases. Firms have the option to use the centralized procedure for other new drugs that offer a significant therapeutic benefit or otherwise serve the public interest. The process has strict deadlines for timely reviews, which include a limit of 210 days for the scientific committees to evaluate drug applications and reach a decision, 30 days for EMA to finalize and then transmit that decision to the European Commission, who grants the marketing authorization (within 90 days).

In a second track, firms can seek approval in a limited number markets through a decentralized procedure based on a process of mutual recognition. Under this track, the EMA receives a NDA from a firm, but then forwards it to a single member country recommended by the firm for review. Following approval, that country refers the drug application to other member countries designated by the firm. If mutual recognition or approval is not granted, firms can pursue arbitration through the EMA.

Both the EMA and FDA share common objectives to protect public health by ensuring the safety, efficacy, and quality of new medicines. Both agencies have special procedures to facilitate the approval of orphan medicines and the accelerated approval of medicines for life-threatening illnesses with few therapeutic options. Both the agencies are also funded by industry user fees. There has been increasing communications, information sharing, and cooperation between the EMA and the FDA over time as well as increasing convergence in regulatory procedures. Along with their Japanese counterparts, the US and EU agencies have worked together with drug firms toward the global harmonization and improvement of international drug regulations through the International Harmonization Conference (ICH).

There remain interesting differences between the EMA and the FDA. (1) Marketing authorization in the EU is valid for 5 years, whereas FDA approval allows firms to market drugs in the US indefinitely. The EU requires sponsors to submit a reevaluation of the risk–benefit balance after 5 years to renew a product's marketing authorization. (2) Marketing authorization is not the final hurdle in the EU because the firm must negotiate a drug's price and reimbursement with the government before marketing. Unlike the US, which does not negotiate drug prices with firms, price negotiations in EU countries can further delay the entry of new drugs into the market (3) The EMA typically requires comparator-controlled clinical trials or three-arm studies, which compare the target drug to a comparator drug and placebo, whereas the FDA is satisfied with placebo control except when it is ethically unfeasible. This amounts to a regulatory standard of comparative efficacy for EU approval and provides one reason why the two agencies might reach different approval decisions or different risk–benefit conclusions for the same drug. (4) Assessments in the FDA are conducted by its own reviewers in a single agency, whereas assessments in the EMA are conducted by the national agencies in different Member States. Unlike the FDA, the EMA is predominantly a coordinating office that depends on scientific input from a large network outside experts from different Member States who participate on its scientific committees. This suggests that differences in culture and practices among the Member States may potentially influence drug authorization decisions. (5) Negative approval decisions along with the accompanying product assessments are published in the EU in contrast to the US where such information is considered proprietary for drugs that are not approved by the FDA. (6) The FDA is responsible for enforcement of its policies, whereas enforcement of marketing authorizations, licensing, control sales, and promotional activities in the EU is left to the Member States, which could potentially create problems of coordination with the EMA. Still EMA has historically possessed more authority to deal with drug safety issues in the postmarket, including the power to suspend marketing authorization while drug safety issues are investigated, and it has access to a wider range of financial penalties for issues of noncompliance with EU requirements.

Evidence about the Effects of Pharmaceutical Regulation

Evaluating pharmaceutical regulation requires evidence about the cost trade-offs of unsafe or ineffective drugs versus delay in the approval of useful drugs. The social costs of approving unsafe or ineffective drugs include deaths or reductions in health experienced by patients exposed to these drugs, including the opportunity costs of wasted time and resources in the ineffective fight against diseases. The incidence of some of these costs may fall on payers/citizens/taxpayers who ultimately pay for medical care as well as firms who pay liability costs. The social costs of delay or increased requirements for drug approval include the health benefits foregone by patients from delayed treatment and any increase in drug development costs due to increased regulation.

Few studies have examined the health effects of pharmaceutical regulation. The social costs of unsafe drugs become visible when drug-related tragedies occur in which patients are harmed. The social costs in terms of forgone health benefits of approving ineffective drugs are also difficult to observe and assess. With little systematic evidence about the health effects of unsafe or ineffective drugs, highly visible drug tragedies have played an important role in shaping pharmaceutical policy and regulator behavior over time. Quantifying the social value of drugs delayed or prevented from the market by pharmaceutical regulation has also proved challenging. Little research has attempted to estimate the gains in patient health arising from faster access to a new drug instead of an older one.

Much of the early research focused on the extent to which regulation affects the number of new drug approvals, the costs and length of drug development, and regulatory delays in the FDA review, with particular emphasis on the effects of the

1962 Amendments. These Amendments constituted the largest single change in the US regulatory policy and hence provide a natural experiment to examine the effects of pharmaceutical regulation.

Studies showed that the time spent on testing new drugs increased substantially after the 1962 Amendments. Wardell *et al.* (1982) found that the period of preclinical and clinical testing increased from 30 months in 1960 to 100 months in 1970 to 120 months in 1980. The cost of developing new drugs also increased since 1962. Hansen (1979) who examined 1963–75 drug approvals estimated a total capitalized cost of US$54 million (in 1976 dollars) per new drug approval. DiMasi *et al.* (1991) who examined 1970–82 drug approvals estimated a cost of US$231 million (in 1987 dollars) per new drug approval. Using more recent drug approvals in 1989–2001, DiMasi *et al.* (2003) estimated a total cost of US$802 million (in 2000 dollars). These studies are also reviewed in this volume. Studies also suggested that the 1962 legislation resulted in a substantial reduction in the number of US drug approvals, but some questioned the relative role of regulation versus other factors in explaining that reduction.

Peltzman (1973) argued that all decline in US new drug approvals following 1962 was due to the new regulations, whereas others, such as Grabowski *et al.* (1978) argued that regulation could explain roughly half of the decline with the rest attributed to other factors, such as the depletion of research opportunities and increased industry restraint after the thalidomide tragedy. Temin (1980) and Wiggins (1984) found that much of the decline in approvals occurred in a few therapeutic areas (i.e., central nervous system tranquilizers, antiinfectives) and concluded regulation may be even less important than other market factors in explaining the trend. Temin further noted that demand as well as supply of tranquilizers diminished after the public became aware of the thalidomide tragedy. Because the purpose of the 1962 Amendments was to prevent dangerous or ineffective drugs from reaching the market, more evidence is needed to conclude that fewer new drug approvals necessarily reduced social welfare.

In one of the first cost–benefit studies of the FDA regulation, Peltzman (1973) used the growth in market shares for new drugs approved before and after the 1962 law to estimate the gains in consumer welfare from the efficacy requirement and compared these to the losses of welfare from reduced drug approvals. He concluded that gains were far smaller than the costs of the law. His estimates have been the subject of much debate and criticism. Peltzman's analysis did not consider the welfare benefits from a possible reduction in unsafe drugs. Temin (1980) argued that Peltzman's estimate of the benefits of the efficacy requirement was not fully captured in the analysis, whereas the costs were overestimated because all of the observed reduction in drug approvals was attributed to the Amendments.

It is difficult to measure the social value of the drugs that were delayed by more stringent regulation. Data from the FDA showed that much of the decline in approvals following 1962 occurred among drugs offering little to no therapeutic gain over existing remedies, known as me-too drugs, whereas other data showed little decline among drug approvals offering important therapeutic gains. Abraham and Davis (2005) showed that drug withdrawal rates were twice as high in the UK as in the US between 1971 and 1992. They argued that more stringent premarket review and testing standards in the US prevented the approval of some unsafe drugs compared to the UK and showed that US regulators had identified the same safety problems upon which drug withdrawal ultimately occurred in the UK. Although the evidence is compelling that FDA regulation prevented the approval of some dangerous drugs, questions remained about the social costs arising from regulatory delays of potentially useful new drugs.

More recent research has examined the effects of PDUFA, which provides another natural experiment to examine how increased drug review speed has affected drug safety and social welfare. Several studies have documented the increase in drug review speed and approval observed under PDUFA. Other research has investigated whether PDUFA and increased drug review speed led to reductions in drug safety.

Measuring drug safety has posed a challenge in research. Some studies investigated whether the rate of drug withdrawals changed under PDUFA. Friedman *et al.* (1999) compared the rates of drug withdrawals by approval year between 1970 and 1999 and found little difference before and after PDUFA. A study by the General Accounting Office (2002) examined the rates of safety-related drug withdrawals between 1985 and 2000 and found that drug withdrawal rates increased under PDUFA. Berndt *et al.* (2005) found no significant difference in drug withdrawal rates pre- and post-PDUFA, but they note results are sensitive to the time periods selected and potential problem of censoring. They also suggest that as safety-related drug withdrawals are relatively rare, it is inherently difficult to detect significant differences in these rates pre- and post-PDUFA. Some have noted that studies focusing on drug withdrawals provide little information about the possible effects of the reform on the safety or risks of drugs that remain on the market.

Other research has examined the rates of black box warnings given to the FDA-approved drugs before and after PDUFA to determine if drug risks have increased. In an National Bureau of Economic Research working paper, Begosh *et al.* (2006) found no significant difference pre- and post-PDUFA in the rate of new black box warnings received after a drug's approval. Carpenter *et al.* (2008a,b), however, found that the new drugs approved before their PDUFA review deadlines had a higher probability of being withdrawn and receiving a postapproval black box warning compared to other drug approvals. They conclude that PDUFA deadline pressures may in some cases compromise drug safety. Although more frequent than drug withdrawals, black box warnings have until very recently been discretionary FDA actions that must be negotiated with firms. Legislation passed in 2007 first gave the FDA statutory authority to require such warnings after approval.

ADR data from the FDA have also been used in studies to investigate the effects of faster FDA drug review times on drug safety. Most ADR reports are submitted by health professionals when patients experience serious adverse reactions to drugs, but anyone can submit an ADR report. ADRs are much more frequent than drug withdrawals or boxed warnings and they provide signals of potential drug safety problems. However, ADRs are generally underreported and the reports do not include evidence of causation. Olson (2002) examined drug

approvals in 1990–95 and found that faster reviews were significantly associated with increased counts of serious ADRs. A subsequent study by Olson (2008), which included more data and additional controls, also found that faster drug review times were significantly associated with increased counts of serious ADRs among the 1990–2001 new drug approvals. Grabowski and Wang (2008) examined drug and biologic approvals in 1992–2002 found no significant effect of the FDA review speed on ADRs. However, Olson (2008) notes that Grabowski and Wang's study excluded drugs approved before the PDUFA, used three annual ADR counts for each drug instead of a single aggregate count, and included ADR reports listing secondary suspect drugs that were weighted equally with reports that listed a primary suspect drug, which Olson argues increases the noise in ADR count measures.

Olson (2004b) assesses the benefits and costs of therapeutically novel drugs, which the FDA targets for faster drug reviews. She found that therapeutically novel drugs had significantly more serious ADRs after approval, including drug reactions resulting in hospitalization and death. Because priority review status is granted primarily to drugs that are novel, there is an inevitable confounding of drug novelty with fast review. With that caveat and after controlling for a drug's review time, the results showed that for an average drug, novelty is associated with a 60% increase in serious ADRs, 45% increase in ADRs that require hospitalization, and 61% increase in ADRs that result in death in the first 2 years after approval. To measure the health benefits of novel drugs, Olson draws on Lichtenberg's (2005) estimate of the increase in life expectancy due to increases in the stock of priority (novel) drugs, 292 000 life years per year. The benefits of novel drugs are then compared to the life years lost from increased ADR deaths over the period. For drug approvals in 1990–95, and assuming no underreporting of ADRs, results showed that ADR deaths (in the first 2 years after approval) reduced the net longevity gains due to novel drug approvals by approximately 8%. This estimate is subject to two potential biases. First, because ADRs are underreported, the number of ADR deaths may be larger than predicted. For instance, Olson's study shows that 30% underreporting of ADR deaths reduces the estimate of net longevity gains of therapeutically novel drugs by 11%. Second, the life years gained from novel drugs may be overstated in Lichtenberg's study because he does not control for increases in the stock of medical devices and other changing health technologies that may have also increased longevity over time. Both biases would result in a reduction in estimated net longevity gains from novel drugs.

Philipson *et al.* (2008) assessed the benefits and costs of the PDUFA by comparing the increases in drug sales due to faster reviews under the PDUFA to ADR deaths reported for the PDUFA-related safety removals. The reform's benefits are measured as the present value of the increase in producer and consumer surplus due to increased review speed for all approved PDUFA drug submissions over each product's 15-year sales life cycle, whereas the reform's costs are measured as the value of the life years lost from reported ADR deaths among the subset of the PDUFA-related drug withdrawals. They examine data for all drug and biologic approvals in 1979–2002 to develop their estimates and use data for drug sales in 1998–2002 and IMS life cycle year to peak percentages from IMS Health to compute drug sales in all other years. They estimate the gains in producer surplus due to the PDUFA to be US$7–11 billion, and the gains in consumer surplus to be US$7–20 billion depending on assumptions about the amount of surplus captured by consumers during patent protection. However, this estimate assumes that the demand curve for drugs measures marginal benefit and hence consumer surplus, which is likely to lead to a large overstatement because it ignores the effect of insurance on drug prices. They also do not account for any offsets in drug sales from the substitution of newer drugs for older ones, which may further bias benefits upward. With these caveats, they estimate that the benefits from faster reviews are 140 000–310 000 life years gained (assuming the value of a life year is US$100 000), whereas the costs of ADR deaths among withdrawn PDUFA drugs are 56 000 life years lost and conclude that the benefits of PDUFA exceed the costs.

Philipson *et al.*'s cost estimates are subject to some potential biases. They assume that all reported ADR deaths among the subset of withdrawn PDUFA drugs are due to the reform and that this set of withdrawn drugs have no benefits, which they argue results in an extreme upper bound on costs. However, they do not include costs resulting from nonfatal hospitalizations associated with the withdrawn drugs, which 'if valued substantially, could potentially lead to offsets larger than the gains of greater speed induced by PDUFA.' They also do not include as part of the reform's costs any reductions in life years due to increased ADR risks among the drugs remaining on the market, which received faster reviews. Olson's (2008) study finds that a 19-month reduction in review time is associated with an increase of 11 ADR deaths in a drug's first 2 years on the market, which translates into an extra 131 501 life years lost among the drug approvals in 1990–2001. Both factors would increase the estimated costs of the reform.

Philipson *et al.*'s study may also not fully capture the dynamic welfare effects of faster drug reviews and increased producer surplus on pharmaceutical R&D and innovation, which would increase the estimated benefits of the reform. Using survey data from seven large pharmaceutical firms, Vernon *et al.* (2009) estimated that the PDUFA increased R&D spending by US$3.2–4.6 billion. By extrapolating to the entire industry, they predict that pharmaceutical industry R&D spending increased by US$10.8–15.4 billion from 1992 to 2002. If the additional R&D results in more drug innovation, then increased R&D could yield additional benefits of the reform. The authors note, however, that sample selection bias could be a serious issue in making industry-wide projections because of a low survey response rate and incomplete surveys. In addition, their study does not control for other important factors that could have increased R&D spending and incentives for R&D, such as the development of the EU's centralized process for drug approval and increasing competitive pressures.

Evidence about the welfare effects of pharmaceutical regulation has increased over time, but important challenges remain, particularly in the estimation of consumer benefits. Studies that use drug sales data to estimate the dollar magnitude of benefit to consumers are subject to some important limitations. Although drug sales are a good indicator of producer benefit, they convey little information about the actual

improvements in patient health arising from faster access to a new drug instead of an older one. Benefits among drugs that are ineffective, equivalent to existing drugs, or those that have serious side effects would be overestimated with this measure. Further, such studies ignore the important effect of third-party insurance on drug prices. With insurance, consumer demand will not reflect the marginal benefit to consumers and consequently estimated consumer welfare is likely to be overstated. *In lieu* of these limitations, future research may want to utilize measures such as life years or other health impacts of drugs to determine more reliable estimates of consumer benefits from reform.

Recent Reforms of Food and Drug Administration Regulation

There continues to be controversy and conflict over pharmaceutical regulation. In the US, drug-related tragedies and concerns about conflicts of interest in the FDA and in its advisory committees have led to increased public and political scrutiny of the agency in recent years. The withdrawal of Vioxx in 2004 and other controversies, which include those relating to pediatric antidepressants and suicide risks in children, called into question various aspects of the agency's oversight structure and processes and reduced the public's confidence in the FDA. In response to calls for reform, in 2005 the FDA asked the Institute of Medicine (IOM) to convene a committee of experts to conduct an assessment of the US drug safety system and make recommendations to improve risk assessment.

The Institute of Medicine Committee (2007) report identified important shortcomings in the US drug safety system including insufficient regulatory authority and tools for addressing drug safety problems that emerge in the postmarketing period after approval, chronic underfunding of postmarketing drug safety activities, organizational and coordination problems among pre- and postmarketing drug safety teams, limited postapproval drug safety data, insufficient monitoring and ineffective communication of new risk information in the postmarket, and insufficient transparency/public access to a drug's benefit and risk information. The committee found that these problems had become more pronounced over time in part due to the pressures created by the PDUFA; increases in the number, complexity, and potency of prescription drugs marketed; increases in drug utilization for chronic conditions; and other changes that affected the way in which drugs were being promoted, used, and prescribed. The report noted that as more people were taking more drugs over extended periods of time, many drug risks were not becoming known until well after approval. In response to these changes, the IOM committee recommended that the FDA be given new duties, authorities, and resources to ensure drug safety, especially in the postmarketing period, and that the agency develop a life cycle approach for the assessment of a drug's risks and benefits.

Many of the IOM report's recommendations were included in the 2007 FDA Amendments Act (PDUFA IV). This Act reauthorized the user fee program and included numerous provisions to strengthen and modernize the US drug safety system. It is interesting to note that the Act does not weaken the agency's policies for accelerating drug approval or improving drug access. Instead, the 2007 Act gave the agency new powers to identify and address drug safety problems that may emerge after approval including the power to require safety labeling changes, postmarketing studies, and Risk Evaluation and Management Strategies. Marking an important change from the past, the PDUFA IV also allowed the agency to more user fee revenues to build the agency's postmarketing drug safety activities and staff. The Act also contains provisions to increase the availability of risk–benefit information for the public, to improve communications of risk information to the public, and to increase transparency of its decision making.

More specifically, the 2007 Act provided the FDA with the new authority to require and enforce the Phase IV postmarketing studies and clinical trials. Before 2007, the FDA lacked such authority. Although the agency could request that firms conduct a Phase IV study as a condition of approval to address unanswered safety questions to help speed up a product to market, the FDA lacked the enforcement tools (other than the most extreme action of withdrawal of approval) to ensure the completion of such studies. Data showed that the completion rate for Phase IV studies of postapproval risks or efficacy studies was low. Among the 88 new molecular entities (NME) approved in 1990–94, Sasich *et al*. (2005) find that 87% had not completed their Phase IV studies 5–10 years after approval by 1999 and none of the 107 NME approved in 1995–99 had completed their Phase IV studies by December 1999. In September 2004, there were 1191 open postmarketing study commitments reported in the Federal Register (2005) of which only 18% were ongoing, whereas 68% were pending or not yet initiated. Failure to complete Phase IV studies, especially those studies agreed to as a condition of approval or accelerated approval, prevents patients and physicians from receiving necessary information about a drug's risk–benefit profile to make informed prescribing decisions. The 2007 Act allowed FDA to require such studies to identify or assess potential serious risks before or after approval and gave the agency power to levy new fines and penalties against firms who did not complete their study commitments. Firms could receive a civil penalty of US$250 000 for each violation and up to US$10 million for an ongoing violation. The FDA was also made responsible for tracking and reporting on the status of these studies annually to consumers in the Federal Register.

The 2007 Act also provided the FDA with new authority to manage risks among marketed drugs through REMS. The REMS are developed by manufacturers with the input of the FDA to ensure that the benefits of a drug or biological product outweigh its risks. They may be required at the time of approval or after approval if new risk information emerges. The REMS may take many forms, but some of the most common include plans to ensure appropriate risk communication (Medication Guides), safe use conditions, adequate prescriber education or training, adequate pharmacy education or certification, and required patient monitoring or patient registries. All REMS must have a timetable for the submission of assessments to determine if the plan has been effective. With the increase in REMS following the 2007 Act, concerns were voiced at the subsequent PDUFA stakeholders meetings about the lack of standardization of REMS.

Pharmacists sought greater involvement in the development of these plans and suggested that the growing number of unique REMS was placing burdens on the health care system. In response, the agency is reaching out to stakeholders and moving toward greater standardization of REMS.

The 2007 Act required FDA to improve its postapproval risk assessments and the timely communication of those assessments to the public. First, the agency must conduct regular, biweekly screening of the Adverse Event Reporting System (AERS) database and post a quarterly report on the AERS website of any new safety information or potential signal of serious risks identified through AERS. The first report was posted in September 2008. Second, the agency must conduct systematic evaluations of the safety of new drug approvals since September 2007, 18 months after approval or after a drug's use by 10 000 patients. These reviews draw upon a range of pre- and postapproval data including the drug's preapproval safety profile, new adverse event reports, other sources of new risk information, trends in drug utilization, etc. to analyze any potential new safety issues involving these drugs. Summaries of the agency's findings are posted on its website to help improve information about emerging drug risks to patients and physicians.

The 2007 Act required the FDA to increase active postmarket risk identification and analysis as a complement to its existing passive AERS. In 2008, the FDA launched the Sentinel program, which is a national electronic system for monitoring product safety that is linked to automated health care data systems, such as electronic health record systems, administrative and insurance claims databases, and registries. The law required the FDA to work with partners from public, academic, and private entities to develop this system. The FDA can use this system to make queries about potential drug safety issues, which can then be explored quickly and securely using these broad data networks. The law set a goal of being able to query data from 25 million patients by 1 July 2010 and data from 100 million patients by 1 July 2012. The agency reported that it met its data access goal for 2010 and is on target to meet the data access goal by 2012 deadline.

In an effort to improve public access to safety data for drugs involved in premarket studies, the 2007 Act requires that firms must register within 21 days of the enrollment of the first patient, all Phases II–IV drug trials in the publically available National Institutes of Health (NIH) online database, www.clinicaltrials.gov. The Act also mandates the creation of a clinical trials results database for approved drugs that, when fully implemented, will include a set of Internet links to key FDA documents, summary tables of primary and secondary outcomes, expanded results, and information about serious adverse events. Clinical trials results must be submitted within 1 year of the trial completion or within 30 days of drug approval. There are few studies of the overall compliance of firms with the new registration requirements although Lester and Godlew (2011) who examine evidence from NIH and FDA suggests that there is a high level of noncompliance.

The recent reforms of FDA regulation reflect a new awareness of the fact that even with stringent preapproval evaluation procedures some drugs that reach the market could potentially harm consumers. The provisions contained in the 2007 Act suggest that current policy makers are not willing to sacrifice the gains made in facilitating drug access to patients under PDUFA or under accelerated approval or fast-track programs to try to prevent future drug tragedies. Instead, policy makers have strengthened the FDA's authorities and resources for addressing drug safety problems that emerge in the postmarketing period. The adoption of REMS, the Sentinel program, and the new postmarketing authorities allow the FDA to limit or reduce risk exposure among patients by gathering and acting on postlaunch risk information more quickly than they have in the past and in some cases by restricting a drug's distribution (through REMS) to those patients who are the most likely to have a positive benefit–risk profile. This appears to be a significant shift from the 1962 Drug Amendments, but puts the FDA more in line with regulators in the EU countries, who already required risk management plans and risk assessments as part of the approval process and generally had more developed postmarketing safety systems and authorities. The success of these reforms for reducing patient exposure to drug-related risks will depend on the effectiveness of the agency's new tools and programs for identifying and addressing new safety problems quickly and efficiently in the postmarket and the effectiveness of the agency's risks communications to the public. Success will also depend on firm compliance with the agency's policies including the completion of required postmarketing study commitments, including confirmatory studies for accelerated approvals. If proven successful, these reforms could allow regulators to further expand programs used to provide early access to important new medicines.

The most recent renewal of the PDUFA program, the 2012 FDA Safety and Innovation Act (PDUFA V), takes a step in this direction with its provisions to stimulate the development and accelerate approval of new antibiotics and drugs serving unmet medical needs for life-threatening or rare diseases. PDUFA V, which also authorizes user fees for medical devices, generic drugs, and biosimilars, increases further prescription drug user fees and allows some of those funds to build the scientific capacity of the agency, to develop standardized, fully electronic application submissions, to standardize REMS, and to use Sentinel to investigate drug safety issues. PDUFA V continues the trend of fostering greater FDA-sponsor communications during the development and review process. It calls for new staff and two new required meetings during the mid and late stages of review to make sponsors aware of required REMS and advisory committee issues earlier in the review process to prevent unnecessary delays later. In a shift from past programs, PDUFA V essentially extends the review targets for NMEs, NDAs, and original biologics license applications by 2 months by adding a 60-day filing period before the review clock begins to give the agency more time to process increasingly complex drug applications and coordinate its interactions with sponsors and advisory committees. The reasoning is that better coordination up front and communication with firms throughout the process will prevent unexpected delays, which might lead to another review cycle, and thus allow for a shorter overall drug review.

See also: Biosimilars. Patents and Regulatory Exclusivity in the USA. Pharmaceutical Marketing and Promotion. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Pricing and

Reimbursement of Biopharmaceuticals and Medical Devices in the USA. Research and Development Costs and Productivity in Biopharmaceuticals

References

Abraham, J. and Davis, C. (2005). A comparative analysis of drug safety withdrawals in the UK and US (1971–1992): Implications for current regulatory thinking and policy. *Social Science and Medicine* **61**, 881–892.

Begosh, A., Goldsmith J., Hass E., et al. (2006). Black box warnings and drug safety: Examining the determinants and timing of FDA warning labels, NBER Working Paper 12803.

Berndt, E. R., Gottschalk, A., Philipson, T. and Strobeck, M. (2005). Industry funding of the FDA: Effects of PDUFA on approval times and withdrawal rates. *Nature Reviews: Drug Discovery* **4**, 545–554.

Carpenter, D., Zucker, E. J. and Avorn, J. (2008a). Drug review deadlines and subsequent safety problems. *New England Journal of Medicine* **358**(13), 1354–1361.

Carpenter, D., Zucker, E. J. and Avorn, J. (2008b). Errata and corrected estimates. *New England Journal of Medicine* **351**(1), 95–98.

DiMasi, J. A., Hansen, R. and Grabowski, H. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics* **22**(2), 151–185.

DiMasi, J. A., Hansen, R., Grabowski, H. and Lasagna, L. (1991). The cost of innovation in the pharmaceutical industry. *Journal of Health Economics* **10**, 107–142.

Federal Register (2005). *Report on the performance of drug and biologics firms in conducting postmarketing commitment studies*; Availability. 70 Federal Register 8379 (18 February 2005). Available at: https://federalregister.gov/a/05-3221 (accessed 05.02.13).

Friedman, M. A., Woodcock, J., Lumpkin, M., et al. (1999). The safety of newly approved medicines: Do recent market removals mean there is a problem? *Journal of the American Medical Association* **281**(18), 1728–1734.

General Accounting Office (2002). Effect of User Fees on Drug Approval Times, Withdrawals, and Other Agency Activities. Washington, DC: United States General Accounting Office GAO-02-958.

Grabowski, H., Vernon, J. and Thomas, L. G. (1978). Estimating the effects of regulation on innovation: An international comparative analysis of the pharmaceutical industry. *Journal of Law and Economics* **21**, 133–163.

Grabowski, H. and Wang, Y. R. (2008). Do faster Food and Drug Administration drug reviews adversely affect patient safety? An analysis of the 1992 Prescription Drug User Fee Act. *Journal of Law and Economics* **51**(2), 377–406.

Hansen, R. (1979). The pharmaceutical development process: Estimates of development costs and times and the effects of proposed regulatory changes. In Chien, R. (ed.) *Issues in pharmaceutical economics*, pp. 151–187. Lexington: Heath, Lexington Books.

Institute of Medicine Committee on the Assessment of the U.S. Drug Safety System (2007). *The future of drug safety: promoting and protecting the health of the public*. Washington, DC: National Academy Press.

Kaitin, K. I. and Cairns, C. (2003). The new drug approvals of 1999, 2000, and 2001: Drug development trends after the passage of the Prescription Drug User Fee Act of 1992. *Drug Information Journal* **37**, 357–371.

Lester, M. and Godlew, B. (2011). ClinicalTrials.gov Registration and results reporting: Updates and recent activity. *Journal of Clinical Research and Best Practices* **7**(2), 1–5.

Lichtenberg, F. R. (2005). Pharmaceutical knowledge-capital accumulation and longevity. In Corrado, C., Haltiwanger, J. and Sichel, D. (eds.) *Measuring capital in the new economy, studies in income and wealth*, vol. 65, pp. 237–274. Chicago: University of Chicago Press.

Olson, M. K. (2002). Pharmaceutical policy and the safety of new drugs. *Journal of Law and Economics* **45**(2), 615–642, Part 2.

Olson, M. K. (2004b). Are novel drugs more risky for patients than less novel drugs? *Journal Health Economics* **23**(6), 1135–1158.

Olson, M. K. (2008). The risk we bear: The effects of review speed and industry user fees on drug safety. *Journal Health Economics* **27**(2), 175–200.

Peltzman, S. (1973). An evaluation of consumer protection legislation: The 1962 drug amendments. *Journal of Political Economy* **81**, 1049–1091.

Philipson, T., Berndt, E. R., Gottschalk, A. and Strobeck, M. W. (2008). Cost-benefit analysis of the FDA: The case of the prescription drug user fee acts. *Journal of Public Economics* **92**(5–6), 1306–1325.

Sasich, L., Lurie, P. and Wolfe, S. M. (2005). *The drug industry's performance in finishing post-marketing research (phase IV) studies*, A Public Citizen's Health Research Group Report, http://www.citizen.org/hrg1520 (accessed 05.02.13).

Temin, P. (1980). *Taking your medicine: Drug regulation in the United States*. Cambridge: Harvard University Press.

Vernon, J. A., Golec, J. H., Lutter, R. and Nardinelli, C. (2009). An exploratory study of FDA new drug review times, prescription drug user fee acts, and R&D spending. *The Quarterly Review of Economics and Finance* **49**(4), 1260–1274.

Wardell, W., May, M. and Trimble, G. (1982). New drug development by United States pharmaceutical firms. *Clinical Pharmacology and Therapeutics* **32**, 407–417.

Wiggins, S. N. (1984). The effect of US pharmaceutical regulation on new introductions, In Lindgren, B. (ed.) *Pharmaceutical economics*, pp. 191–205. Liber Forlag: Stockholm, Swedish Institute for Health Economics.

Further Reading

Comanor, W. (1986). The political economy of the pharmaceutical industry. *Journal of Economic Literature* **24**(3), 1178–1217.

Faden, L. B. and Kaitin, K. I. (2008). Assessing the performance of the EMEA's centralized procedure: A comparative analysis with the US FDA. *Drug Information Journal* **42**, 45–56.

Faden, L. B. and Milne, C. L. (2008). Pharmacovigilance activities in the United States, European Union and Japan: Harmonic convergence or convergent evolution? *Food and Drug Law Journal* **63**(3), 683–700.

FDANews.com (2008). European medicines agency regulations do not mirror FDA's. *Food and Drug Letter*. Issue No. 805. http://www.fdanews.com/ext/files/FDL.pdf (accessed 05.02.13).

Food and Drug Administration (2009). Report to Congress: Changing the Future of Drug Safety: FDA Initiatives to Strengthen and Transform the Drug Safety System. July 2009. Available at: http://www.fda.gov/downloads/Safety/SafetyofSpecificProducts/UCM184046.pdf (accessed 05.02.13).

Food and Drug Administration (2010). Prescription Drug User Fee Act Public Meeting. Available at: http://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM211617.pdf (accessed 05.02.13).

Food and Drug Administration (2011). Report to Congress: The Sentinel Initiative-A National Strategy for Monitoring Medical Product Safety. Available at: http://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM274548.pdf (accessed 05.02.13).

Grabowski, H. (1976). *Drug regulation and innovation*. Washington, DC: American Enterprise Institute.

Kulynych, J. (1999). Will FDA relinquish the "gold standard" for new drug approval? redefining "substantial evidence" in the FDA Modernization Act of 1997. *Food and Drug Law Journal* **54**, 127–149.

Olson, M. K. (2004a). Managing delegation with user fees: Reducing delay in new drug review. *Journal of Health Politics, Policy, and Law* **29**(3), 397–430.

Psaty, B. and Korn, D. (2007). Congress responds to the IOM drug safety report-in full. *Journal of the American Medical Association* **298**(18), 2185–2187.

Shulman, S. R. and Brown, J. S. (1995). The Food and Drug Administration's early access and fast-track approval initiatives: How have they worked? *Food and Drug Law Journal* **50**, 503–531.

Wiggins, S. N. (1981). Product quality regulation and new drug introductions: Some new evidence from the 1970s. *Review of Economics and Statistics* **63**(4), 615–619.

Woodcock, J. (2011). PDUFA V: Medical innovation, jobs and patients. Statement of Janet Woodcock before the Subcommittee on Health, Committee on Energy and Commerce, US House of Representatives. Available at: http://www.fda.gov/NewsEvents/Testimony/ucm261396.htm (accessed 05.02.13).

Research and Development Costs and Productivity in Biopharmaceuticals

FM Scherer, Harvard University, Cambridge, MA, USA

Introduction

Substantial gains in human health and longevity have been achieved, especially since the 1930s, through the development and introduction of new pharmaceuticals into clinical practice, ranging inter alia from early antibiotics through anti-cholesterol agents to anticancer medicines. Most of the detailed development of new pharmaceutical entities has been conducted, at least in capitalist nations, by private enterprises, typically subject to detailed regulation by government agencies that monitor clinical testing activities and determine whether a proposed new drug is safe and efficacious enough to permit marketing. The pharmaceutical industry is one of the most research intensive of all private industries. During the early years of the twenty-first century, however, there was evidence of sharply rising research and development (R&D) costs underlying the average new pharmaceutical entity introduced into commercial use and hence reduced research productivity. This article explores the evidence and the issues, with a focus mainly on the US, which has played a leading role in drug development and on which the most complete data are available.

Quantitative Overview

Figure 1 presents an overview of inputs and outputs for the drug R&D process. The solid line traces input trends – notably, reported R&D expenditures (right-hand scale, in billions of dollars) by members of the principal US trade association, the Pharmaceutical Research and Manufacturers of America (PhRMA). The data have important limitations. They are adjusted to year 2000 average purchasing power levels using the US gross domestic product (GDP) price deflator, although R&D cost inflation (measured from US National Institutes of Health studies) has probably proceeded slightly more rapidly than general economy-wide price inflation. Most of the leading pharmaceutical producers are multinational firms, but Figure 1 includes only the R&D expenditures of PhRMA members within the US. Counting overseas outlays of the members, many with home bases elsewhere, would add roughly 25% to the cost. Not all private-sector company pharmaceutical R&D outlays are made by PhRMA members. A particularly important exclusion is for biotechnology specialists, many of which do not publicly report their R&D outlays. Several biotech companies were members of PhRMA in 2008, so their data are included in the PhRMA tallies. But most of them joined only after incurring the R&D underlying successful drug developments, and it would appear that the reported PhRMA R&D totals were not recalculated backward to hold membership constant, in which case the addition of new members overstates actual growth rates. Recognizing these limitations, one can estimate from Figure 1 that inflation-adjusted R&D outlays grew between 1970 and 2007 at an average annual rate slightly below 7.4%.

The dash-dash line in Figure 1 estimates the number of new molecular entities (NMEs) (left-hand scale) approved each year for prescription use in the US. The count of NMEs includes new therapeutic organic chemical molecules – the so-called 'small-molecule drugs,' excluding new uses of already-approved molecules and different formulations of pre-existing

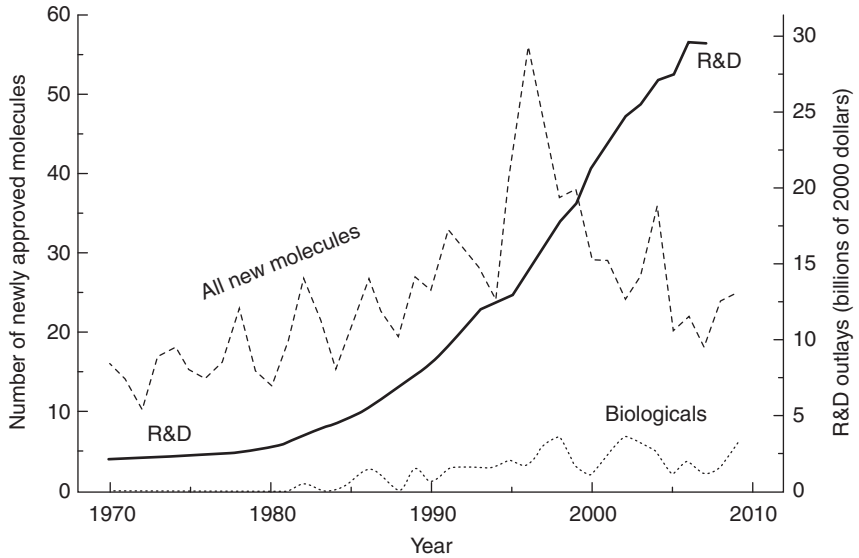


Figure 1 Trends in newly approved molecules and R&D spending US, 1970–2009. Data are drawn from Pharmaceutical Research and Manufacturers of America, *Profile 2008*; Bethan Hughes, 2009 FDA Drug Approvals, *Nature Reviews Drug Discovery* **9**, 89–92 (February 2010); and US Food and Drug Administration, Center for Drug Evaluation and research, statistical tabulations, various years.

molecules, plus the typically much larger molecules derived by gene splicing and related biological processes (but excluding vaccines, blood products, and the like). The large molecule drugs, conveniently called 'biologicals,' are also broken out for separate reporting with the dotted line in Figure 1, beginning in 1982 with the first such new entry, a synthetic human growth hormone. Because source counts vary, a slight estimation error cannot be avoided. It is clear with any set of definitions that the number of new drug approvals varies widely from year to year. The spike around 1996 is artificial, resulting from a sharp fee-induced reduction in the Food and Drug Administration's backlog of drugs awaiting approval. When that peak is redistributed over subsequent years, one finds a modest upward trend of approximately 2.1% per year.

With inflation-adjusted R&D expenditures rising at roughly 7.4% per year and the approval of new pharmaceutical entities increasing at only 2.1% per year, it appears likely that the average R&D cost of NMEs has been rising over time.

The R&D Phases

The discovery and testing of potential new drugs follow a fairly regular sequence of stages characterized in Figure 2. The horizontal time axis is calibrated at zero for the year when testing in humans begins. The vertical axis smooths impressionistically annual spending levels in year 2000 dollars, approximating averages reported by DiMasi *et al.* (2003) for drugs emerging mainly during the 1990s. The costs assumed are those of a project that goes all the way from preclinical work to regulatory approval. No adjustment is made for uncompleted phases, for example, abandonments or failures. There is a long discovery period in which basic and applied research seeks to find and/or synthesize new molecules and identify through theory and *in vitro* testing which drugs might actually work in human beings. In the early years of active pharmaceutical research, the discovery process entailed primarily a random 'try every bottle on the shelf' search, but as scientific knowledge has advanced, theory has come to play an increasingly important role. Once a promising molecule has been identified, it is tested on animals for possible toxicity.

If that hurdle is cleared, a stylized set of human testing phases begins, with appreciable attrition rates at each phase. In Phase I, the drug is administered to a typically small sample of humans to determine the safety of various dosages and in some cases to secure preliminary insight into whether the molecule can alleviate the target disease. If those tests yield promise, targeted Phase II tests for efficacy are conducted in larger cohorts. Success in Phase II is typically followed by considerably more extensive Phase III tests carefully designed with double blinds to infer at reliable levels of statistical confidence whether the drug is safe and effective relative to placebos or, less often, relative to the best-accepted approved drug in the relevant therapeutic category. Phase III tests, typically divided into at least two distinct protocols, may encompass from a few hundred human subjects (only for diseases with no known cures) to more than 10 000 individuals. If the results from Phase III are promising, the drug developer (usually a private pharmaceutical company) applies for marketing approval – in the US, for a new drug approval issued by the Food and Drug Administration; and in Europe since 1995, to the European Medicines Agency. On average, only one-fifth to one-fourth of the small-molecule drugs entering Phase I testing emerge approximately 8 years later with marketing approval. For biological therapies developed during the 1990s, the survival probabilities appear to be higher – for example, roughly 0.3 from a survey by DiMasi and Grabowski (2007) and even higher for the earliest approved biologicals, which mainly emulated naturally occurring substances.

The relevant regulatory agency may after approval insist on additional tests to clarify remaining uncertainties, in which case, further trials continue into a Phase IV. Or the company developing the drug may seek to illuminate more exactly the differences between its drug and existing competitors, embarking on its own initiative into further Phase IV testing.

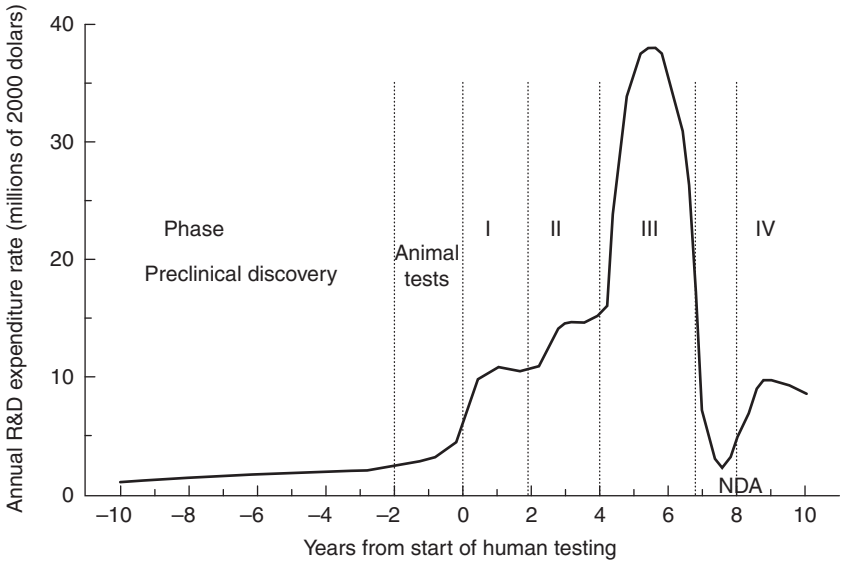


Figure 2 Schematic history of a successful new drug development. Adapted from data in DiMasi *et al.* (2003).

For wholly new vaccines (as compared to minor variants adapted annually to new strains of influenza) even larger human test samples are often needed. The basic problem is, once a subject acquires the target disease, it may be too late for vaccine administration. For preventative vaccines, tests are conducted on populations that might be afflicted in the future, and to keep trial periods within reasonable time bounds, given small probabilities that any given sample member will actually acquire the disease, samples numbering in the tens of thousands may be required to achieve acceptable levels of statistical discrimination along with detecting adverse reactions. To ensure accuracy, subjects might alternatively be injected with the target organism after vaccine administration, as was done, for example, during the eighteenth century in the discovery of the first cowpox-based smallpox vaccine, but this approach violates medical ethics and is now avoided.

Estimating R&D Cost per Successful New Drug

Given this broad picture and its many variations, interest has been focused on the productivity of research and development in yielding new pharmaceutical therapies, i.e., the cost per successful new molecule. There have been numerous quantitative investigations. The leading efforts, and those most highly cited in both the scientific and popular literature, have come from collaborating economists at Tufts University, the University of Rochester, and Duke University (DiMasi *et al.*, 2003). Their methodology, which aims to minimize the proliferation of names will be called the Tufts University studies, enlisted deep cooperation from a handful of major pharmaceutical companies (in the most comprehensive recent effort, 10) operating in the US. The investigators began by identifying a set of clinical testing programs undertaken by the cooperating enterprises on 68 so called 'self-originated' molecules first tested in humans between 1983 and 1994. Their sample excluded 'licensed in' drugs whose early development was performed by companies other than the survey respondents. Once the molecules entered clinical testing, detailed data on individual test program costs and failure rates were obtained so that they could be aggregated into estimates of the average cost per successful molecule, i.e., the actual out-of-pocket cost of the ultimate successes, into which were loaded the probability-adjusted estimates of preclinical and clinical phase failure costs. If, for example, only one molecule out of five entering Phase I testing ultimately secured marketing approval, the cost of an average Phase I test, successful or unsuccessful, was multiplied by $1/0.2=5$ to obtain the average Phase I success cost. Similar probabilistic adjustments were made for later stages. For the most recent of the comprehensive Tufts University studies, the estimated average out-of-pocket cost per successfully approved molecule, including the prorated costs of failed tests, all measured in year 2000 purchasing power levels, were as follows:

Preclinical	$121 million
Clinical testing	$282 million
Total cost per approved drug	$403 million

The mean clinical testing estimates, which are undoubtedly more reliable than preclinical estimates, can be compared with the analogous costs from three earlier studies summarized by Scherer (2010), each adjusted to year 2000 purchasing power levels:

Source	*Test period*	*Average out-of-pocket cost per approved new drug*
Mansfield	Late 1950s	$5.4 million
Clymer	Late 1960s	40.2 million
Tufts I	1970–early 1980s	65.7 million
Tufts II	1983–late 1990s	$282 million

It seems clear that the clinical R&D costs of new drugs have exploded over time. The particularly large multiplier between the estimates of Edwin Mansfield and Harold Clymer is explained by the fact that after 1962, constrained by new and tougher legislation, the US Food and Drug Administration enforced considerably more stringent rules for the evidence it would accept before approving new drug applications. The subsequent sizeable increase between Tufts I and II is now discussed later.

The estimates of success probability-adjusted preclinical R&D costs by the Tufts group are more problematic. For the Tufts II sample, it is seen in the table above, the mean value was $121 million, or 30% of total estimated mean cost per successful molecule. For the Tufts I sample (from the 1970s), it was $90 million (in year 2000 dollars), or 57.8%.

The striking reduction over roughly 15 years in preclinical cost shares, not explained by the Tufts researchers, is probably attributable to radical changes in the way new drugs have been discovered. The science of drug action in the human body advanced by leaps and bounds in the time interval separating the two studies, leading among other things to so-called 'rational drug design,' i.e., the structured synthesis of molecules targeted to interact in particular ways with known receptors in the human body. (A detailed chart of biological pathways is revised and published periodically by the Boehringer-Ingelheim Co. (Michal, 1993).) Much of the research underlying such insights was conducted not in drug company laboratories, but in universities and hospitals supported by grants, most notably, from the US National Institutes of Health. Between 1983 and 2000, the research budget of 'NIH' rose from roughly $2.7 billion (at year 2000 GDP price levels) to $14.4 billion, or two-thirds of US R&D outlays by PhRMA member firms in 2000. Additional research support came from the US National Science Foundation and private philanthropic institutions. An unknown but undoubtedly substantial fraction of such outlays generated basic knowledge helpful in the design of new pharmaceutical entities and in many cases identified specific molecules eventually brought into clinical testing by private sector enterprises (Scherer, 2010; Stevens *et al.*, 2011).

Also, the first drug synthesized using radically new gene splicing methods was introduced commercially in 1982, spurring the explosive growth of a new biotechnology industry, mostly in new companies initially financed by venture capital. Although 90% of the entities comprising the Tufts II sample were small molecules (as compared to biologics), it

cannot be ruled out that the sample companies saved some preclinical R&D expenditures by building on research done inter alia in biotech enterprises. However, DiMasi and Grabowski (2007) report quite similar constant-dollar R&D cost estimates for their Tufts II sample and a slightly later sample covering only biological entities.

Critiques

The Tufts estimates and their predecessors have been widely cited by pharmaceutical industry advocates to argue that drug testing is both risky and costly, and, with additional evidence, that government agencies ought not to intervene in pharmaceutical companies' controversial price-setting process (which in fact many national governments do through various price control mechanisms). Given this, the estimates have been criticized as biased and excessive. Diverse and conflicting critiques are found in Love (2003), Angell (2004), and Light and Warburton (2011). The criticisms have several foci.

Capitalization

More widely cited than the out-of-pocket averages presented above are estimates from the Tufts research of average drug discovery costs, capitalized to 'present value' at the time of product approval to reflect the cost of capital tied up during the R&D period. In the 1983–late 1990s estimates presented above, for example, out-of-pocket costs were capitalized to the time of marketing approval at an implied 11% cost of capital. To illustrate, suppose that 10 years before a new drug's approval date, for example, at year −2 in **Figure 2**, out-of-pocket costs amounting to $5 million (with adjustments for failed trials) are observed. The capitalized figure becomes $5 million $\times (1.11^{10}) = \$5$ million $\times 2.84 = \$14.20$ million, which is the value incorporated into the capitalized R&D cost sums. Here, 1.11^{10} is the amount to which $1 grows over 10 years at compound annual interest. Such adjustments are made for each year to take into account the 'opportunity cost' of companies' investable funds on the assumption that if the money were not invested in R&D, investors could allocate it to other comparably risky assets that over time would yield 11% inflation-adjusted annual returns (derived from standard finance sources using the so-called 'Capital Asset Pricing Model'). For years nearer the time of marketing approval, the adjustment is of course smaller; for example, 5 years out, $1.11^5 = 1.685$ rather than 2.59. When these capitalization adjustments are made, among other things giving relatively greater weight to preclinical as opposed to clinical testing costs, the $403 million Tufts II average successful drug development cost reported above nearly doubles to $802 million. For the earlier Tufts I study, average out-of-pocket costs (preclinical plus clinical) rise in year 2000 dollars from $156 million uncapitalized to $318 million capitalized.

This capitalization assumption, typically reported in the popular press without explanation, has been criticized by, for example, Light and Warburton (2011) on both conceptual and numerical grounds. To ensure correctness, most estimates of investment outlays for research and development as well as physical facilities, advertising, and much else, are typically publicized in unadjusted form for the year of incurrence rather than with capitalization, and consistency in reporting practice would argue for avoiding capitalization, unless the rationale is clearly explained. Nevertheless, it is clear that R&D outlays do have opportunity costs, and in drug discovery and testing, with their long time lags between outlay and the return of profits, the opportunity costs are more significant than for investments with quicker paybacks. Public controversy over the capitalization issue became sufficiently intense in the early 1990s that a specially created US government agency study team focused on it, among other things obtaining consulting assistance from prominent finance theorists. Its report was in US Congress, Office of Technology Assessment (OTA) (1993). The study group concluded that the three most important components of R&D investment are 'money, time, and risk' and that 'the practice of capitalizing costs to their present value in the year of market approval is a valid approach to measuring R&D costs....' Given the lack of public understanding; however, it would undoubtedly be good practice for journalists to report out-of-pocket costs along with capitalized cost estimates.

The higher the interest rate used in capitalization, the larger is the multiple between out-of-pocket and capitalized costs. Light and Warburton (2011) argue that the 11% interest rate used by the Tufts group was too high, given that US Government Office of Management and Budget guidelines in 2003 called for applying a 3% interest rate in evaluating public capital outlays. This criticism is clearly wrong. Governments like in the US (at least up to the year 2012) financed their deficits with what were widely considered 'risk-free' bonds that indeed often bore quite low interest rates. But the common stock with which corporations are financed is riskier and bears considerably higher implicit interest rates. Addressing this issue, finance experts advising the US OTA found (p. 67) that the cost of capital (i.e., the implicit interest rate) for established pharmaceutical companies in the 1980s and early 1990s was on the order of 8–10% after stripping away inflation premia. They also found that R&D-intensive activities were more risky than ordinary corporate investments, calling for interest rate premia on the order of 4.5 percentage points, or approximately 13–14% overall. Recognizing this, the 11% implicit interest rate used by the Tufts group appeared consistent with broader knowledge and perhaps even conservative for the time period covered.

Tax Benefits

Some critics have argued that tax savings realized by corporations as a result of their R&D outlays (treated as current expenses under prevailing tax accounting) ought to be deducted in estimates of what drug development costs. It is true that tax offsets exist. Considering first only the corporate income tax, when a corporation spends an incremental dollar on R&D, that dollar reduces its current pretax profits by a dollar (assuming profits to be positive), and at the 34% US corporate income tax rate prevailing at the time of the most recent Tufts study, a savings of 34 cents is achieved. The problem with adjusting for this saving is that it applies for any incremental expenditure in a positive-income regime – for the cost of hiring an additional worker, for the cost of fuel, for the cost of

environmental cleanup activities, and so on. But to apply such adjustments for each expenditure requires distinctions between optional and mandatory outlays and runs into the difficulty that, if every expenditure were treated as less costly than its out-of-pocket cost, expenditures could rise to exhaust the profits against which savings are claimed. Also, multinational pharmaceutical companies have been adept at shifting their reported profits to nations with low marginal income tax rates, so any attempt to offset R&D outlays by tax savings would have to cope with a multiplicity of savings rates.

A slightly better case can be made for adjusting R&D outlays for tax benefits specific to R&D. These were of two main relevant forms. Under US law since the 1980s, credits against income tax liability have been offered for increases in R&D expenditures relative to the amount expended in specified base years. The provisions of the law have varied from time to time, so adjustments would be complex. Because the credits apply only to incremental outlays above a base year value, it would be difficult to determine which outlays in a large R&D budget are incremental and which are within the no-credit baseline. Special 50% federal income tax credits have also been offered under US law since 1983 for costs incurred testing so-called 'orphan' drugs, i.e., those expected to serve small patient populations. Because the credits are targeted at specific molecules, adjustments to orphan drug R&D costs would be more feasible than adjustment for generalized tax savings. Other complexities of estimating orphan drug R&D costs are considered later.

More generally, the genuine issues posed by capitalization and tax benefits are best judged in policy evaluations of pharmaceutical companies' aggregate net profitability, not with respect to specific drug discovery and testing cost estimates. There too issues arise, although they are beyond the scope of this article. The OTA study concluded (1993) that established pharmaceutical firms' rates of return on net capital averaged 2–3 percentage points higher than their cost of capital, estimated to be roughly 10% after taxes. The OTA group refrained from rendering a clear value judgment as to whether such a premium was problematic, given the risks of new drug development and the desirability of attracting new investment.

Sample Representativeness

Without doubt the most compelling criticism of the Tufts methodology is that their samples may not have been representative of the entire drug development universe. For the Tufts II estimates, the unnamed product sample and cost data came from 10 pharmaceutical firms, 8 of them from the top 20 in terms of sales volume – i.e., the representatives of what many call 'Big Pharma.' It is conceivable that the drugs chosen for development by those companies differed from those developed by smaller firms or even misrepresented the respondents' typical portfolios. In particular, with vast sales pipelines to fill, the companies may have emphasized candidates with a large sales potential – i.e., with luck, the 'blockbusters.' For example, drugs that address widespread health conditions and that are prescribed for chronic as contrasted to acute symptoms tend to have better sales prospects than those targeting relatively rare and/or acute conditions – e.g., those with the mandate, "Take two tablets per day for ten days and if the symptoms persist, see your doctor." Higher sales prospects, both theory and statistical analyses reveal, induce more lavish R&D outlays (Scherer, 2010).

The testing strategies mandated by the Food and Drug Administration or favored by the companies may also have differed. For drugs that will be taken daily for years on end, regulators tend to be more wary of rare and/or cumulative adverse side effects and require larger samples to impart additional statistical confidence on what might otherwise be seen as clinical testing flukes. And for drugs alleviating chronic medical problems of long standing, new drugs will often have to compete with existing therapies that may arguably be less effective, but the differences are foreseen to be sufficiently small that tests are authorized not against placebos, but against established molecules, with unusually large clinical populations to obtain evidence bolstering marketing claims that the new drug is in fact superior to existing alternatives.

A case history at the opposite extreme is seen in the first drugs effective against human immunodeficiency virus/acquired immune deficiency syndrome (AIDS), recognized as a threat by physicians only in the 1980s. The lethality of AIDS was so shocking, and its spread so rapid, that clinicians and regulators accepted major shortcuts to ensure that weapons against the disease were immediately available. The first candidate, azidothymidine (AZT) (also known as zidovudine), was approved by the Food and Drug Administration in March 1987 – only 25 months after the start of human testing, breaking post-1962 speed records. Although comparative placebo tests were conducted, the decisive trial included only 282 patients, and instead of waiting to see whether or how long AZT recipients lived, FDA evaluated the drug's efficacy mainly on the basis of 'surrogate endpoints' – i.e., measures of retroviral levels in trial subjects' blood. Clinical trials were conducted jointly by Duke University, Burroughs-Wellcome, and the National Institutes of Health, with substantial financial support from the NIH. Another AIDS drug, Nevirapine, with the remarkable ability significantly to inhibit transmission of the disease from infected mothers to newborn children, was approved in June 1996 after trials spanning 76 weeks on a total of 549 patients, one branch conducted by the US NIH in parallel with other tests by the drug's inventor, Boehringer-Ingelheim of Germany (Love, 2003).

The initial AIDS drug developments shared two distinguishing bureaucratic characteristics. First, the early AIDS population was sufficiently small that the first therapeutic candidates were ruled at the outset to be 'orphan drugs' – i.e., mainly targeted toward conditions afflicting 200 000 or fewer individuals in the US. Second, they were also accorded 'priority' status by the Food and Drug Administration – i.e., for molecules offering potentially major improvements over already marketed therapies, as distinguished from 'standard' drugs yielding more modest therapeutic gains.

As has been seen, private funds in the US devoted to orphan drug testing have been accorded in the US especially favorable tax status, and clinical testing support by Federal government entities is also common. Recognizing the possibly small market potential of orphan drugs, the Food and Drug Administration has tended to accept smaller clinical trial samples than for drugs targeting wider markets. Also, because

of the tax implications, data are publicly available on the total amount spent for orphan drug testing. For 16 new orphan chemical entities approved in the US between 1998 and 2000, the average clinical trial cost per approved orphan, prorating the costs of failed tests, was $34 million (Love, 2003). This is far below the $282 million out-of-pocket for the most recent Tufts sample, the bulk of whose testing outlays occurred in years earlier and hence were less inflated than those gleaned by James Love. Although DiMasi *et al.* (2003) do not elaborate the point, the OTA reported (1993, p. 232) that roughly two-thirds of orphan drug designations went to companies that were not PhRMA members.

Orphan drugs are also more likely to obtain priority rankings from the Food and Drug Administration than standard drugs. Thus, for new chemical entities approved by the FDA during the first 5 years of the twenty-first century, 89% of the orphans had priority ratings, as compared to 38% for the standard drugs (Scherer, 2010). Since the early 1990s, the Food and Drug Administration has tended to process nonorphan priority drug approval requests more rapidly than standard requests. It is also possible that FDA demands fewer and less costly clinical trials for priority drugs, but on this the evidence is sparse. DiMasi *et al.* (2003) report that in their Tufts II sample, the out-of-pocket clinical testing costs of priority drugs exceeded the average cost of standard drugs by a statistically insignificant amount. The difference was even smaller for capitalized costs, implying that test-to-approval lags were shorter for the priority drugs. DiMasi *et al.* (1991) suggest that priority drugs may have been more costly to test because they break newer scientific ground, requiring more learning-by-doing, and also (p. 172) because "firms have the incentive to do more wide-ranging and costly testing on drugs that have the potential to be both clinically and commercially significant." Whether this inference carries over to the non-orphan priority drugs tests of smaller companies is unknown.

An Independent Test of the Evidence

There are other fragments of evidence suggesting average out-of-pocket costs lower for drugs outside the Tufts sample than for in-group molecules. But now an alternative approach is explored. Several authors, such as Adams and Branter (2010), have pursued more aggregative approaches to the problem of estimating drug development costs. Here the author report the result of his own broad-brush approach. The methodology is simple: dividing annual counts of new therapeutic entity approvals in the US into the reported intra-US research and development spending of PhRMA members. It is bound to be incomplete and inexact for at least four reasons. First, PhRMA's membership includes companies with a home base outside the US, and by excluding overseas R&D outlays, the full costs of their drugs approved in the US are certain to be underestimated. Second, many of the drugs approved in the US come from non-PhRMA members, and although such firms' innovations are included in the denominator of cost/drug calculations, their R&D outlays are excluded from the numerator, again resulting in an underestimate. Over the years 2001–05, the non-PhRMA share of approved new medical entities was 51%, implying a sizeable downward bias. Third, the R&D expenditures of PhRMA members are focused not only on developing and testing NMEs but also testing to see whether existing molecular entities are effective against additional disease conditions, developing vaccines and other biological products, and reformulating inert binders that control the timing of a drug's release into the blood stream. And much Phase IV research undertaken by major pharmaceutical firms is aimed not at complying with regulatory agency mandates, but to strengthen evidence used in field marketing of already approved molecules. By excluding such projects from the denominator count, the cost per drug, new and old, is overestimated. And finally, as we have seen in **Figure 2**, the R&D expenditures underlying new entities precede by as much as a decade the date of approval. Lags must be accounted for, but are inherently variable.

Recognizing that perfection is unattainable, the following methodology was pursued. The **Figure 1** time series of NME approvals for the years 1974–2007, including both small-molecule drugs and some biologicals (but not vaccines and the like) was used as the denominator of the cost calculation. Total reported R&D expenditures of PhRMA members in the US were used, adjusted with the GDP deflator to constant year 2000 price levels, to measure costs. To reflect the fact that approvals lag the incurrence of testing costs, the R&D series was prelagged by 4 years relative to approvals, for example, approvals in the year 2000 were related to 1996 R&D expenditures. This convention reflects in a crude way the central tendency of the outlay flow shown in **Figure 2**, with outlays peaking 3 years before approval but with early outlays weighted more heavily because of attrition.

The Tufts II analysis focused on drugs whose clinical test expenditures were mostly incurred between 1983 and 1999. Within that restricted sample of years, the computed average out-of-pocket cost R&D per lagged NME approved, using the methodology described above, was $306 million in year 2000 dollars (also used as the measuring basis for Tufts group's summary estimates). When the exercise was repeated without the inclusion of biological entities, the average was $390 million. The Tufts II estimate, including seven biologicals (10% of the sample) was $282 million. The difference is small, suggesting that for an intrinsically difficult measurement, the Tufts estimates are both credible and perhaps even conservative.

For the full 1974–2007 molecular approval series, the average growth rate of constant-dollar R&D costs per molecule was found by regression analysis, which smooths year-to-year variations, to be 6.5% per year with biologicals included and 7.2% per year with them excluded. DiMasi *et al.* (2003) estimated the growth rate between their Tufts I and II studies, spanning slightly shorter intervals, to be 7.4%. Again, the conclusion seems inescapable that there has been substantial growth in R&D costs per new approved molecule, or in other words, a decline in research productivity.

Reasons for Change

Several hypotheses vie to explain the apparently continuous increase in R&D costs per molecule approved. Despite advances in the technology of preclinical small-molecule screening, one might suppose that diminishing returns would

set in after seven or more decades of active discovery, among other things forcing companies to focus on more difficult therapeutic targets. During the 1990s it was thought that the perfection of large-molecule gene-splicing techniques would reverse any such tendency and usher in a new golden age of pharmaceutical discovery. However, the observable changes thus far have been less than revolutionary.

There is definite evidence that clinical trial sizes have risen over time, partly as a result of tougher standards established by the US Food and Drug Administration. Also, as individual therapeutic classes became more crowded, companies may have elected to increase sample sizes to improve the statistical significance of results touted in competitive marketing. For three therapeutic categories studied by the OTA (1993), average enrollment in Phase I through III clinical trials rose from 2237 for drugs approved in 1978–83 to 3174 for 1986–90 entities, implying a median year growth rate of 4.7%. The average number of subjects drawn into Phase IV grew considerably more rapidly, from 413 to 2000 (sic), or 21% per year. Using publicly available data, DiMasi *et al.* (2003) estimate that average trial sizes in the 1980s and 1990s rose at a rate of 7.47% per year. In addition, the complexity of trials rose. DiMasi *et al.* (2003) report from an outside data source that the number of procedures administered per trial subject increased between 1990 and 1997 by 120% for Phase I trials, by 90% for Phase II trials, and by 27% for Phase III trials. Weighting the phase growth percentages by the fraction of out-of-pocket costs incurred per phase, this implies an average growth of 50% in 7 years, or 5.8% per year.

Clinical trials are mostly conducted in hospitals and similar medical centers. Over the period 1970–90, a day of hospitalization costs in the US rose at an average rate of 11% per year – nearly twice the rate at which the GDP price index was increasing. It seems reasonable to assume that in-hospital test costs rose commensurately. There is also reason to believe that major hospitals view their clinical testing activities as a 'profit center' and dump some of their soaring overhead costs onto the well-heeled pharmaceutical firms sponsoring clinical trials.

A more speculative hypothesis is that 'Big Pharma' companies have allowed organizational slack to accumulate in their R&D activities, especially after numerous large-company mergers failed to achieve substantial increases in the output of new therapeutic entities (Munos, 2009). A correction against this trend may have begun in the second decade of the twenty-first century as pharmaceutical giants such as Pfizer and Merck, acknowledging disappointment over the lagging productivity of their innovation efforts, cut back their R&D staffs in the wake of major new mergers.

Conclusion

In summary, the research and clinical testing costs underlying pharmaceutical innovations have risen considerably over recent decades to levels measured in hundreds of millions of dollars per approved new molecule. The most widely publicized estimates of R&D costs, sometimes poorly understood, are consistent with alternative estimates. There would probably be less controversy over those estimates if more detailed data on sample composition were disclosed, but confidentiality constraints imposed in exchange for access to company microdata may preclude this. It is clear that clinical success may be achieved at substantially lower cost with alternative models of pharmaceutical development and testing, but embracing those alternatives requires streamlined regulatory and organizational approaches and sacrifices in the richness of the evidence on the basis of which physicians must make subsequent prescription choices.

See also: Biosimilars. Health and Its Value: Overview. Pharmaceutical Marketing and Promotion. Regulation of Safety, Efficacy, and Quality. Time Preference and Discounting

References

Adams, C. P. and Branter, V. V. (2010). Spending on new drug development. *Health Economics* **19**, 130–141.

Angell, M. (2004). *The truth about the drug companies*. New York: Random House.

DiMasi, J. A. and Grabowski, H. G. (2007). The cost of biopharmaceutical R&D: Is biotech different? *Managerial and Decision Economics* **28**, 469–479.

DiMasi, J. A., Hansen, R. W. and Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics* **22**, 151–185.

DiMasi, J. A., Hansen, R. W., Grabowski, H. G. and Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. *Journal of Health Economics* **10**, 107–142.

Light, D. W. and Warburton, R. (2011). Demythologizing the high costs of pharmaceutical research. *BioSocieties* **6**, 34–50.

Love, J. P. (2003). *Evidence regarding research and development investments in innovative and non-innovative medicines*. Washington: Consumer Project on Technology.

Michal, G. (ed.) (1993). *Biochemical pathways*, 3rd ed., 2 parts. Mannheim: Boehringer-Mannheim.

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery* **8**, 959–968.

Scherer, F. M. (2010). Pharmaceutical innovation. In Hall, B. H. and Rosenberg, N. (eds.) *Handbook of the economics of innovation*, pp. 539–574. Amsterdam: North-Holland.

Stevens, A. J., Jensen, J. J., Wyller, K., Kilgore, P. C., et al. (2011). The role of public-sector research in the discovery of drugs and vaccines. *New England Journal of Medicine* **364**, 535–641.

U.S. Congress, Office of Technology Assessment (1993). *Pharmaceutical R&D: Costs, risks, and rewards*. Washington: Government Printing Office.

Further Reading

Lichtenberg, F. (2007). The impact of new drug launches on U.S. longevity and medical expenditures. *American Economic Review Proceedings* **97**, 438–441.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203–214.

Resource Allocation Funding Formulae, Efficiency of

W Whittaker, University of Manchester, Manchester, UK

Glossary

Actual allocations The allocations Primary Care Trusts (PCTs) in England receive for the forthcoming financial year.

Allocative efficiency Producing the maximum output subject to a constraint.

Department of Health The ministry responsible in England and Wales for determining PCT allocations for each financial year.

Disability free life expectancy (DFLE) It is the expected years of life with no limiting long-term illness or disability.

Market forces factor (MFF) It is an index used to weight PCTs in accordance to the differing costs of delivering health care across England.

National Health Service It is the provider of publicly funded healthcare in the four countries of the United Kingdom (England, Wales, Scotland and Northern Ireland).

Need Capacity to benefit from additional expenditure, often proxied by utilisation. Additional need represents need over and above the age structure of the population. Legitimate and illegitimate need are often operationalized as utilization that do/do not reflect need. Unmet need represents under-utilization of services by groups of the population. General unmet need affects all groups in the population proportionately. Specific unmet need affects particular groups of the population.

Pace of change The rate at which PCTs are moved toward their target allocations from previous allocations.

Payer The institution or body responsible for distributing the healthcare budget.

Primary Care Trust The geographic organizations responsible for the spending of an area of England.

Production possibility frontier It is the maximum amount of health generated for differing levels of expenditure.

Pure efficiency Allocations made that maximise risk-adjusted health gains.

Resource Allocation Working Party It was set up to attain a resource allocation formula that objectively, equitably and efficiently responded to relative need.

Social welfare function A function that maps from the levels of utility attained by members of society to the overall level of welfare for society.

Target allocations Allocations expected to meet the needs of the PCT in accordance to the PCT population size, age structure, additional needs (over and above age), MFF, and DFLE.

Technical efficiency A given output is produced using no more inputs than are technically necessary – there will normally be a wide variety of different combinations arising out of their substitutability.

Weighted capitation Payment per individual, weighted by risk, more risky populations hence receive higher payments.

Introduction

Publicly funded health care systems require some form of resource allocation funding principles, usually in the form of formulae, to enable the payer (typically a government body) to distribute health care budgets across population groups. Population groups are typically defined by geography and a population within each geographic boundary is likely to have variations across individuals within it both in terms of health and in the utilization of healthcare services.

To distribute health care budgets efficiently requires knowledge of how population groups differ, which in turn relies on up-to-date data on population variations in the need for health care. Need, taken here as the capacity to benefit, is difficult to measure. Poor data availability usually results in need being proxied by service utilization, with the assumption that populations with higher needs will have higher rates of health care services use. Precision is therefore not really to be expected in matching budgets to local needs, because data on neither can be perfect. Perhaps the best that can be expected is that formulaic solutions push the system in the right directions.

Utilization, however, reflects access to care which may be the product of both demand (influenced by an individual's need for care, how affordable care is to the individual and whether the individual is willing to accept care) and supply factors (whether the individual has providers of care in their area and whether care is available). If access varies by geographic population groups, funding formulae using utilization measures alone to allocate budgets could reinforce inequalities and inefficiencies in both health and access to health care.

An alternative way of measuring need is to use demographic and/or health related population characteristics as a proxy (e.g., a mortality ratio, which assumes populations with higher needs have higher mortality ratios). Although such proxies can serve as an alternative to measuring need as utilization, they are most frequently seen as complementary measures, the assumption being that they detect different aspects of need.

Weighted capitation methods are typically used to apply a needs-based approach to resource allocation in health care. These methods weight populations by indices of need and are used to determine each population group's share of the health care budget.

The main challenge in designing resource allocation formulae is that they ought to allocate health care budgets in accordance with the payer's objectives. The payer's objectives typically relate to efficiency, equity, or more likely, a mixture of

 doi:10.1016/B978-0-12-375678-7.00211-X

both. The prime concern lies with whether the funding formulae used to derive allocations is efficient, i.e., pushes the system toward maximizing the health of society given the resources available, or whether there is potential for inefficiencies in the funding formulae such that the push is in the wrong direction. This article investigates how weighted capitation approaches, and the payer's efficiency-equity objective, impact on the efficiency of the resource allocation funding formulae used to distribute health care budgets.

The article is structured as follows. First, what is meant by efficiency in the resource allocation formulae is explained and examples provided of the conflict between efficiency and equity in resource allocation using two population groups. Second, the extent to which the formulae impact or are impacted by technical efficiency is looked at. Examples of how the resource allocation formulae have developed since the introduction of the National Health Service (NHS) in England, and the impacts of these changes on efficiency are then given. Much of the discussion presented on the efficiency of using a weighted capitation approach in England is applicable to other international settings. While the methods for the financing of health care in the developed world vary widely, there has been increasing use of capitation payments. The final section provides the article summary.

Efficiency

A Production Possibility Frontier Approach to Resource Allocation

Economic efficiency consists of two types of efficiency, allocative and technical efficiency. Allocative efficiency concerns producing the maximum output subject to inputs, i.e., it is not possible to increase output simply by reallocating resources, and it is achieved by equalizing the marginal capacity to benefit from additional funds across all inputs, while technical efficiency concerns utilizing a specified combination of inputs to produce maximum output. From a resource allocation perspective economic efficiency concerns maximizing the value of output (health gains and/or prevention) from given resources.

The problem with measuring efficiency in health care is that it is not obvious how to define the output of interest: is it health care use? Access to health care? Or measures of health gains or outcomes? Further, how can such data be obtained? Hollingsworth (2008) provides a review of the literature surrounding the measurement of production frontier efficiency in health care. The answers to these questions are not obvious, and will likely invoke some normative judgment.

Figure 1 gives a graphical representation of a production possibility frontier (PPF) for a public service, that is, the plot of health outcomes associated with spending on a population group (this example, and what follows is adapted from Smith (2007)). The aim is to determine the optimum payment for the population group given the payer's efficiency and equity objectives. The PPF relates expenditure (X) on the population group to the health outcomes (Y), and is a cumulative outcome for all individuals within the population group. The more productive the population group is in generating health,

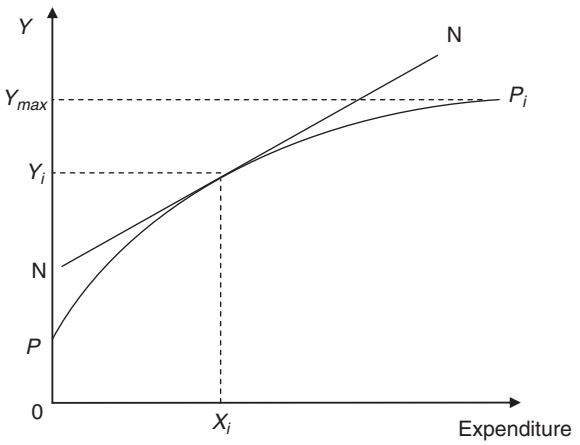


Figure 1 Production possibility frontier (PPF). Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

the higher the PPF lies. In this simple example it is assumed that no historic spend feeds into present health, and that population groups with zero public health care expenditure may have a positive health outcome. This may be because of private utilization of health care services, for example. This figure, and all subsequent figures are illustrative, and not much should be read into the intercept.

The PPF for population group i (PP_i in Figure 1) represents outcomes for payments where technical efficiency holds (efficient production of the health care by the providers serving the population group). Further assumptions are that there are decreasing returns to expenditure (the PPF is concave), that there is only one input (or alternatively that all inputs can be aggregated to one measure, in this example, budget allocations) and one output (some measure of health such as life-years gained, whose value is independent of the identity of those in whom it is embodied), and that other factors such as different providers, complimentary services, the environment, personal characteristics, and societal influences are exogenous to the shape of the PPF. The time period is assumed to be 1 year (that, in reality, is typically the time frame used for allocated budgets).

The need of population i can be measured as the difference between the level of health at a particular level of expenditure, and the maximum health attainable. In the case with unlimited funds, one may expect to observe expenditure up to the point where the benefit (additional health) of additional expenditure is zero (i.e., where further need is zero). This is where the PPF flattens out (Y_{max}). Unfortunately, resources are scarce, and the available budget allocated to health care (which in itself is also an issue for efficiency at the State level) may not be enough to ensure that maximum health is attained. With an allocated health care budget payers have to determine how to distribute this across geographic populations.

Figure 1 highlights, in this simple example, how determining the amount of expenditure determines the outcome achieved. With this in mind, and with knowledge of each geographic population's PPFs, the payer can optimize a social welfare function.

The payer determines the social welfare function and hence marginal social value (slope of NN in Figure 1), essentially a

cut-off cost above which no more treatment is offered. The payer provides expenditure up until the point where additional expenditure results in less health produced than is valued by society. Only at expenditure X_i does the marginal social value of the payer meet the marginal gain in health outcome (benefit). Expenditure less than X_i corresponds to a higher marginal benefit than marginal social value (the slope of the PPF is steeper than the slope of the welfare function, NN). Expenditure greater than X_i corresponds to a lower marginal benefit than marginal social value (the slope of the welfare function, NN, is steeper than the slope of the PPF). A steeper social welfare function would have a higher marginal social value of expenditure, and this would imply optimum expenditure at a higher marginal gain in health, which would lie to the left of X_i. To be allocatively efficient the payer should equate the marginal social value to the marginal benefit, this would be at expenditure X_i and corresponds to the pure efficient solution to resource allocation. The pure efficient solution maximizes aggregate outcomes, and is where expenditures are allocative and technically efficient. Alternative allocations would result in a reduction in total health produced.

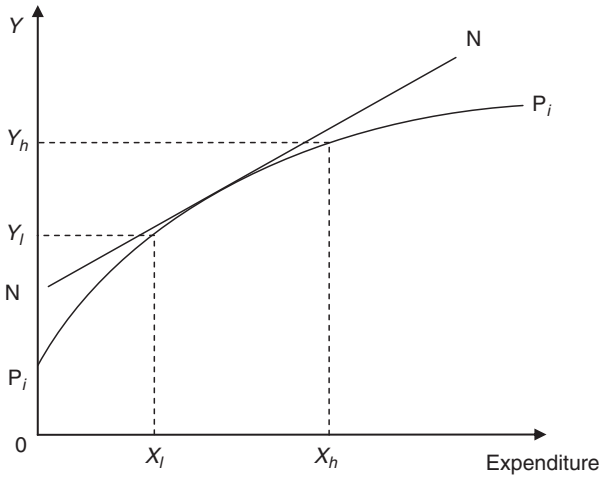


Figure 2 Inefficient allocations of expenditure across two populations with the same PPF. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

The Divergence from Pure Efficiency

In reality there are two main reasons why the pure efficiency solution above is never met. First, allocations may not be set at the point where the marginal capacity to benefit from expenditure equals the marginal social value. This may be because of inaccurate needs measurement, differences in the cost of delivering health care across population groups, or as a result of an alternative equity objective held by the payer. Second, although the payer may distribute budgets efficiently, budgets might not be spent efficiently by providers serving the geographic populations (technical inefficiency). This section will look at each case.

Inaccurate needs measurement

Where expenditures do not reflect need, the implication is that populations of equal (risk adjusted) marginal need receive different budgets. Figure 2 gives a graphical example. In this case, two populations have equal need and are equally productive (the PPFs for the two populations are the same and overlap), but receive different allocations, X_l and X_h. One population receives X_l and resulting outcome Y_l, while the other receives X_h and outcome Y_h. Allocative efficiency does not hold here since redistributing funds so each population group receives the same expenditure would increase outcomes in total. This is because the increase in expenditure to the low-resource population (l) results in greater gains in outcome than the loss in outcome from decreasing expenditure to the high-resource population group (h). The four main ways in which inaccuracies arise are described below.

Utilization methods to model need

Utilization data is largely employed in resource allocation formulae to help model the needs of different geographical populations. This assumes that higher utilization of services reflects a higher capacity to benefit from expenditure. There are a variety of methods available to weight geographic populations based on utilization data. For example, one can weight populations based on the share of total utilization, or apply average utilization rates to the age and/or gender breakdown of geographic populations. Detailed econometric methods can also be applied, enabling a multitude of population characteristics to explain variations in utilization across population groups. This approach regresses geographic population level utilization on variables thought to identify needs. The estimates from the regression are then used to weight the geographic population's share of the budget to reflect the differences in need across each geographic population.

Analyzing utilization data to direct budget allocations poses two potential problems, the first being that only observed (met) needs are modeled. Under- (and over-) utilization by population groups will be sustained in this setting. The second concerns the effect of supply-side impacts on utilization. Variations in utilization may be because of differences in provider provision of services rather than need.

Unmet need

Unobserved (unmet) need can either be general or specific. General unmet need is nondiscriminatory across all geographic population groups, such as a lack of health care services for all. This does not bias the formulae since the unmet need represents a common proportionate lack of health care meaning relative needs weights to population groups remain valid. Specific unmet need is discriminatory, affecting certain geographic population groups because of their population makeup. For example, if utilization is lower for minority ethnic groups, population groups with higher rates of minority ethnic groups would receive a lower weighting than would be if the needs of minority ethnic groups had been observed. To control for specific unmet need, estimates with the 'wrong' sign in regressions explaining utilization may be held back from being part of the funding formulae weights (for example, negative estimates for minority ethnic groups where no clinical explanation for an ethnic difference is identified).

Illegitimate supply

Illegitimate supply-side factors are where variations in utilization reflect providers' provision of services rather than the needs of population groups. For example, utilization approaches could reinforce any existing inefficiency in maintaining service provision levels that may in part reflect an access to health care issue (such as differences in waiting times for hospital surgery or in the availability of General Practitioners across geographic population groups). Methods to control for any illegitimate supply-side factors include the addition of geographic population group dummies in a regression to control for group average variations that are over and above the observed variations across geographic population groups.

Age and gender weighting

Age and/or gender weights may be applied to the resource allocation formulae, in addition to, or in place of, utilization modeling. The motive for such an inclusion is the notion that health needs vary over the life cycle and by gender. To be accurate these must fully explain variations in need. This, however, is unlikely, given numerous other factors including socioeconomic status, education, and sociodemographic characteristics have been found to explain variations over and above age and gender.

Incorporating concerns for health inequalities

Geographic populations may differ, in part, because of demographical or epidemiological factors. Differences in geographic populations may mean that the same level of expenditure will result in different health outcomes between geographies. That is, differences in productivity result in different PPFs across geographic populations. How productive a population group is determines the height and/or slope of the PPF. Differences in the height of PPFs across population groups do not alter the allocations made under pure efficiency. However, differences in the slope of the PPF across geographic populations have important implications for the pure efficient allocation. For the payer to maximize outcomes, it has to adjust expenditure and outcomes by marginal need. Needs adjustment will make the marginal social benefit from expenditure equivalent across geographic populations, meaning the same sloped line is applied to all geographic populations, irrespective of their PPF. Applying the same sloped line means across each geographic population the marginal social value and marginal benefit are equal at the allocative efficient levels of expenditure. At these levels an additional increase in expenditure results in the same increase in benefit (health) for each geographic population. This can be seen as the pure efficiency solution. Any variation from this leads to a net reduction in the sum of outcomes.

Assuming needs have been fully captured in the resource allocation formulae, the pure efficient solution is consistent with different geographical populations attaining different levels of health. For example, one population may be less productive, with a lower PPF, shown by P_1 in **Figure 3**. With no funding formulae, the payer may divide the budget into two, say X^*. At X^* the capacity to benefit from expenditure is greater for the less productive population. Utilization rates may reveal that the less productive population has higher utilization because of the population being relatively older than the more productive population. Weighting X^* on the basis of differences in utilization by age may direct allocations to X_1 and X_2. Under the pure efficiency outcome (where N_1 and N_2 have the same gradient), although the less productive population group receives a higher budget allocation than the more productive population, $X_1 > X_2$, there are differences in health outcomes across the two populations, $Y_2 > Y_1$.

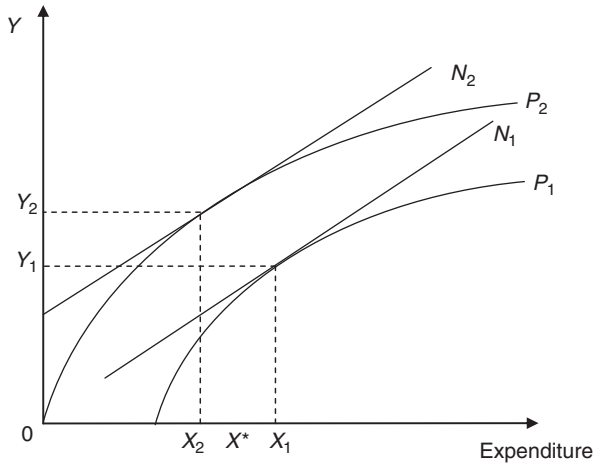


Figure 3 Heterogeneity in PPF across population groups. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

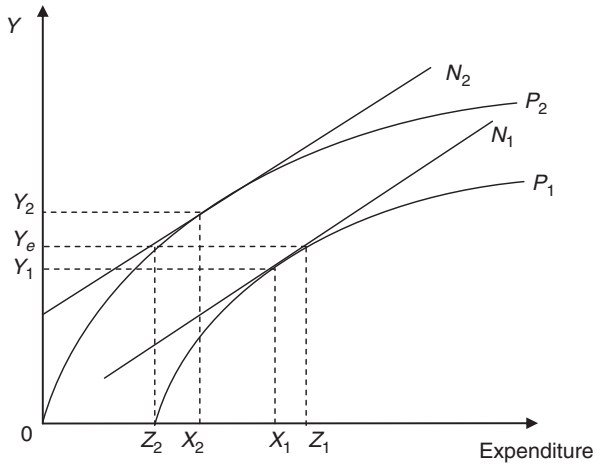


Figure 4 Changes in expenditure to achieve equal outcomes for two populations with different PPFs. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

However, differences in health outcomes across population groups may not be seen as equitable. Policy aims have an important impact on how this inequality should be targeted. **Figure 4** gives the expenditure required to ensure both population groups in the example attain the same health, Y_e. The pure efficient solution has differing health outcomes and payments, so to attain Y_e requires increasing allocations to the less productive population ($Z_1 > X_1$), and decreasing

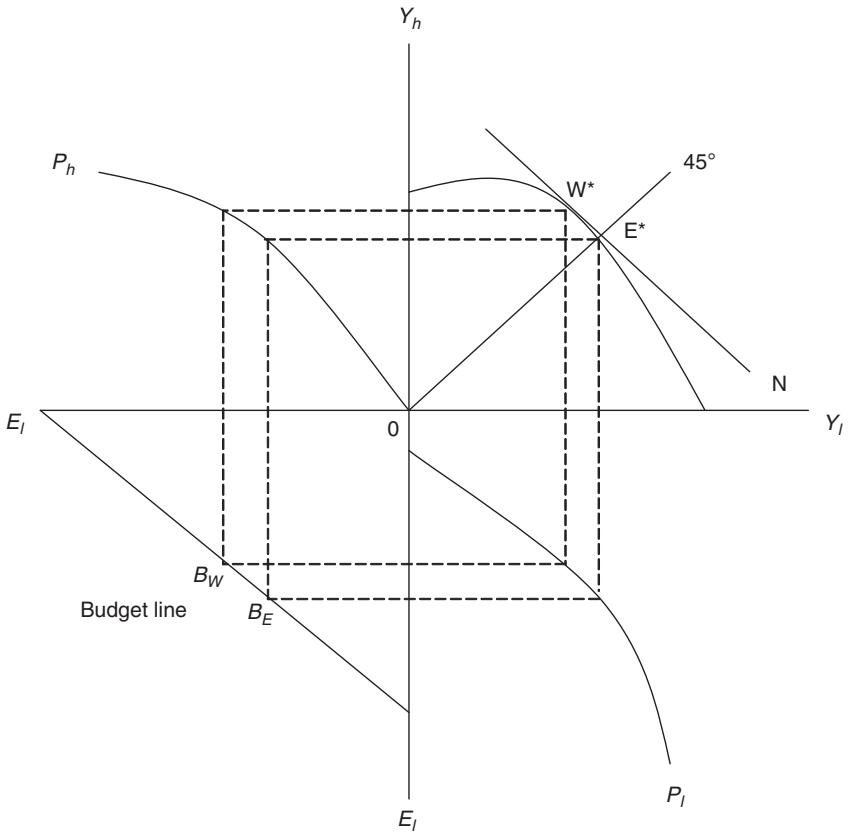


Figure 5 Efficiency-equity trade off. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

allocations for the more productive population ($Z_2 < X_2$). As both population allocations are now allocatively inefficient (but still technically efficient), the allocation is less efficient ($Y_1 + Y_2 > 2Y_e$).

Equity versus efficiency

Figure 4 highlights the trade-off in health outcomes when equity or efficiency are maximized. An alternative way to present the trade-off is given in Figure 5. The North-West and South-East quadrants give the PPF for the high productive and low productive populations respectively; the North-East quadrant maps the two production frontiers together to attain total output possibilities between the two populations. The South-West quadrant gives the budget line, with all points on the line sum to the total budget available. Weighted capitation would provide the weight applied to each population. The total outcomes for the two populations are obtained by: allocating expenditure between the two in the south-west quadrant, mapping these to the PPFs for each population in the North-West and South-East quadrants, and finally mapping these to the production possibility frontier in the North-East quadrant. The line N in the North-East quadrant gives the relative marginal social value of each population's outcome. Here it is assumed to be symmetric. The payer's welfare function (and hence slope of N) will dictate how allocatively efficient the allocation is (note it is still assumed technical efficiency holds). The allocations under the pure efficiency solution, where total outcomes are maximized and the marginal capacity to benefit from additional expenditure is equivalent across population groups (W^* in the North-East quadrant), would be point B_W. The alternative equal outcomes objective, where equal outcomes are attained between the two groups (E^* in the North-East quadrant), would be under allocations at B_E. Total outcomes will be inefficient (lower) under the equal outcomes solution (E^*) than the efficient solution (W^*).

Avoidable inequalities in health

One implication of populations having different PPFs is that some populations may never be able to attain a state of health that equals the average across the total population. Figure 6 gives an example where the maximum possible health achievable is lower for one population than the average of the two (Y_e is above the PPF for the less productive group at all points). Figure 7 gives the four quadrant representation. In this example, any attempt to achieve equal outcomes in health will fail (no amount of allocation could ensure the less productive population will reach Y_e). The best the payer can hope to achieve is to ensure the removal of *avoidable* inequalities in health outcomes. In terms of Figure 7, this means allocating at the point where the PPF for the less productive population begins to flatten (the maximum that population group can achieve). Note how, compared to the pure efficiency solution (W^*), the equal health outcome (AE^*) requires a relatively

greater loss in potential health for the more productive population than the negligible gain in health for the less productive population.

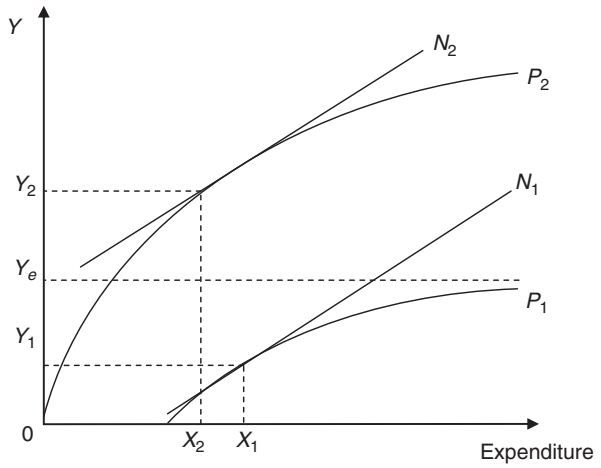


Figure 6 Avoidable inequalities in health. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

Differing costs in the delivery of health care

The costs of delivering health care may vary across geographic population groups, and not controlling for this will inflate (deflate) allocations for low (high) cost population groups. Effectively, differing costs mean allocated budgets are worth more (or less) in different geographic populations. This may be seen as a reduction (increase) in the budget allocations for high (low) cost geographic population group, making allocations no longer efficient. This is particularly problematic if there exists a correlation between costs and productivity. For example, if less productive population groups face higher costs then allocations will not be at the level sufficient to reflect the expenditure required to deliver the health care for the respective need.

Measuring allocative inefficiency

The issue concerning the resource allocation formulae is one of determining how to distribute the budget across population groups. Figure 8 gives the four quadrant representation of Figure 3, where it is assumed that the payer's objective is the pure efficient outcome with the resulting maximization of output, W^*. Assume however, that the weighted capitation formulae have failed to identify needs of the population groups accurately. The result is higher payments to the more productive population ($X_{2A} > X_2$) and lower payments to the

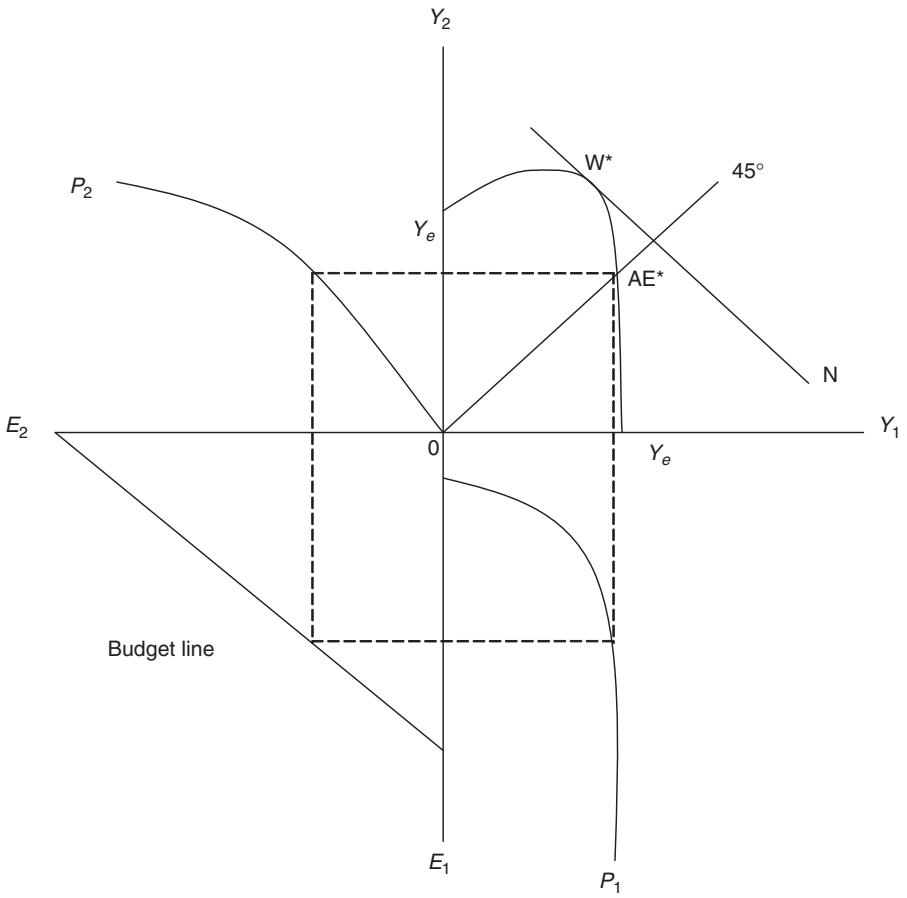


Figure 7 Avoidable inequalities in health. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

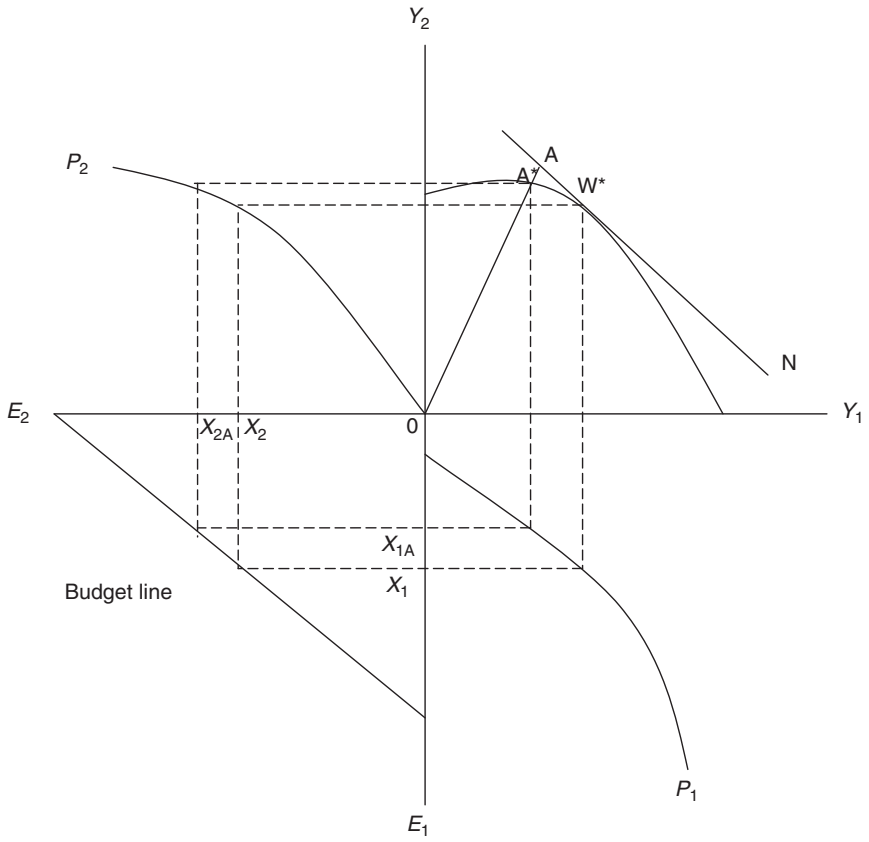


Figure 8 Allocative efficiency in society. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

less productive population ($X_{1A} < X_1$) than would have been if needs were correctly identified. The resulting outcome is at A*, where the level of allocative inefficiency is $0A^*/0A$. The failure to identify needs correctly results in a lower output of health $A^* < W^*$.

Assuming the more productive population has lower costs in the delivery of health care, the effect of differences in the cost of delivering health care across population groups could also be seen in Figure 8. Here the differences in costs inflate the budget allocated to the high productive population ($X_{2A} > X_2$), and deflate the budget allocated to the low productive population.

Technical inefficiency

Until now it has been assumed that the budgets allocated to population groups are spent efficiently, utilizing the budget to attain the maximum health possible. There is, however, an ever growing number of studies that investigate the efficiency of providers. Hollingsworth *et al*. (1999) provided a review of methods used to model technical efficiency. The potential for technical inefficiency arises because of factors within the formulae, within the structure of the health care system, and external to health care.

Within the formula: Budget risk

A key issue arising with budgeting health care funds is that of budget risk. Because of uncertainty in the demand for health care, variations in practice across providers, and potential errors in the capitation formulae, health care expenditure is unlikely to match budgets allocated. This may lead to resources being used on the basis of budget availability rather than need. Budget risks are likely to increase the smaller the geographical area used, the shorter the time horizon, and for more limited types of care.

There may be the possibility that the resource allocation formulae present incentives to be technically inefficient. For example, if budgets are directed toward high need populations, this may present an incentive to undertreat populations to sustain high(er) budgets.

Further complications with the budgets arise since any budget underspend is not redistributed. There is an incentive for underbudget providers to behave inefficiently, that is, spend over and above what they (efficiently) need to provide health care to use up the budget.

Since most resource allocation formulae incorporate past usage (as used in utilization models) any inefficiencies in provision is reinforced. The needs weights will be generated from utilization that is potentially inefficient and inequitable. If past allocations were inefficient future allocations based on these prior allocations will also be inefficient. To ensure inefficiencies are not sustained, the causes of the inefficiencies should be targeted, for example, access to services. A potential way forward would be to incorporate measures of within-provider efficiency and equity into the formulae. For example,

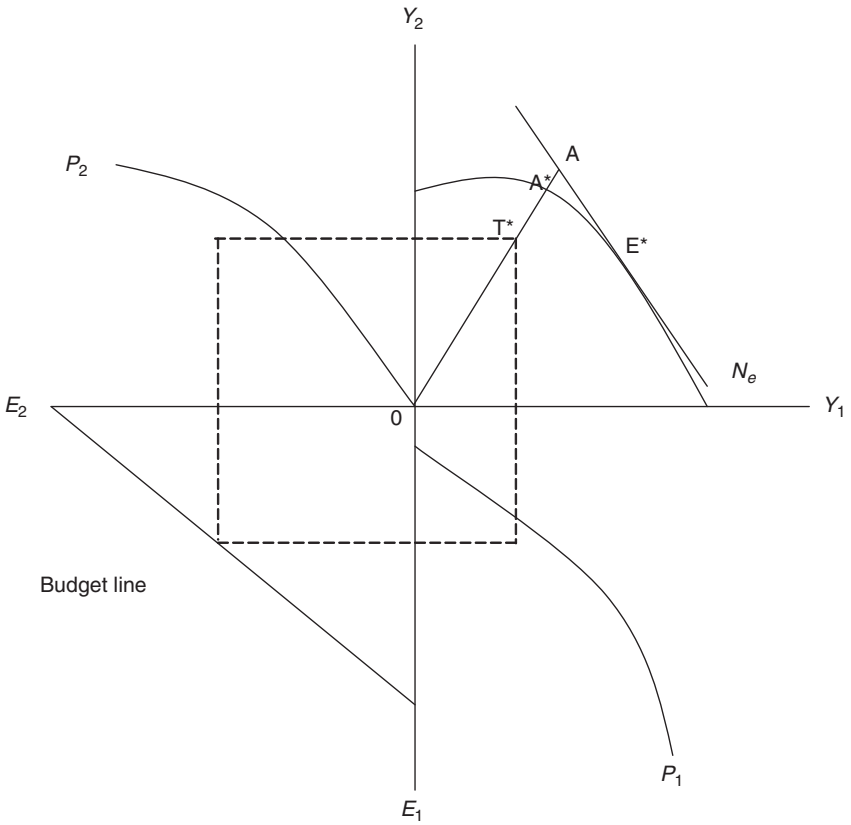


Figure 9 Total efficiency of the resource allocation formulae. Adapted from Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

budget allocations may include a weight for relative efficiency (however this is measured) or for relative equity in health (however defined), which could penalize providers who do not channel budgets to target inefficiency and poor access.

Within the health care system: Market failure

The market structure of the health care system could lead to inefficiencies. For example, in England the NHS is essentially a monopolistic employer and provider of health care resources. The NHS is not subject to competitive market forces that may provide pressure to ensure the NHS system works efficiently, at lowest cost with most impact. For example, the frequency of restructuring the administration of the NHS highlights the struggle to streamline NHS administration and 'red tape' present in the system: allocations were originally made to 14 geographic population groups, but this figure has grown over the years: for 1996–97, to 100; for 2003–04, to 303; and since 2006, to 152 geographic population groups.

External factors

There are also economic influences that may affect the overall delivery of health care. The recent recession, for example, has led to pressure on publicly funded health care systems' budgets, which have been cut in real terms. The effects of this could, for example, lead to the encouragement of better (more) efficient procedures and structures both between and within the sectors and services of the health care system.

Impact on efficiency

In terms of the example, Figure 9 shows the implications of technical inefficiency. Now the outcome (T^*) in the North-East quadrant is not on the PPF of both population groups. Relaxing the assumption of technical efficiency would mean allocations do not produce the maximum health for populations. One moves south and/or westwards in the north east quadrant reflecting reductions in output because of technical inefficiency.

The Current NHS Resource Allocation Formulae in England

In England the amount of public funds allocated to health care from taxation is politically determined. The Treasury sets aside a total spend on the NHS health care in the annual Budget. Of this, the Department of Health (DH) then assigns the total budget for the three largest 'programmes' of NHS service. These are Hospital and Community Health Services (79% of the budget), Prescribing (11%), and Primary Medical Services (10%). Within these budgets, weighted capitation methods are applied to generate a target share of the budget for each Primary Care Trust (PCT), of which there are currently 152 across England. PCTs receive a lump sum to spend across the three programmes as they deem fit.

The DH has used variations of weighted (risk adjusted) capitation payments for distributing budgets for NHS health

care across England since 1977–78. The aim of the current formulae used to determine the capitations is to enable PCTs "to commission similar levels of health services for populations with similar need, with the further objective, since 1999, of helping to reduce avoidable health inequalities" (Department of Health, 2011, p.7).

The services provided by the NHS are not free to the user, because, while free at the point of consumption, individuals pay indirectly via taxation. The taxes used to fund the NHS health care services are substantive: the current weighted capitation approach is used to allocate some $132 billion to PCTs for the 2011–12 financial year, and the NHS as a whole accounted for over 8% of UK Gross Domestic Product for 2010–11.

The weighted capitation formula generates target allocations based on PCT populations, adjusted for: the age distribution; additional need over and above observed age structure; and differences in costs of delivering the services across PCTs – the Market Forces Factor (MFF). A separate index is given to each of these three components, and these are multiplied together to give a weighted population for each PCT in each program (see eqn [1]).

$$\begin{aligned}\text{PCT weighted population} = {} & (\text{population}) \times (\text{age index}) \\ & \times (\text{additional needs index}) \\ & \times (\text{MFF index}) \end{aligned} \quad (1)$$

These weighted populations are then combined according to the share of each program to the total budget. The weighted population for each PCT is then multiplied by the total budget to give a target allocation. PCT allocations, however, are not solely determined by weighted capitation methods. Actual allocations are obtained by taking the difference between target allocations and the previous year's allocation (adjusted for any transfers in responsibilities). A 'pace of change' policy then sets the differential growth in allocations that PCTs receive in addition to the previous year's allocation. The growth component is determined by national and local priorities, and the distance between last year's allocations from target allocations.

Different weighted capitation models are used both within and between each program. Within the Hospital and Community Health Services program, for example, the measurement of need is done separately for five sectors: acute care, maternity services, mental health services, human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS) treatment, and HIV prevention. For prescribing budgets, a MFF is not required because of nationwide prescribing costs. Utilization models account for 90% of each needs index, with the remaining 10% accounted for by an avoidable health inequalities factor.

Changes in the NHS Resource Allocation Formulae over Time

There have been significant changes to the England NHS resource allocation formulae since the NHS inception in 1948.

Geographic population groups (the population index)

The population index provides the basis of the capitation formulae. Population size was introduced into the formulae from 1971, and methods used to obtain these weights have become increasingly more accurate.

First, the differing administrative boundaries for the allocation of budgets have become more refined, from the original allocations across 14 Regional Health Authorities in 1971–95, to: 100 Health Authorities in 1996–02, 303 PCTs in 2003–08, and now 152 PCTs. A higher level of disaggregation increases the accuracy in identifying need by providing greater variations across population groups.

Second, the methods used to obtain population data have also become more advanced. From 1999 population size has been derived from 'constrained' population data. This calculates the number of people registered with General Practitioner (GP) surgeries and unregistered patients living in the area. Constrained population data direct budgets to the population the area serves, reducing the inefficiency of directing budgets on the basis of residence (which had been the main population measure), which may under- or over-estimate the populations served in the area.

From 2006 population projections from the Office of National Statistics (ONS) have been used, which given allocations are based in advance, are thought to better reflect the population served by an area. Population projections incorporate trends in birth, death, and migration.

The changes in the population index would have meant a more accurate measurement of PCT populations. It is important to ensure accurate population modeling since any inaccuracy can divert allocations away from need.

Need measurement (age and additional needs indices)

There have been a number of reviews of the resource allocation formulae, where each has mainly been because of developments in the needs weights. Below the major developments are picked out.

The UK NHS was introduced in 1948, and at the time there was no defined procedure in place to allocate healthcare budgets across the country. The NHS was faced with funding the hospitals, beds, and staff that it had taken over. The initial method for resource allocation was to sustain funding of these services, irrespective of the differing need for healthcare services across populations of England. Funding continued in this way until 1971.

By 1971, a new funding formula, 'The Crossman Formula', was implemented. The formula contained three elements: population size weighted by average bed days by age and gender; bed days weighted by a national cost per bed per year; and cases (inpatient, outpatient and day cases) weighted by the national average cost per year. The inclusion of a proxy measure of need such as bed days meant that resources were being directed toward areas where they were most needed. The inclusion of bed and case volumes, however, still maintained a proportion of the budget that was allocated based on current/inherited supply.

Criticism of the lack of a needs-driven approach resulted in the introduction of the Resource Allocation Working Party (RAWP) formula in 1976. RAWP was set up to attain a resource allocation formula that objectively, equitably and efficiently responded to relative differences in need. The RAWP formula was a weighted capitation approach to allocating resources across the country. The formula now derived

capitation payments weighted by need. Need was measured by age and gender, and noting how age and gender alone would be insufficient to approximate need, 'additional need' was calculated using Standardized Mortality Ratios (SMRs). For the first time, the resource allocation formulae now contained non-utilization measures for need.

The RAWP formula was in place until the 1988 review of the formula was implemented in 1990. For the first time, empirical estimations on the variations of need factors on health care utilization were modeled using regression analysis of hospital utilization. Regression analysis was creating a shift from value judgments on the weighting of needs and permitted the weights on the needs variables to be adjusted for supply factors.

With the release of 1991 Census data, a further review was made in 1994 which was carried out at the University of York (Carr-Hill *et al.*, 1994). To remove the potential endogeneity of supply, significant supply variables were removed from the regression and the regression was then reestimated.

In 1995, a Resource Allocation Group (RAG) was set up to assess resource allocation in primary care. RAG was replaced in 1997 by the Advisory Committee on Resource Allocation (ACRA), an independent expert body to ensure resources for primary and secondary care fully reflect local population needs. The Technical Advisory Group (TAG) was also set up to provide technical support.

In 2002, a further review was conducted. The new formula incorporated more updateable deprivation measures (the Indices of Deprivation), and, for the first time, a measure of unmet need. Unmet need was modeled by maintaining the coefficients with incorrect signs in the models for the descriptive regression analysis but not using these variables as prescriptive weights in the weighted capitation formulae. Supply-side variables were now maintained in the regression but like those measuring unmet need, were not used to obtain weights for need.

There are four key factors in the measuring of need: observing legitimate met and unmet need, controlling for illegitimate need by the modeling of supply-side factors on utilization, and the inclusion of a health inequalities adjustment. Econometric techniques and updates in data availability have improved the accuracy of measuring met need, and the removal of illegitimate supply-side measures of need have arguably lead to greater efficiency in the allocation of budgets.

Unmet need was not modeled until 2002, and is intrinsically difficult to capture, particularly since most resource allocation formulae use utilization data to model need. Even if unmet need is accurately modeled, this may still be inefficient unless procedures are put in place to target the removal of under-utilization of these groups. With no stipulation of how providers should spend their allocated budgets, unmet need 'premiums' could implicitly generate technical inefficiency if these premiums are spent on services that would be relatively less productive in health.

Differing costs in the delivery of health care (the MFF index)

The RAWP formula was amended in 1980–81 to weight capitations on the basis of unavoidable differences in the costs of delivering health care across the country. The MFF was applied to account for differences in staff costs. The introduction of this factor removed the inefficiency of under/over funding because of the differences in the cost of health care by region. The MFF has been updated and expanded regularly. In 2002, the number of pay zones (used to calculate the MFF index) increased from 78 to 117 and a smoothing technique was applied to reduce sharp drops in wage rates for neighboring areas. In 2006, there were further increases in pay zones to 303. In 2008 ,more up- to- date data were used, doctors and dentists were given their own weighting, and further smoothing techniques were implemented. Another review in 2011 updated the MFF with more recent data.

The MFF aims to correct for differences in the cost of delivering services across PCTs. Assuming high cost PCTs serve relatively more productive (in health) populations, the MFF moves budgets to a more efficient allocation, through aiming to be more equitable by adjusting for unavoidable differences in the cost of delivering health care between PCTs. This diverts budgets away from less productive populations, which again, could raise concerns over equity.

Payer's equity concerns

The Department of Health and Social Security (1976) review of the resource allocation formulae set the objective "to secure, through resource allocation, that there would eventually be equal opportunity of access to health care for people at equal risk." This objective is consistent with the pure efficiency solution where equal risk translates to equal marginal benefit from additional expenditure.

In 1998, ministers announced a new objective of the formulae: that of "contributing to the reduction in avoidable health inequalities" (Department of Health, 2011). This was incorporated into the 2002 review of the formula by the introduction of a health inequalities adjustment (based on years of life lost, measured as deviations from the average mortality rate in England). In 2008, ACRA introduced a new health inequalities adjustment, Disability Free Life Expectancy (DFLE). The DFLE combines mortality and morbidity (measured by limiting long-standing illness) data to generate expected years from birth that are free from disability or limiting long-term illness, and is compared to a baseline of 70 years.

The formulae have moved from the pure efficiency solution toward a more equitable solution (from W^* to AE^* in **Figure 7**). The health inequalities adjustment is currently given a weight of 10% in the needs indices, but this has no statistical basis.

Overall Impacts of the Formulae Changes on Efficiency

The improvements in modeling geographic population groups, need and the incorporation of the MFF are likely to have reduced inefficiency in budget allocations enabling a better possibility of maximizing health (the pure efficient solution: W^* in **Figure 7**). The inclusion of an avoidable health inequalities measure from 2002, however, means it is unlikely the pure efficient solution is met. Rather, the best the payer can hope to achieve would be somewhere between W^* and AE^*in **Figure 7**, resulting in relatively lower level of health (the point between W^* and AE^* will depend on how much weight is applied to the health inequalities adjustment). In addition, the inclusion of

past allocations and the pace of change policy to formulate actual allocations make any movement toward the desired outcome slower. Including past allocations also creates issues about technical efficiency, in particular, budget risk. Under spending providers have an incentive to inflate their spending and such activities could become self-perpetuating.

Conclusion

The most efficient resource allocation funding formula requires the relative needs of the populations the payer serves to be identified. Identifying needs accurately enables the weighted capitation approach (commonly used in developed countries to allocate health care budgets) to direct budgets to those areas most efficient in producing health. This approach is consistent with higher budgets allocated to less productive populations on the basis of need, but recognizes that the maximization of health output may lead to inequalities in output between each population group.

There are a variety of reasons why the funding formulae may not be efficient. The reliance on accurately measured needs is a key concern, requiring the ability to disentangle supply-side factors, and understand and recognize the impacts of unmet need. Another key issue surrounds the adjustments that may be included for health inequalities. Equity and efficiency aims do not necessarily have to be in conflict – but they often are. Perhaps the most striking and complex impact on efficiency lies in the activity of the providers of health care. Technical inefficiency may be caused by a number of factors, and each detracts from the resource allocation funding formulae being efficient.

See also: Adoption of New Technologies, Using Economic Evaluation. Efficiency and Equity in Health: Philosophical Considerations. Efficiency in Health Care, Concepts of. Evaluating Efficiency of a Health Care System in the Developed World. Health and Health Care, Need for. Health and Its Value: Overview. Production Functions for Medical Services

References

Carr-Hill, R. A., Hardman, G., Martin, S., et al. (1994). *A formula for distributing NHS revenues based on small area use of hospital beds*. York: Centre for Health Economics, University of York.

Department of Health. (2011). *Resource allocation: Weighted capitation formula*. 7th ed. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/152060/dh_124947.pdf.pdf (accessed 24.06.13).

Department of Health and Social Security. (1976). Sharing resources for health in England. *Report of the Resource Allocation Working Party*. London: Her Majesty's Stationery Office. Available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4121873 (accessed on 24.06.13).

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics* **17**, 1107–1128.

Hollingsworth, B., Dawson, P. and Maniadakis, N. (1999). Efficiency measurement of health care: A review of non-parametric methods and applications. *Health Care Management Science* **2**(3), 161–172.

Smith, P. C. (2007). *Formula funding of public services*. London and New York: Routledge Taylor & Francis Group.

Further Reading

Culyer, A. J. and Wagstaff, A. (1993). Equity and equality in health and health care. *Journal of Health Economics* **12**, 431–457.

Hauck, K., Shaw, R. and Smith, P. C. (2002). Reducing avoidable inequalities in health: A new criterion for setting health care capitation payments. *Health Economics* **11**, 667–677.

Jacobs, R., Smith, P. C. and Street, A. (2006). *Measuring efficiency in health care*. Cambridge, UK: Cambridge University Press.

Relevant websites

http://www.dh.gov.uk/en/Managingyourorganisation/Financeandplanning/Allocations/index.htm
Department of Health Website for allocations funding.

http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_124949
Exposition book (current allocations).

http://www.dh.gov.uk/en/Managingyourorganisation/Financeandplanning/Allocations/DH_076396
Weighted Capitation Formula.

Risk Adjustment as Mechanism Design

J Glazer, Boston University, Boston, MA, USA, and Tel Aviv University, Tel Aviv, Israel
TG McGuire, Harvard Medical School, Boston, MA, USA

Glossary

Managed competition A policy toward individual health insurance markets using both competition among plans and regulation of premia and benefits. Usually includes some risk adjustment of plan payment.

Mechanism design A field of game theory in which a principal tries to motivate an agent by designing a mechanism for rewarding the agent.

Optimal risk adjustment Refers to a risk adjustment system in which a regulator finds the form of risk adjustment of plan payments to maximize some economic objective, such as minimization of efficiency problems due to adverse selection, a form of mechanism design problem.

Service-level selection Action by a health plan to provide more or less of certain services in order to attract or deter potential enrollees who would lead to profits or losses.

Introduction

In many countries, residents choose a health plan or sickness fund through which to receive health insurance benefits. These choices are regulated and at least partially paid for by governments and employers. Collective financing of health care redistributes the burden of cost from the sick to the healthy and from the poor to the rich. At the same time, societies seek the virtues of markets: choice, innovation, and price and quality competition from their health insurance plans. Melding the desires for both a fair yet controlled and efficient yet innovative health insurance sector is a central problem facing all developed nations. A common approach to this problem consists of national governments collecting the funds to pay for health care, but then passing responsibility for the purchasing of health care to a local organization, a private insurance plan as in the federal Medicare program in the US, a local government in the UK, Canada and Australia, or sickness funds as in Germany, Israel, the Netherlands and Belgium. Governmental involvement aims at a fair distribution of the cost burden, and competition among the decentralized participants is intended to promote efficiency. A critical element of the policy is paying more for the sick and less for the healthy as they join plans – the job of risk adjustment.

Defining Risk Adjustment

In this article 'risk adjustment' is referred to as a formula relating payment to a provider or a health plan to observable characteristics of a person (such as age, previous diagnoses). In the Medicare program in the US, Germany, Netherlands, Israel, and other countries, risk adjusted payments flow from governments to health plans when individuals enroll in private plans. Large provider groups are increasingly asked to bear risk, (this is the primary direction of US health policy) and risk adjusted payments come into play there as well. As provider payment and risk bearing become more central in health policies, the statistics and economics of risk adjustment take on increasing importance.

The evolution of health policy has put new demands on risk adjustment. This article reviews the two basic methods for deriving a risk adjustment, one primarily statistical, and a second introducing an economic objective into the statistical analysis. In the second method, finding the right risk adjustment is a problem in mechanism design. The authors begin with the basic model of adverse selection and derive the implications for risk adjustment. They then consider two applications with a close connection to the empirical methods that can be used to estimate risk adjustment weights.

The most common approach to risk adjustment is statistical. Conventional risk adjustment sees the goal of risk adjustment as matching payments to expected cost as closely as possible. If an older enrollee is expected to be twice as expensive as a younger enrollee, conventional risk adjustment would pay twice as much for the older enrollee. Many factors other than age matter for expected costs. Research on conventional risk adjustment is statistical and data oriented. Researchers seek to find the right combination of variables (referred to as risk adjustors) to include in regression models so that the explained variation in health care costs is high, without relying on risk adjustors that are difficult to collect in practice or can be manipulated by providers seeking to increase revenue. There is also discussion of whether certain variables should be recognized as part of risk adjustment, for example, if an individual chooses to smoke, whether he or she should be 'rewarded' by a higher risk adjusted payment. The premise behind this research – sometimes regarded to be so obvious as to not require justification or analysis – is that the healthcare market in question will function better, the better the job the regression model can do in predicting healthcare costs of enrollees.

By contrast, optimal risk adjustment views risk adjustment as a set of incentives to address an economic problem. Calculating the optimal risk adjustment begins with an explicit conception of how the relevant market functions, which relates the risk-adjusted price (e.g., the payment for young and old) to the behavior of payment recipients. The economic objective (usually efficiency) is also stated explicitly. Then, using mechanism design, the optimal risk adjustment is derived as the prices for young and old which maximize the

efficiency of the health care market. Optimal risk adjustment does not refer to particular weights, but rather to a methodology by which the optimal weights are obtained. Optimal risk adjustment also relies on data, however, the optimal weights are not, in general, regression coefficients but a solution to a problem of economic maximization.

The Basic Adverse Selection Problem and the Role of Risk Adjustment

A health plan can underprovide some services and overprovide others, attracting the low risks and deterring the high risks. The basic idea draws on the early analysis of insurance by Rothschild and Stiglitz. Demand for treatment of chronic conditions, for example, may be much better anticipated, and more unevenly distributed in a population, than demand for acute care. In such a case, the health plan has a financial incentive to distort the mix of its care away from chronic care and toward acute illness, in order to deter/attract the high/low risks. Nearly all writers on the efficiency of health insurance markets with managed care, acknowledge this effect, though they vary in the emphasis they put on it. When a plan can set premia as well as quality, a version of this strategy is to provide low quality overall, and set a low price, to attract the low-risks.

This quality distortion problem has received a good deal of attention in health economics literature. The basic adverse selection model is presented here. Suppose that there are two types of individuals, L and H, who can contract two illnesses, a and c. Illness a we call an acute illness and both types of people have the same probability of contracting this illness, $p_a > 0$. The two types are distinguished in their probability of contracting the chronic illness c. Let p_i, $i \in \{H,L\}$ denote the probability that a person of type i contracts illness c. Then, $p_H > p_L > 0$. The proportion of H types in the population is λ, $0 < \lambda < 1$. Let $p_c \equiv \lambda p_H + (1-\lambda)p_L$ denote the (expected) probability that a person randomly drawn contracts the chronic illness. Throughout the analysis it is assumed that each individual knows their type. It is also assumed that each individual must choose one plan.

If a person (of either type) has illness j, $j \in \{a,c\}$, their utility from treatment will be increased by $V_j(q_j)$, where $q_j > 0$ denotes the 'quality' of the services devoted to treat illness j, with $V'_j > 0$ and $V''_j < 0$. If a person has both illnesses, their utility, if treated, will simply be increased by $V_a(q_a) + V_c(q_c)$. Treatment services are provided by health plans. A health plan is characterized by a quality pair (q_a, q_c). Thus, if a person of type i, $i \in \{H,L\}$ joins a plan with a quality pair (q_a, q_c), their expected utility will increase by:

$$U_i(q_a, q_c) = p_a V_a(q_a) + p_i V_c(q_c) \quad [1]$$

Throughout the analysis it is assumed that each plan gets to choose its quality pair and a plan can offer only one quality pair. All plans have the same cost function. A plan's cost of treating a person with illness j, $j \in \{a,c\}$ at a quality level q_j is $C_j(q_j)$, where $C'_j > 0$, $C''_j > 0$. Thus, if a person of type i, $i \in \{H, L\}$ joins a plan that offers a quality pair (q_a, q_c), the plan's costs are expected to increase by:

$$C_i(q_a, q_c) = p_a C_a(q_a) + p_i C_c(q_c) \quad [2]$$

The 'socially efficient' quality pair (q_a^*, q_c^*) equalizes marginal benefit of treatment to marginal cost, thus solving the following pair of equations:

$$\begin{aligned} V'_a(q_a^*) &= C'_a(q_a^*) \\ V'_c(q_c^*) &= C'_c(q_c^*) \end{aligned} \quad [3]$$

High and low risk types have different probabilities of becoming ill, but once ill, receive the same utility from treatment. Thus, the efficient level of quality is independent on the probability of becoming ill and is the same for both types.

It is assumed that the Regulator can enforce an open enrollment policy. The order of moves in our model is as follows: First the Regulator/payer announces r^*, the premium (paid by the Regulator/payer) a plan will receive per enrollee. Next, plans (simultaneously) choose their quality pair (q_a, q_c), then individuals choose plans and plans collect a revenue of r^* per enrollee, finally each individual's health state (whether she has illness a and/or c) is realized and plans pay the costs of treatment. A 'competitive equilibrium' in this market is a set of quality pairs such that, when individuals choose plan to maximize expected utility, (1) no quality pair in the equilibrium set makes negative expected profit, and (2) there is no quality pair outside the equilibrium set that, if offered, will make a positive profit.

Following Rothschild and Stiglitz it is known that if the proportion of the H types in the population is sufficiently large, then a competitive equilibrium exists and is characterized by two quality pairs. H types choose the plan(s) that offer the quality pair:

$$\begin{aligned} (q_a^H, q_c^H) &= \text{argmax } U_H(q_a, q_c) \\ \text{s.t.} \quad & C_H(q_a, q_c) = r^* \end{aligned} \quad [4]$$

and L types choose the plan(s) that offer the quality pair:

$$\begin{aligned} (q_a^L, q_c^L) &= \text{argmax} \quad U_L(q_a, q_c) \\ \text{s.t.} \quad & C_L(q_a, q_c) = r^* \\ \text{and} \quad & U_H(q_a, q_c) = U_H(q_a^H, q_c^H) \end{aligned} \quad [5]$$

The equilibrium is described in Figure 1. The curves r_i^*, $i = H, L$ represent all plans, i.e., pairs of (q_a, q_c), that break even if the plan attracts only individuals of type i, when the premium is r^*. The points denoted by q_i, $i = H$ or L, depict the plan chosen by type i in equilibrium, i.e., $q_i = (q_a^i, q_c^i)$. Curves u_i, $i = H, L$, represent type i's indifference curves that goes through the point q_i. The curve r^* represents all plans that break even if the plan attracts a random sample of the population, and the point q^* depicts the socially efficient levels of quality. It can, therefore, be seen that plans will not offer the socially efficient quality profile in equilibrium.

When the Regulator/payer is using risk adjustment, the premium paid to the plan (often referred to as 'capitation') is conditioned on observable characteristics of the enrollee. The capitation payment might be based, for example, on the enrollee's age, with older enrollees having higher payments associated with them because they are expected to cost more.

It can be illustrated how conventional and optimal risk adjustments are calculated using this model. Assume that the

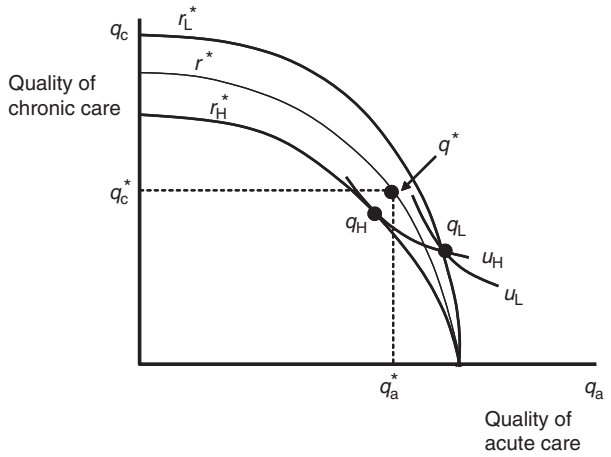


Figure 1 The basic adverse selection result.

Regulator gets a signal s about each consumer's type. The signal could be, for example, the consumer's age. Suppose that s can take a value of 0 or 1 ('young' or 'old'). The signal contains information in the sense that type H person is more likely than type L person to get the signal 1. Let γ_i, i=H or L, be the probability that consumer of type i gets the signal 1. It is assumed that $\gamma_H > \gamma_L \geq 0$. (Note that if $\gamma_H = 1$ and $\gamma_L = 0$, the signal is perfect i.e., the Regulator knows the individual's type.)

Let λ_s be the posterior probability the consumer is of type H given the signal s. Because the signal is informative, using Bayes' rule one can show that $1 \geq \lambda_1 \geq \lambda_0 \geq 0$. Thus, if a person got the signal 1, that person is more likely to be of type H than a person who got the signal 0. Let,

$$P_s = P_H \lambda_s + P_L(1 - \lambda_s) \quad \text{for } s = 0, 1. \qquad [6]$$

and

$$r_s = C(q_a^*) + P_s C(q_c^*) \quad \text{for } s = 0, 1. \qquad [7]$$

P_s is the probability that a person with signal s will contract illness c and r_s is the expected health care costs of such a person at the efficient level of quality of care. Clearly $P_1 > P_0$ and $r_1 > r_0$. One can readily confirm that if plans are paid r_s for each person who got the signal s, and consumers are randomly distributed across plans, plans break even providing the efficient level of care.

The capitation payment r_s is what the authors mean by 'conventional' risk adjustment. It can be shown, however, that conventional risk adjustment does not implement the socially desired outcome, i.e., at the competitive equilibrium, plans do not provide the socially efficient quality. The same forces that break the efficient pooling equilibrium when premiums are not risk adjusted will also break the efficient pooling equilibrium when premiums are conventionally risk adjusted. Market equilibrium under conventional risk adjustment will still be a separating one where the H types and the L types choose different plans with a different quality profile. This separating equilibrium is more efficient (i.e., it induces a higher expected utility) than the one without risk adjustment, but it is not the best the Regulator can do. As demonstrated below, an optimal risk adjustment can be constructed to implement precisely the socially desired quality.

Let

$$C_H^* = C(q_a^*) + P_H C(q_c^*) \qquad [8]$$

and

$$C_L^* = C(q_a^*) + P_L C(q_c^*) \qquad [9]$$

C_i^* is the expected costs of an individual of type i at the efficient quality profile.

The authors are now ready to discuss the conditions under which risk adjustors implement the socially desired contract:

Proposition: Let r_s^*, $s=0,1$ be solution to the following system of equations:

$$\gamma_H r_1^* + (1 - \gamma_H) r_0^* = C_H^* \qquad [10]$$

$$\gamma_L r_1^* + (1 - \gamma_L) r_0^* = C_L^* \qquad [11]$$

then if plans are paid a premium r_s^*, $s=0,1$ for each individual who got the signal s, all plans will offer the socially desired quality in equilibrium.

The left hand side of eqn [10] is the expected premium a plan receives for each enrollee of type H, under the risk adjustment scheme r_s^*. The right hand side of eqn [10] is the plan's expected cost of an enrollee of type H, under the socially desired quality bundle. Equation [10] states the condition for the expected premium for a type H individual to be equal to the individual's expected cost. Equation [11] does the same thing for a type L individual.

It can be easily verified that, if $\gamma_H < 1$ and $\gamma_L > 0$ then $r_1^* > r_1$ and $r_0^* > r_0$. Conventional risk adjustment redistributes some, but not enough, resources from the low-cost to the high-cost types. In Figure 1 this redistribution would appear as a shift in the zero-profit curves relative to the curves in the no risk adjustment case. As the proposition above shows, the Regulator may shift the zero-profit curves even further than is implied by conventional risk adjustment, by 'overpaying' for a consumer who got the signal 1, compensated by 'underpaying' for consumers who got the signal 0, and by so doing, bring the market closer to the socially desired outcomes. 'Overpaying' and 'underpaying' are in comparison to the conventional risk adjustment premiums. Figure 2 illustrates the equilibrium under optimal risk adjustment.

Intuitively, this result can be understood as follows: If the signal is not very precise, the difference in premium conventional risk adjustment pays, for a consumer who got the signal 1 and a consumer who got the signal 0, will be small. Furthermore, the proportion of consumers who got the signal 0 among the L-types is not much larger than the proportion of consumers who got this signal in the entire population. Thus, by offering a quality profile that attracts only the L-type consumers, a plan can reduce its cost by a significant amount relative to the reduction in the premium it is expected to receive. If, however, the premium for an individual who got the signal 0 is significantly lower than the premium for an individual with the signal 1, the plan is severely punished for attracting only individuals of type L.

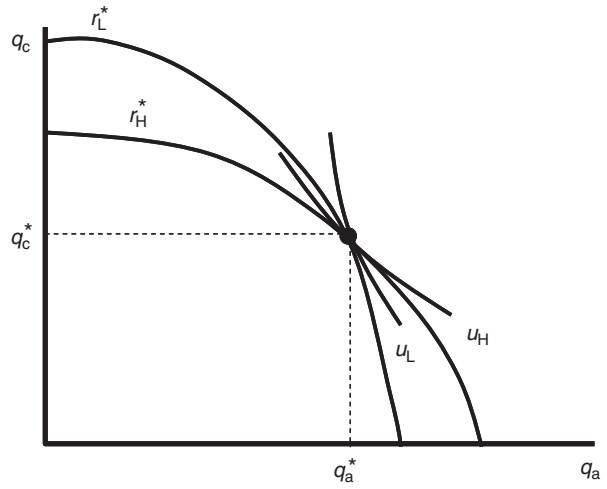


Figure 2 Market equilibrium under optimal risk adjustment.

Multiple Services and Maximizing Fit

The analysis just laid out can be made operational in a more realistic setting in which a plan is providing multiple services (not just two) and in which the Regulator's objective in conventional risk adjustment of 'best fit' is also recognized.

When a plan decreases the stringency of rationing on a service, i.e., increases the level of spending on a service, costs are affected because spending goes up for existing enrollees and spending is incurred on enrollees newly attracted by the spending increase. The idea of optimal risk adjustment is to make sure that for all services a plan is making a decision about, from those used mostly by the healthy and cheap to those used mostly by the sick and costly, these cost increases are balanced against revenue gains to the same degree. This is done by recognizing that enrollees have a payment made on their behalf and the terms of that payment are described by the risk adjustment parameters. There is, thus, a relation between the marginal cost and marginal revenue for each of the services in a plan. Balancing incentives across services essentially amounts to an equation for each service, the level of rationing being the variable and the risk adjustment weights being the parameters.

To equalize incentives in rationing all services, the covariance of the risk adjusted payment with the use of every service must track the covariance of the total predicted costs associated with the increase in use of the service. Intuitively, the optimal risk adjustment formula must have the property that by spending on a service, the cost consequences to a plan relate to the revenue consequences in the same way for all services. It is important to stress that the result for optimal risk adjustment says how a given average payment should be risk adjusted, but does not answer the question of how high or low on average the payment should be.

The optimal risk adjustment emerges as a set of linear equations one for each service, with unknowns equal to the variables available for risk adjustment. An interesting feature of this optimal risk adjustment scheme is that the number of parameters available for risk adjustment could be greater or less than the number of services a plan is deciding about. (Some risk adjustment systems have scores of weights.) If the number of available risk adjustment parameters is larger than the number of services whose quality the plan decides on, there may be many risk adjustors that achieve optimality in the sense of incentive balance across services.

This observation raises the question: Among the ways to set risk adjustment to achieve efficient incentives, which way is best? A natural way to answer this question is to reintroduce the original statistical objective of risk adjustment, best fit. In an earlier paper, the authors show that conditions describing efficient service provision are linear in the risk adjustment weights. These can be introduced as constraints on a least squares risk adjustment regression. This risk adjustor is referred to as the minimum variance optimal risk adjustment (MVORA). It is minimum variance by properties of least squares estimators, and it is optimal because the linear constraints on incentives for efficient service provision are satisfied.

Application to Managed Competition and Enrollee Premiums

Risk adjustment is recognized as an integral part of managed competition policy, with the general objective of making sure plans are willing to accept and serve ailing expensive enrollees as well as healthy low cost enrollees. Managed competition policy also relies on premiums paid by enrollees. If competition is to have its desired effects, plans that are able to provide good care at a lower cost, or provide worthwhile benefit enhancements, need to transmit incentives to consumers through premium competition. If a plan has a better product that consumers are willing to pay for, the logic goes the plan will be able to charge an incremental premium and attract enrollment.

Premium-based incentives are an integral part of health plan payment in the US in two important emerging health policy contexts – the Medicare Advantage (MA) program and Medicare Part D offering private plans in Medicare, and the new state-run 'Exchanges' created as part of the Affordable Care Act (ACA) – plan payments come from two sources at once: risk-adjusted payments from a Regulator and premiums charged to individual enrollees. Premiums also play a role in the Netherlands and Germany, though again important institutional details describe the relationship.

Here the authors consider a general setting of a premium support policy in which a Regulator has a budget to pay plans, and has the ability to risk adjust that budget. In addition to risk adjusted payments, plans must also collect revenue from enrollees through premiums. How should a Regulator set risk adjustment weights if only part of the funding for plans is coming through public budgets and being risk adjusted, and the balance will be set by plans in a managed competition market?

Suppose the Regulator collects some public funds to pay plans, and must risk adjust 75% of costs based on age, gender and previous diagnoses. Enrollee premiums must cover the other 25%, and premiums are conditioned on another,

possibly overlapping set of variables, age, smoking status, and geography. The key insight is that the risk adjustment mechanism adopted by the Regulator affects premiums, because what a plan would want to (from profit-maximization) and would be able to (due to competition) charge enrollees as a premium depends on how the Regulator sets risk-adjusted payments. 'What happens' in this plan market, from a number of perspectives, depends on how these premiums work out. The Regulator needs to consider the effect of the risk adjustment on premiums as part of the answer to how to set risk adjustment weights. The Regulator's problem in the case of premiums conditioned on several variables differs from the case when the Regulator has the same budget but premiums are to be the same for everyone.

The first step is to describe how risk adjustment weights affect premiums. Let the total number of people be N and health care costs of individual i be x_i. People vary in two observable dimensions, according to health status, the basis of risk adjustment, and according to another set of characteristics the authors refer to as personal, the basis of premiums. Health status is indexed by h, $h=1,\ldots,H$; personal characteristics are indexed by t, $t=1,\ldots,T$. For notational simplicity, it is assumed that each of these categorizations is one dimensional and the information is mutually exclusive so that each person is characterized by an (h,t) pair the authors will refer to together as a 'type.'

Define x_{ht} to be the average cost of person of type (h,t), and n_{ht} to be the number of people of type (h,t). Health care costs are plan costs (which must be covered by plan payments) and are fixed (do not depend on risk adjustment or premiums). The risk adjusted payment by the Regulator can only depend on $h{:}r_h$. The premium can only depend on $t{:}p_t$. It is assumed that competition among health plans forces premiums to be zero-profit, meaning, for each premium category, t, the premium for that category is determined by the following condition:

$$\sum_h n_{ht}(r_h + p_t - x_{ht}) = 0 \qquad [12]$$

Equation [12] shows how premiums depend on risk adjustment weights. There are T of these expressions, one for each premium category. In each expression there are H parameters, the risk adjustment weights on category h. The authors now move on to the consideration of a second step in a managed competition context, choosing the risk adjustment weights in light of the presence of premiums in plan payment. There are many applications dependent on the objectives of the Regulator. The authors begin with the most basic.

Suppose the Regulator seeks to maximize fit of the payment system with respect to choice of the risk adjustment weights r_h on the H health status factors, subject to a per-person budget for risk adjustment and subject to how the market will set premiums to be zero profit conditional on risk adjustment, described in eqn [12].

In general, maximizing an objective of the Regulator involving premiums subject to eqn [12] can be addressed as a problem in mechanism design, setting the payment parameters (risk adjustment weights) to maximize the objective subject to the constraints. The budget constraint on risk adjustment is one linear constraint. The set of linear equations in eqn [12] are also constraints. It is important to note in this regard that the constraints in eqn [12] are equivalent to the so-called normal equations in least squares with respect to premium variables. Thus, a least squares regression in which premium categories are added as variables and the Regulator's budget is added as a constraint will find the risk adjustment weights that lead to the best fit (by properties of least squares).

One could also change the objective of the Regulator, introducing concerns for efficient service provision as discussed in the earlier application. Premium support policies raise other issues as well. For enrollees to sort themselves efficiently across plans, the premiums' differences they face should be close to the cost differences they would impose on the plans. This efficiency objective could also be expressed as a set of constraints on premiums.

Final Comment

The most basic implication of economic analysis of risk adjustment is this: When considering design of any mechanism to deal with problems of adverse selection, the nature of the underlying inefficiency and an anticipation of how plans and providers react to the policy should provide the foundation for the analysis. This observation implies that 'conventional' approaches to risk adjustment are not in general optimal, and encourages researchers and policy makers to consider alternatives. To design an efficient risk adjustment payment scheme, one needs to know how plans/providers/patients will react to it. Economic theory can help shed light on this question, but ultimately, plan behavior is a matter of empirical research.

See also: Risk Equalization and Risk Adjustment, the European Perspective. Risk Selection and Risk Adjustment

Further Reading

Breyer, F., Bundorf, M. K. and Pauly, M. V. (2012). Health care spending risk, health insurance, and payment to health plans. In Pauly, McGuire and Barros (eds.) *Handbook of health economics*, vol. 2, pp. 691–762. Amsterdam: Elsevier.

Cutler, D. and Zeckhauser, R. (2000). The anatomy of health insurance. In Culyer, A. and Newhouse, J. (eds.) *Handbook of health economics*. Amsterdam: North-Holland.

Ellis, R. P. and McGuire, T. G. (2007). Predictability and predictiveness in health care spending. *Journal of Health Economics* **26**(1), 25–48.

Enthoven, A. and Kronick, R. (1989). A consumer-choice health plan for the 1990s. *New England Journal of Medicine* **320**, 29–37.

Glazer, J. and McGuire, T. G. (2000). Optimal risk adjustment of health insurance premiums: An application to managed care. *American Economic Review* **90**, 1055–1071.

Glazer, J. and McGuire, T. G. (2002). Setting health plan premiums to ensure efficient quality in health care: minimum variance optimal risk adjustment. *Journal of Public Economics* **84**, 153–173.

Glazer, J. and McGuire, T. G. (2011). Gold and silver plans: Accomodating demand heterogeneity in managed competition. *Journal of Health Economics* **30**(5), 1011–1019.

Newhouse, J. P. (1996). Reimbursing health plans and health providers: Selection versus efficiency in production. *Journal of Economic Literature* **34**(3), 1236–1263.

Van de Ven, W. P. and Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In Culyer, A. and Newhouse, J. (eds.) *Handbook of health economics*. Amsterdam: North Holland.

Risk Classification and Health Insurance

G Dionne, HEC Montréal, Montreal, QC, Canada
CG Rothschild, Wellesley College, Wellesley, MA, USA

Glossary

Adverse selection A situation in which high-risk individuals tend to 'drive out' low-risk individuals from insurance when premium averaging leads the former to detect a bargain and the latter no bargain at all.
Asymmetry of information A situation in which one party in a transaction has more or superior information compared to another, for example, physicians have a greater knowledge than patients of the likely effectiveness of drugs while the patients have greater knowledge of the likely impact of drugs on their family circumstances, or people seeking insurance have more reliable expectations of their risk exposure than insurance companies.
Classification risk The possibility of being initially misclassified and later reclassified as high-risk with limited access to or high premiums for insurance as information is revealed overtime.
Community rating Setting health care insurance premia according to the utilization by a broad population (for example, one defined by employer type or geography).
Coverage mandates Regulations requiring insurance policies that cover certain health conditions.
Cream skimming A form of selection in private health insurance markets by which the insurer obtains a higher proportion of good risks (people with a low probability of needing care or who are likely to need only low-cost care or both) in their portfolio of clients than is assumed in the calculation of the insurance premiums also called 'cherry-picking' and 'creaming'.
Death spiral The complete unraveling of an insurance pool by a feedback loop between rising premiums and exit by the healthiest remaining individuals in the pool that forms an extreme adverse selection.
Distributional equity in insurance The notion that a fair insurance system should equalize premiums and insurance access across higher and lower-risk individuals.
***Ex-ante* efficiency** The best outcome from the point of view of an individual behind a notional veil of ignorance, before one knows what risk type or risk class they will belong to.
Financial equity The notion that two individuals facing the same risks and the same coverage should pay the same premium.
Group equity The notion that no identifiable group should be required to cross-subsidize any other identifiable group.
Horizontal equity The notion that two individuals facing the same risks should have access to the same coverage at the same premium. Treating equally those who are equal in some morally relevant sense that commonly meet horizontal equity principles 'equal treatment for equal need' and 'equal treatment for equal deservingness'.
Incentive contracting The contracts designed to mitigate problems arising from informational asymmetries.
Interim efficiency/interim Pareto efficiency The outcomes that cannot be improved on for any risk type within a risk class without harming another individual.
Moral hazard There are two main types. Ex ante moral hazard refers to the effect that being insured has on behavior, generally increasing the probability of the event insured against occurring. Ex post moral hazard arises because being insured reduces price of care to the patient and hence leads to an increase in demand from insured persons.
Perfect risk classification A risk classification based on observable characteristics that generates insurance premiums that fully reflect the expected cost associated with each class of risk characteristics.
Risk adjustment A technique for adjusting a payment to an individual's risk characteristics usually achieved through government-run transfers across insurers that depend on their insurance pool's risk-related characteristics.
Risk classification in health insurance The use of observable characteristics by insurers to group individual risks with expected medical costs while underwriting insurance policies.
Screening insurance The contracts that are designed to induce individuals with different private information to self-sort into distinct contracts.
Underwriting The process of measuring risk exposure of a potential client and determining the insurance premium to be charged.

Introduction

Risk classification refers to the use of observable characteristics, such as gender, race, age, and behavior, to price or structure insurance policies. Risk classification potentially has undesirable consequences, including adverse effects on distributional equity. In dynamic settings, risk classification can also increase classification risk, which refers to the risk that an individual faces of being reclassified into a higher-cost class at a later date.

A perfect risk classification system should, using actuarial rules and principles, generate an insurance premium that

 doi:10.1016/B978-0-12-375678-7.01316-X

reflects the expected cost associated with a given risk. Two clients with the same risk level should pay the same, actuarially fair premium. This is known as the financial equity criteria. In health insurance, premiums are most commonly determined by age, sex, and smoking behavior. Current medical conditions (high cholesterol, diabetes, etc.) and medical histories of older clients are often added as criteria because they can affect the medical expenses covered by the insurance plan. Information on lifestyle, diet, and exercise can also be considered.

Market forces push competitive insurers toward employing risk classification whenever it is legal (and permissible according to social norms) to do so. For example, age is an easily observable characteristic that is often correlated with expected health care expenditures. If insurers do not price their insurance products on the basis of age, they will find that, on average, selling policies to the lower-risk young is more profitable than selling policies to the higher-risk old. Individual firms will therefore have an incentive to cream skim – to offer a lower-priced insurance product only to the young and thereby to attract only the most profitable risks.

Selection and pricing activity based on individual characteristics is subject to concerns about social fairness (or equity) and potential discrimination. This is particularly true in the medical and disability insurance markets. Policy makers who dislike the consequences of risk classification may therefore find it desirable to use regulatory restrictions to limit its use. Indeed, risk classification is restricted or banned outright in various markets, for example, via community rating laws that require insurers to offer all individuals in a given community the same policies at the same premiums, as the compulsory public health system in Canada and several US states.

The policy decision to restrict the use of risk classification often involves a trade-off between financial and social equity. This trade-off is policy relevant because departures from financial equity can lead directly to inefficient insurance provision. For example, risk pooling arising from legal restrictions on risk classification variables may lead to a situation in which lower-risk individuals are faced with higher premiums than those corresponding to their true risk, whereas higher-risk individuals pay lower premiums. Low-risk individuals may leave the pool, driving premiums higher and causing even more individuals to leave the pool. This inefficient market unraveling is known as an adverse selection death spiral.

Understanding the cost and benefits of risk classification more generally is challenging for at least two reasons. First, there are a number of interrelated and overlapping effects of risk classification that are difficult to disentangle. Second, the relative importance of these various effects depends strongly on the institutional details of the insurance market. In some markets, permitting risk classification facilitates efficient insurance provision without compromising concerns about equity. In some others, banning risk classification has beneficial equity effects without imposing any efficiency costs. In others, the decision to ban risk classification involves a nontrivial trade-off between equity and efficiency goals.

This article provides a simple framework for identifying the types of markets in which these three cases arise. One of the key determinants is the presence or absence of residual asymmetric information in the market. This article reviews empirical tests for asymmetric information in health insurance markets.

Potential Welfare Effects of Risk Classification

Public policy typically involves trade-offs between equity and efficiency. Public policy regarding risk classification in insurance markets is no exception. It is further complicated by the presence of several conceptually distinct but overlapping notions of equity and efficiency.

Equity

There are at least four potential notions of equity in the risk-classification context: horizontal equity, financial equity, group equity, and distributional equity.

Horizontal equity refers to the idea that any two individuals facing identical insurable risks should be treated identically: they should, for example, have access to the same policies and at the same prices. Group equity, however, refers to the idea that each identifiably distinct group (e.g., males, 25-year-olds) should not, as a group, be required to cross-subsidize other groups. A desire for this type of equity is sometimes referred to as subsidy aversion. Financial equity is a special case of group equity in which there is no heterogeneity within the group.

The goals of horizontal and group equity are frequently in tension with each other. By way of illustration, consider a population consisting of otherwise homogenous 30-year-old men and women seeking individual health insurance policies. Suppose that there is only one type of policy available; the only question is, what price individuals will be charged for it? Suppose further that the expected cost to an insurer of providing coverage to a woman is higher, on average, than the cost of providing coverage to a man.

If insurers risk-classify using gender, then women in the population will face higher insurance prices. Group equity will be satisfied at the level of gender, as each gender will be charged an appropriate premium. Insofar as not all women in the group are identical, however, financial equity will not be satisfied. Moreover, some women, perhaps those in good health with no interest in bearing children, are likely to have lower than average expected costs and, similarly, some men are likely to have higher than average expected costs. So it is likely that there are some men and some women in the population with exactly the same expected costs to insurers. Because insurers are risk-classifying by gender, these two identical risks will be charged different premiums for identical coverage. This violates horizontal equity. On the other hand if insurers do not risk-classify by gender, then horizontal equity will be trivially satisfied. Group equity will be violated, however, because the lower-on-average-risk men will be charged the same as women, men as a group will effectively be cross-subsidizing women as a group.

Group and financial equity are founded on an actuarial notion of fairness: What is fair to an individual or group is that they be charged prices in relation to their true cost to an insurer. Like horizontal equity, distributional equity is a nonactuarial notion. It refers to the idea that, at least in some

circumstances, two individuals should be charged the same price in spite of the fact that they, or groups they are members of, face different risks. Bans on risk classification on the basis of genetic conditions such as Huntington's disease, or on the basis of preexisting conditions more generally, are primarily motivated by a concern for distributional equity.

Distributional equity also encompasses attempts to use policy for the explicit purpose of redistributing from a historically advantaged class (e.g., males) to a historically disadvantaged class (e.g., females). This is distinct from concerns for actuarial group equity – i.e., that one group should not subsidize another in an actuarial sense. Because riskiness rather than group membership is the more fundamental characteristic vis-a-vis insurance provision, it is not obvious why providing distinct groups of heterogeneous risks actuarially equally would be a desirable policy goal. This article therefore focuses primarily on horizontal, financial, and distributional equity.

Efficiency

There are at least two distinct notions of efficiency that are relevant in the risk-classification context: interim efficiency and ex-ante efficiency. There are two distinct types of interim efficiency: the efficiency of outcomes and the efficiency of institutions.

Insurance outcomes in market A are said to be more interim efficient (or interim Pareto efficient) than insurance outcomes in market B when every individual is at least as happy with the insurance policy they would get in market A as with the insurance policy they would get in market B, and someone is strictly happier. Equivalently, market B's outcomes are interim inefficient if nobody would object to replacing the market with market A's outcomes, and at least somebody would strictly prefer the switch.

The notion of interim efficiency of institutions is applied when there is a range of possible insurance outcomes consistent with different policy institutions. The range of possible outcomes consistent with a policy regime in which risk classification is legal, for example, may depend on the extent to which the government also imposes taxes on the contracts sold to different risk classes. Similarly, there may be a range of insurance outcomes consistent with a regime in which risk classification is banned. Saying that the institution of legal risk classification is interim efficient means that for every potential banned classification outcome, there is some legal classification outcome that makes every individual at least as happy (and some strictly happier).

Interim efficiency involves evaluating insurance markets from the point of view of individuals who know their type (which could include intrinsic riskiness or tastes) and risk class. Because there are typically many types and classes within a given population, there will typically be many different possible insurance outcomes which cannot be compared on interim efficiency grounds, as some individuals would be better-off with one of these outcomes, and other individuals would be better-off with another.

In contrast, ex-ante efficiency evaluates efficiency from the point of view of a representative individual behind a veil of ignorance about their risk type or class. Insurance outcomes in market A are thus said to be more ex-ante efficient than insurance outcomes in market B if a hypothetical individual who did not yet know what risk type or class they will belong to would prefer the market A outcomes.

The notions of distributional equity and ex-ante efficiency are closely related. One might reasonably use the notion of distributional equity to argue that individuals with the gene for Huntington's disease should be able to purchase insurance covering the costs associated with its treatment for the same premium as someone without the gene. One basic argument is that it would be unfair to charge an individual for something entirely out of their control. Alternatively, one could make the same arguments on the grounds of ex-ante efficiency: A risk-averse representative individual behind the veil of ignorance, who did not yet know whether or not they would be born with the gene, would strictly prefer to be born into a world in which premiums do not depend on the presence of the gene.

Distributional equity can potentially be invoked for other unrelated reasons, but this article focuses on the particular distributional equity concerns arising from the point of view of a representative, risk-averse individual behind a veil of ignorance about their type and class. In other words, it regards as beneficial policies which redistribute toward risk types that are relatively disadvantaged from an ex-ante point of view.

Another way of framing the desire for, for example, gene-independent pricing is in terms of a desire for insurance against classification risk. Because individuals are either born with the Huntington's disease or not, individuals cannot directly insure themselves against the risk of having the gene and being in a bad risk class. Preventing insurers from genetic discrimination is potentially desirable insofar as it effectively provides otherwise unavailable insurance against this classification risk.

A similar argument potentially applies more generally to bans on risk classification on the basis of preexisting conditions like cancer or diabetes. The primary conceptual distinction is that one could, in principle, have insured oneself against such classification risk by purchasing a long-term, or guaranteed renewable contract before the condition developed. Insofar as the market for long-term contracts or other forms of insurance against classification risk functions poorly, however, restricting risk classification has potentially beneficial distributional equity effects insofar as it reduces classification risk.

It is important to note that interim inefficiency of outcomes implies ex-ante inefficiency as well. If outcomes in market A are better than market B outcomes at the interim stage for each type, then the representative agent behind the veil of ignorance will necessarily prefer market A. Interim inefficiency in the institutional sense implies ex-ante inefficiency in a somewhat more subtle sense. As a stand-alone policy, for example, a ban on risk classification might reduce interim efficiency (in the institutional sense) yet raise ex-ante efficiency through beneficial distributional equity effects. Nevertheless, interim inefficiency in the institutional sense implies the existence of some alternative intervention, such as government-coordinated risk adjustments, that is even better than a ban from an ex-ante perspective.

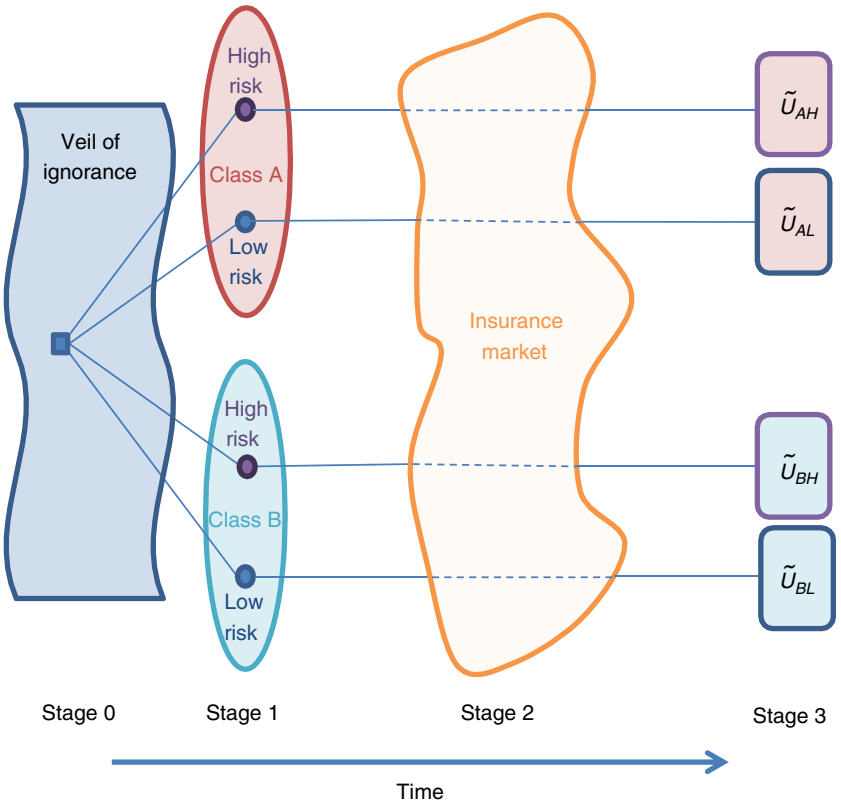


Figure 1 A conceptual framework.

The Equity and Efficiency Trade-offs of Risk Classification

Risk classification will typically have implications for both efficiency and equity. The particular trade-offs between efficiency and equity implied by the decision to allow or ban risk classification in insurance markets are context-dependent. Figure 1 provides a simple framework for sorting these contexts.

Figure 1 depicts the timing of an abstract insurance market. At stage 0, a notional representative individual contemplates the future from behind a veil of ignorance. At stage 1, individuals are born and learn both their true risk type and their class. The diagram depicts a case with two risk types, high and low risks, and two classes, labeled A and B, which might represent male and female, white and nonwhite, or Huntington's positive and negative, for example.

At stage 2, individuals enter an insurance market and potentially purchase insurance. At stage 3, the health outcomes are realized, and individuals with insurance receive their coverage and choose treatment levels. These outcomes result in a random utility, or well-being, level denoted by $\tilde{U}_{ij}$ that will potentially depend on risk type j and class i.

This framework can be used to explore the consequences of risk classification in a variety of situations. For example, if classes A and B are 30-year-old females and males, respectively, then it encompasses the example above illustrating the trade-off between individual and group equity when there is a lower but nonzero fraction of high-risk types within class A and a lower but nonzero fraction of low-risk types within class B. To capture a situation like classification based on the presence of the Huntington's gene in which class is perfectly predictive of risk type, one would simply take the fractions of the high-risk types within classes A and B to be one and zero, respectively.

The framework is best suited for analyzing insurance decisions that take place at one moment in time. It is less well suited to addressing fundamentally dynamic issues, such as the implications of risk classification based on preexisting conditions like diabetes or heart disease. It can be adapted to this application if insurance is sold on an annual basis with no long-term contracting, however. Similarly, it can be fruitfully applied to analyze the implications of age-based risk classification in some contexts.

When this simple framework is applicable, the qualitative implications of risk classification hinge on three basic questions.

Question 1: Is Risk Classification Perfect or Imperfect?

In some cases, as with the BRCA1/2 breast cancer gene, class is only imperfectly correlated with risk: There may be women without the gene who still face a high risk of breast cancer. In other cases, such as the gene for Huntington's disease, class is closer to perfectly predictive of risk type, and all members of either class will have the same risk type. It may then be said that risk classification is perfect.

Whether or not risk classification is perfect is important for two reasons. First, when risk classification is perfect, classes are pools of individuals who are perfectly homogenous from the point of view of health risk. The tension between horizontal and group equity therefore disappears. Second, when risk classification is imperfect, insurers who employ risk classification still face heterogeneity of risks within each class. Employing risk classification therefore reduces, but does not eliminate informational asymmetries. This is important because, in the face of informational asymmetries, insurers may find it useful to employ indirect mechanisms to induce self-sorting of different risks. This is known as screening. Screening can have important implications for efficiency and equity.

Question 2: Are Policies Uniform or Not?

Screening refers to the deliberate attempt to induce self-sorting of individuals through contract design. In the canonical example of screening, insurers offer two types of policies: An expensive comprehensive policy and a less expensive and less comprehensive policy, such as a catastrophic coverage policy with a very high deductible. Individuals who know themselves to be in good health are more likely to find the latter an appealing option, so individuals will be induced to self-sort by riskiness into distinct policies.

Screening relies on insurers' ability to tailor menus of significantly different policy options: it is predicated on nonuniform policies. If regulatory restrictions circumscribe insurers' ability to design such menus, for example, through coverage mandates that require all insurance policies to cover a certain same set of conditions, then screening will be curtailed or eliminated.

To see why the uniformity or nonuniformity of policies can have important implications for the equity and efficiency effects of banning risk classification, consider the effects of banning gender-based risk classification. If insurers find it much more costly to provide health care to women than to men then, absent any coverage mandates, it could potentially circumvent the ban by offering two policies: An expensive and comprehensive policy, and a less expensive one providing comprehensive coverage for everything except childbirth, breast cancer, gynecological examinations, and other gender-specific health care needs. Women faced with such a menu would find it worthwhile to pay the higher premium for coverage of their needs, and men would not. In this case, the insurer would effectively circumvent the risk classification ban, which would consequently have neither efficiency nor equity effects. In contrast, a ban imposed under coverage mandate-induced policy uniformity would likely have welfare effects.

Question 3: Are Insurance Purchases Mandated or Not?

In markets without purchase mandates, individuals who perceive themselves to have the greatest need for insurance, and hence the highest expected costs to insurers, will be differentially more likely to purchase coverage, whereas lower-risk individuals are differentially likely to opt out of buying coverage at all. In this case, the pool of insured individuals is said to be adversely selected relative to the population. An adversely selected risk pool requires higher premiums for firms to break even. In the most severe cases, adverse selection can completely destroy a market via an adverse selection death spiral.

Because purchase mandates and risk classification are two different ways to mitigate adverse selection, the presence or absence of a purchase mandate is crucial for analyzing the equity and efficiency implications of risk classification.

A Quick-and-Dirty Guide to the Equity–Efficiency Trade-offs

The eight distinct answers to the set of three questions above describe eight conceptually distinct institutional contexts. In practice, however, purchase mandates are typically coupled with minimum coverage mandates that limit the degree of policy differentiation: otherwise, individuals could fulfill the mandate by purchasing a low-priced contract providing essentially zero coverage. Therefore the two regimes with mandated purchases and differentiated products are not considered.

Table 1 provides a quick reference guide to the efficiency and equity effects of risk classification in the six remaining institutional contexts. It focuses on interim efficiency and horizontal, financial, and distributional equity, with beneficial distributional equity effects interpreted as those which would be desirable from the point of view of the veiled representative individual at stage 0 in Figure 1.

Consider first environments with a purchase mandate and a uniform contract. Bans on risk classification are the least likely to be controversial in these environments, because they

Table 1 Effect of a ban on risk classification

Institutional context	*Interim efficiency*	*Distributional equity*	*Horizontal equity*	*Financial equity*
Effects of a ban on perfect risk classification				
Mandated purchase and uniform contract	Neutral	Beneficial	Neutral	Detrimental
Optional purchase and uniform contract	Detrimental	Beneficial/neutral	Neutral	Detrimental
Optional purchase and differentiated contract	Detrimental	Beneficial/neutral	Neutral	Detrimental
Effects of a ban on imperfect risk classification				
Mandated purchase and uniform contract	Neutral	Beneficial	Beneficial	Detrimental
Optional purchase and uniform contract	Detrimental (institutionally)	Beneficial/neutral	Beneficial	Detrimental
Optional purchase and differentiated contract	Detrimental (institutionally)	Beneficial/neutral	Beneficial	Detrimental/neutral

improve distributional equity in a horizontally equitable way without harming interim efficiency. This is because only premiums, not insurance coverage, are affected by risk classification, and banning risk classification beneficially (from an ex-ante perspective) redistributes from individuals who were born into the fortunate group with fewer low risks to those who were unlucky enough to be born into the higher-risk group.

When risk classification is imperfect, but purchases of a uniform contract are still mandatory, such a ban also has a beneficial impact on horizontal equity, because it prevents individuals with the same true risk from being charged different premiums by virtue of the group to which they happen to belong. The primary objections to banning risk classification in mandatory-purchase uniform-contract contexts are likely to stem from financial equity effects; some might feel strongly that individuals should not be forced to cross-subsidize others. This objection is likely to be particularly germane for risk classification based on preexisting and preventable conditions, but it may be present more broadly.

In an optional-purchase institutional context, banning risk classification may additionally induce efficiency reducing adverse selection effects. Suppose, for example, that group A consists of people with an expensive-to-treat preexisting condition and group B consists of healthy individuals. With legal risk classification, the market will segment and group B individuals will pay lower premiums. One possibility, if risk classification is banned, is that all individuals will continue to purchase insurance at some intermediate premium. In this case, the welfare effects are exactly as with a mandate. The other possibility is that the market will suffer from a death spiral: Group B individuals will find insurance too expensive at the new premium and will leave the market. Premiums for the group A individuals will then rise to the same as they were before the ban was imposed. In this case, the risk classification ban will be purely efficiency reducing. It will have no beneficial equity effects at all.

Similar adverse selection-driven negative efficiency effects arise with a richer and more realistic set of individual risk types. If the adverse selection is mild, so that only a few of the lowest-risk types are driven from the market by a ban in risk classification, then policy makers will face a genuine trade-off between beneficial distributional equity effects of uniform pricing and the efficiency costs of adverse selection. With sufficiently severe adverse selection, the uniform pricing will have at most mild distributional equity benefits, and ex-ante efficiency will be reduced. If a policy maker wanted to ban risk classification and believed that the adverse selection problem was likely to be severe, introducing a purchase mandate would therefore be essential. This was the primary motivation for policy makers to include a coverage mandate in the Patient Protection and Affordable Care Act passed by the US Congress and signed into law by President Barack Obama in 2010.

A similar trade-off between interim efficiency and distributional equity applies if risk classification is imperfect, but the interim efficiency effects of a ban are detrimental in the institutional sense rather than in the outcome-based sense. In particular, one can show that the outcome with banned risk classification can always be Pareto improved on with legal risk classification and some appropriate risk adjustments, for example, through government-administered transfers across insurers serving different risk classes.

The detrimental interim efficiency effects of banning perfect risk classification are also similar when contracts are differentiated and insurance is not mandated (bottom rows). The mechanism is somewhat different, however: Low-risk individuals will be screened – induced to self-select – into a high-deductible policy providing worse coverage rather than being adversely selected out of the market entirely. There is some disagreement among economists about the precise nature of screening in insurance markets; some widely used models of insurance markets predict the same outcomes with and without risk classification bans, and, consequently, no distributional equity effects. Others predict the potential for beneficial distributional equity effects via pooling of different risk types or cross-subsidies across distinct contracts.

It is clear that even in the simple analytical framework depicted in Figure 1, evaluating the welfare consequences of risk classification is nontrivial and highly context-dependent. Of the three central questions identified above as being useful for understanding these welfare effects, the latter two are observable policy questions. The first question is an empirical one. It can be understood as a question about the presence or absence of asymmetric information: Risk classification is imperfect precisely when there is unused information about risk within a risk class. In these cases, incentive contracting – designing contracts to mitigate the imperfections of the risk classification technology – can play an important role. Screening is the type of incentive contracting that is particularly important when the relevant asymmetric information within a risk class is of the adverse selection type, as has largely been assumed up to this point, but other types are potentially important with other types of informational asymmetries. In part because of its central importance for the welfare analysis of risk classification, the presence or absence of informational asymmetries has been the subject of much recent empirical work.

Risk Classification and Residual Asymmetric Information in Health Insurance Markets

Perfect risk classification should separate individual risks and generate different actuarial insurance premiums that reflect these risks. With actuarial premiums, full insurance should be the optimal contract, and there should not be any correlation between insurance coverage and individual risk. But in the real life of health insurance contracting, there are numerous reasons for imperfections in risk classification. Particularly important among these is the possibility of residual asymmetric information within a given risk class. Recent empirical work has therefore focused on searching for evidence for presence and extent of this sort of residual asymmetric information.

General Tests for Residual Asymmetric Information

Information problems are common in insurance markets. Usually, the insured are better informed about their own

characteristics or actions than are their insurers. The two best-known information problems discussed in the economics literature are adverse selection, discussed above, and moral hazard, where insurance leads individuals to take unobserved actions either before (ex-ante moral hazard) or after (ex-post moral hazard) the realization of health outcomes that raise the costs borne by the insurer. Asymmetric learning over time is a third information problem. Because similar empirical patterns are predicted by these three problems, empirical work on information problems is challenging.

Empirical work has three sequential goals. The first is to determine whether information problems exist, and, if so, how severe they are. The second is to identify which information problem or problems are present when the first test rejects the null hypothesis that there is no information problem. This is important for an insurer because it must implement the appropriate instruments to improve resource allocation. A fixed deductible, for example, efficiently reduces ex-ante moral hazard, but not necessarily ex-post moral hazard. A high deductible can even have an adverse effect and encourage accident cost building.

The third goal is to find ways to improve the contracts and reduce the negative impact of asymmetric information on resource allocation. These resource allocation objectives must take into account other issues, such as risk aversion, equity, and accessibility of services. This last issue is particularly important in health care markets. A decrease in insurance coverage may reduce ex-ante moral hazard because it exposes the insured person to risk, but it also significantly reduces accessibility to health services for sick people who are not responsible for their condition. Although the third goal is ultimately the most important, this article focuses on the first two goals, which have been convincingly tackled in the literature only recently.

Well-constructed theoretical models with carefully established theoretical predictions are essential for achieving these goals. Many theoretical contributions were published in the 1970s which appealed to asymmetric information to account for stylized facts observed in insurance markets. Not all of these accounts were readily adapted to formal tests of information asymmetries, however. For example, partial insurance, such as deductible and co-insurance contracts, can be justified by either moral hazard or adverse selection, but proportional administrative costs can also justify it. So the mere fact that these are common features of real-world insurance policies does not imply the presence of asymmetric information.

The following simple question has motivated most recent empirical work toward the first goal: Do insurers that apply risk classification techniques based on observable characteristics in their underwriting policies also find it useful to employ contract design to further separate risk types within the risk class? In static or one-period contracts, the answer is no unless there is residual asymmetric information within the risk classes. (The reality is, of course, much more complicated because contract duration between the parties can cover many periods, and over time the true risks may become known to both parties.) Finding a residual correlation between chosen insurance coverage and risk within risk classes is therefore a tell-tale sign of asymmetric information. Tests for such a correlation have been the centerpiece of the empirical literature on information problems in insurance markets.

Econometricians analyze two types of information when studying insurers' data. The first type contains variables that are observable by both parties to the insurance contract. Risk classification variables are one example. Econometricians/insurers combine these variables to create risk classes when estimating accident distributions. They can be used to make estimates conditional on the risk classes or inside the risk classes. The second type is related to what is not observed by the insurer (or the econometrician) during contract negotiations or selections, but can explain the insured's choice of contracts or actions. A typical empirical study looks for the conditional residual presence of asymmetric information in an insurer's portfolio by testing for a correlation between the contract coverage and the realization of the risk variable during a contract period. Different parametric and nonparametric tests have been proposed in the literature.

Finding a positive correlation between insurance coverage and risk is a necessary condition for the presence of asymmetric residual information, but it does not shed light on the nature of the information problem. In insurance markets, the distinction between moral hazard and adverse selection boils down to a question of causality. Under moral hazard, the structure of an insurance contract drives the unobserved actions, and hence the riskiness, of the insured. For example, a generous health insurance plan can reduce the incentives for prevention and increase the risk of becoming sick. Under adverse selection, the predetermined riskiness of an individual drives their contract choices: Higher-risk individuals will tend to choose policies providing better coverage. The correlation between insurance coverage and the level of risk is positive in both cases, but the directions of causality in the two cases are exactly opposite.

To separate moral hazard from adverse selection, econometricians need a supplementary step. In insurance markets, dynamic data are often available. Time adds an additional degree of freedom to test for asymmetric information, particularly in the presence of experience rating – whereby future premiums depend on past accident history. Experience rating works at two levels in insurance. Past accidents implicitly reflect unobservable characteristics of the insured (adverse selection) and introduce additional incentives for prevention (moral hazard). Experience rating can therefore directly mitigate problems of adverse selection and moral hazard, which often hinder risk allocation in the insurance market.

The failure to detect residual asymmetric information, and more specifically, moral hazard and adverse selection in insurance data, is often due to the failure of previous econometric approaches to model the dynamic relationship between contract choices and claims adequately and simultaneously when looking at experience rating. Intuitively, because there are at least two potential information problems in the data, an additional relationship to the correlation between risk and insurance coverage is necessary to test for the causality between risk and insurance coverage.

Testing for Asymmetric Information in the Health Insurance Market

Many reviews of empirical studies in different insurance markets have been published, including health insurance and

long-term care insurance. It is observed that the coverage–risk correlation is particular to each market. Accordingly, the presence of a significant coverage–risk correlation has different meanings in different markets, and even in different risk pools in a given market, depending on the type of the insured service, the participants' characteristics, institutional factors, and regulation. This means that when testing for the presence of residual asymmetric information, one must control for these factors as well. Up to now the empirical coverage–risk correlation findings have been equivocal. What characteristics and factors explain the absence of robust coverage–risk correlations in health insurance markets?

Long-term care market and the health care market are analyzed separately, notably because long-term care insurance effectively combines both health insurance and longevity insurance (annuities). It is well documented that private long-term care insurance is very expensive in the US and therefore not very popular. Less than 5% of the elderly participate in this market. Is it due to adverse selection? Those who purchase this coverage do not seem to represent higher risks than the average population. This negative result is explained by a combination of two opposite effects: A pure risk effect and a risk aversion effect. For a given risk aversion, higher-risk individuals buy more insurance under asymmetric information, as do more risk-averse individuals (who are assumed to engage in more prevention to reduce their risk). The net effect on the correlation between risk and coverage is not significant because both high-risk and low-risk individuals buy this insurance. However, it is not evident that more risk-averse individuals put forth more effort. Consequently, the absence of correlation may be explained by factors other than risk aversion.

Many empirical studies in the literature find a positive correlation between poor health condition and generous coverage, whereas other studies do not find this correlation. Some do not reject asymmetric information in the medical insurance market, but do not find evidence of adverse selection. Their results are even consistent with multidimensional private information along with advantageous selection. Indeed, some obtain a negative correlation between risk and insurance coverage. The significant sources of advantageous selection are income, education, longevity expectations, financial planning horizons, and most importantly, cognitive ability.

Other studies offer detailed analyses of health insurance plans. For example, it is shown that when the employer increased the average participation cost of the most generous plan for the policyholders, regardless of the risk they represented, the best risks in the pool with lower medical expenses left this plan for a less generous one with a lower premium. The new insurance pricing clearly generated adverse selection. Even if the age of the insured were observable, the insurance provider did not use this information, and the younger participants abandoned the more generous plan. This is a case in which the absence of a proper risk classification yielded severe adverse selection. This type of constraint, wherein risk classification variables are not used, is often observed in the health care market where the trade-off between efficiency and distributional equity matters.

One potential reason for not observing a significant correlation between coverage and risk is the absence of insured private information on the insured's health status. Young individuals who may not have experienced any health problems may think they belong to the low-risk group. The statistical test should be done within these risk classes, even if the employer does not use age as a risk classification variable. Another reason for the lack of risk–coverage correlation, which may also apply to health insurance, is policyholders' failure to use their private information when selecting insurance policies. It has been found, for example, that the demand for life insurance is not sensitive to insurance price and risk.

It has also been documented that insurance consumption depends on institutions. Moreover, risk classification in the health care market is heavily regulated in many countries. Therefore, the empirical predictions based on the implicit assumption of competitive markets may not be appropriate for many markets, including health insurance. For further discussion on particularities other than efficient risk classification that may generate an absence of correlation between insurance coverage and risk, see the references in further reading.

Conclusion

This article has discussed the complex welfare effects of risk classification in health insurance. The policy decision to permit or ban risk classification may have consequences for efficiency, for equity, or for both. The various relevant notions of efficiency and equity appropriate in health insurance context were reviewed and trade-offs that are likely to arise in various institutional contexts were qualitatively characterized. A key question for this characterization is whether or not there is (or would be) within-class residual asymmetric information when insurers employ risk classification based on observable characteristics. The extensive and growing empirical works on this question were discussed. There remains substantial scope for future empirical work directed toward quantifying the equity–efficiency trade-offs of risk classification.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Long-Term Care Insurance. Moral Hazard. Personalized Medicine: Pricing and Reimbursement Policies as a Potential Barrier to Development and Adoption, Economics of. Private Insurance System Concerns. Risk Selection and Risk Adjustment. Social Health Insurance – Theory and Evidence

Further Reading

Buchmueller, T. and DiNardo, J. (2002). Did community rating induce an adverse selection death spiral? Evidence from New York, Pennsylvania, and Connecticut. *American Economic Review* **92**, 280–294.

Chiappori, P. A. and Salanié, B. (2013). Asymmetric information in insurance markets: Predictions and tests. In Dionne, G. (ed.) *Handbook of insurance*, 2nd ed. Boston: Springer.

Cohen, A. and Siegelman, P. (2010). Testing for adverse selection in insurance markets. *Journal of Risk and Insurance* **77**, 39–84.

Crocker, K. J. and Snow, A. (1986). The efficiency effects of categorical discrimination in the insurance industry. *Journal of Political Economy* **94**, 321–344.

Crocker, K. J. and Snow, A. (2013). The theory of risk classification. In Dionne (ed.) *Handbook of insurance*, 2nd ed. Boston: Springer.

Cutler, D. M. and Zeckhauser, R. J. (2000). The anatomy of health insurance. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, 1, pp. 563–643. Amsterdam: Elsevier Science.

Dionne, G., Fombaron, N. and Doherty, N. A. (2013). Adverse selection in insurance contracting. In Dionne, G. (ed.) *Handbook of insurance*, 2nd ed. Boston: Springer.

Dionne, G. and Rothschild C. G. (2012). Risk classification in insurance contracting, Working paper, Canada Research in Risk Management, HEC-Montréal, 50 pages, SSRN 1958176.

Harrington, S. (2010). U.S. health-care reform: The patient protection and affordable care act. *Journal of Risk and Insurance* **77**(3), 703–708.

Hoy, M. and Polborn, M. (2000). The value of genetic information in the life insurance market. *Journal of Public Economics* **78**, 235–252.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**, 629–649.

Thiery, Y. and Van Shoubroeck, C. (2006). Fairness and equality in insurance classification. *Geneva Papers on Risk and Insurance: Issues and Practice* **31**, 190–211.

Risk Equalization and Risk Adjustment, the European Perspective

WPMM van de Ven, Erasmus University Rotterdam, Rotterdam, The Netherlands

Glossary

Equalization payment The compensation per insured that an insurer receives from or has to pay to the equalization fund. For example, the equalization payment could be equal to the insured's risk-adjusted predicted expenses minus the average expenses per person. If the equalization payment is negative, the insurer has to pay it to the equalization fund.

Out-of-pocket premium Premium minus equalization payment.

Premium The price of insurance. In a competitive market the premiums are risk-rated. An insured's premium then equals the risk-adjusted expected expenses of the insured, plus a loading fee (for administrative costs, the insurer's profit, marketing costs, etc.).

Risk adjustment A technique for adjusting a payment to someone's risk characteristics.

Risk equalization A technique for compensating insurers by means of risk-adjusted equalization payments for the composition of their risk portfolio.

Risk selection Actions (not including risk rating) by consumers and insurers to exploit unpriced risk heterogeneity and break pooling arrangements.

Introduction

Since the 1990s an increasing number of European countries permit periodic consumer choice of insurer in their social health insurance schemes (e.g., Belgium, the Czech Republic, Germany, Israel, the Netherlands, Russia, Slovakia, Switzerland, and Russia). It can be hypothesized that such consumer choice provides the insurers with effective incentives for efficiency and innovation. However, an unregulated competitive health insurance market also tends toward risk-adjusted premiums and rejection by insurers of high-risk applicants. Therefore, governments in these countries interfered with regulation to make health insurance accessible and affordable for everyone. Risk-adjusted equalization payments are a necessary component of any efficient intervention. Other potential interventions have unfavorable effects: premium regulation creates incentives for selection (which may have several unfavorable effects) and ex-post compensations to the insurers reduce their incentives for efficiency. However, because risk equalization is technically complex, policymakers are confronted with a complicated trade-off between affordability, selection and efficiency.

Risk equalization is discussed from a European perspective. First its relevance is discussed in the competitive health insurance markets and some technical complications (Section 'Why Risk Equalization, and How?'). Then a historical perspective of the European experience with risk equalization (Section 'The European Experience with Risk Equalization: A Historical Perspective') and future perspectives including its relevance for provider payments and for countries with a National Health Service (NHS) such as England is provided. Finally the conclusions are summarized in the Section 'Conclusion.'

Why Risk Equalization, and How?

The solidarity principle, which is highly valued in Europe, implies that high-risk and low-income individuals receive a subsidy to make health insurance affordable. Therefore a great challenge for policymakers is: how to combine solidarity with consumer choice of health insurer? In an unregulated competitive insurance market insurers have to break even, in expectation, on each contract, because competition minimizes the predictable profits per contract. Insurers can do so by (1) adjusting the premium to the consumer's risk (premium differentiation), (2) adjusting the product, for example, coverage and benefits designed to attract different risk groups per product and charge premiums accordingly (product differentiation), or, if the transaction costs of further premium and product differentiation are too high, (3) by adjusting the accepted risk to the premium of a given product (risk selection), for example, by excluding certain preexisting medical conditions from coverage or by not accepting high-risk people. Given the average expenses per risk group, unregulated competition could result in premiums that can differ a factor of 500 or more once health status and age are taken into account.

Although premium differentiation makes coverage less affordable for the high risks, risk selection (by excluding certain preexisting medical conditions from coverage or by not accepting high-risk people) makes coverage less available to the high risks. In both ways, guaranteed access to affordable coverage for the high risks is jeopardized.

To simplify the analysis, it is assumed that health insurers are bound to an open enrollment requirement. This implies that insurers must accept each applicant for a standard coverage. In practice, open enrollment is required in all countries with a competitive social health insurance market. As long as insurers are free in setting premiums, this assumption is nonrestrictive, because insurers are allowed to risk-adjust the premium for each applicant and can offer each type of policy in addition to the policy with the standardized coverage. By this assumption the problem of unavailability that would occur in case of rejection or coverage restrictions, is essentially transformed into a problem of unaffordability (high premiums for high-risk individuals) to be solved by cross subsidies. In this article the author focuses only on the so-called risk-solidarity, i.e., cross subsidies from low-risk to high-risk individuals (and not on income solidarity).

An effective way to achieve risk-solidarity without disturbing competition among the insurers is to give the high-risk consumers a subsidy out of a solidarity fund that is filled with mandatory solidarity contributions from the low risks. Ideally these subsidies are risk-adjusted, i.e., the subsidy is adjusted for the risk factors that the insurers use. For practical reasons the subsidy can be given directly to the insurers. In a transparent competitive market, insurers are forced to reduce each consumer's premium with the per capita subsidy they receive for this consumer. By giving risk-adjusted subsidies to the insurers the different risks that consumers represent for them are equalized. Therefore this way of organizing risk-adjusted subsidies is referred to as 'risk equalization.' In practice, all European countries that apply risk-adjusted subsidies do this in the form of risk equalization (see Section 'The European Experience with Risk Equalization: A Historical Perspective').

Sometimes the term risk adjustment is used rather than risk equalization. However, risk adjustment can also be applied to, for example, provider payments. The term risk equalization is used to denote the specific case of 'risk-adjusted compensations to (the consumers via) the insurers.'

Complementary Strategies

Although sufficiently risk-adjusted equalization payments can be an effective strategy to guarantee affordable coverage in a competitive individual health insurance market, in practice the risk equalization payments are still insufficiently risk-adjusted (see below). Therefore, in addition government may implement one or more of the following strategies:

1. A system of ex-post cost-based compensations to the insurers. For example, the insurers are fully or partly compensated by government for an individual's expenses in excess of a certain annual threshold. These compensations will be reflected in premium reductions, in particular for the high risk-risk enrollees.
2. Premium rate restrictions. An extreme form of premium rate restrictions is that the premiums must be community rated, i.e., insurers must charge the same out-of-pocket premium for the same product to each enrollee, independent of the enrollee's risk. All European countries with a competitive social health insurance market require the out-of-pocket premiums to be community rated.

However, each of these additional strategies has substantial drawbacks, resulting in serious trade-offs.

Ex-post cost-based compensations are not optimal because they reduce the insurers' incentive for efficiency resulting in an affordability-efficiency trade-off.

Premium rate restrictions have some major drawbacks as well. Although the goal is to create implicit cross subsidies from the low risks to the high risks who are in the same pool, this pooling creates predictable profits and losses for identifiable subgroups in the pool, and thereby provides insurers with incentives for risk selection, which may threaten affordability, efficiency, quality of care, and consumer satisfaction (see Box 1). Therefore, premium rate restrictions confront policymakers with an affordability-selection trade-off. In addition it is questionable to what extent premium rate restrictions are effective in the long term, because product differentiation may result in indirect premium differentiation. Insurers may offer special products for various risk groups, for example, depending on life-stage, lifestyle, or health status. Such risk segmentation across the product spectrum can be observed in, for example, Australia, Ireland, and South Africa, where premiums must be community rated. In this way 'community rating per product' results in low premiums for low risks and high premiums for high risks, which undermines the goal of 'community rating across the market'.

Product differentiation may not only occur in voluntary health insurance markets, but also in mandatory social health insurance markets. For example, in the Netherlands a substantial variation in the health insurance products is allowed. These products may vary, for example, according to the list of contracted providers, the financial incentives to motivate consumers to use preferred providers, procedural conditions (e.g., yes or no preauthorization by the insurer or by the general practitioner) and the list of covered pharmaceuticals.

The relevance of good risk equalization is that if the equalization payments are sufficiently risk-adjusted (see below) there is no need for the other strategies, each of which confronts policymakers with severe trade-offs. The better the risk-adjusted equalization payments are adjusted for relevant risk factors, the less severe are these trade-offs.

Box 1 Unfavorable effects of risk selection

1. Insurers have a disincentive to respond to the preferences of high-risk individuals. For example, insurers with a good reputation for chronic care would attract many unprofitable patients and would be the victim of their own success. Therefore, insurers may structure their coverage to make the plan unattractive to high risks or choose not to contract with providers who have the best reputation for treating chronic illnesses. This in turn may discourage physicians and hospitals from acquiring such a reputation.
2. Efficient insurers who do not engage in risk selection may lose market share to inefficient risk-selecting insurers, resulting in welfare loss to society.
3. If risk selection generates large predictable profits it will be more profitable than improving efficiency in health care production. At least in the short run, if an insurer has limited resources available to invest in reducing costs, it may prefer to invest in risk selection rather than in improving efficiency.
4. To the extent that some insurers are more successful than others in attracting low risks, selection will result in risk segmentation. High risks will therefore pay higher premiums than low risks, undermining 'community rating across the market.'
5. Selection may induce instability in the insurance market because low risks have a permanent incentive to break the pooling of heterogeneous risks by switching to lower-priced insurers.
6. Selection wastes resources because investments purely aimed at attracting low risks through risk segmentation or selection produce no net benefits to society.

Acceptable Costs; S-Type and N-Type Risk Factors

For the calculation of the risk-adjusted equalization payments it is important to determine the costs and the risk factors on which the payments should be based. The costs of the services

and intensity of treatment that are acceptable to be compensated are denoted as the acceptable costs. For example, acceptable costs may be those generated in delivering a 'specified basic benefit package' containing only medically necessary and cost-effective care. Because the 'acceptable cost level' is hard to determine, in practice the equalization payments are mainly based on observed expenses rather than needs-based costs. However, observed expenses are determined by many factors, not all of which need to be used for calculating the equalization payments. Assume that all risks factors X that determine observed expenses can be divided into two subsets: those factors for which solidarity/subsidy is desired, the S-type factors; and those for which solidarity/subsidy is not desired, the N-type factors. Then the equalization payments should only be adjusted for the S-type risk factors and not for the N-type risk factors.

Decisions about which risk factors should be labeled an S-type or N-type factor, reflect value judgments that differ across countries and among individuals. In most societies health status and gender are likely to be S-type risk factors. Other risk factors may be open for discussion:

- Characteristics of the individual such as lifestyle, taste, income/wealth, religion, being self-employed, race, and ethnicity;
- Characteristics of the contracted providers, such as price level, practice style, utilization review, various health management strategies, and (in)efficiency;
- Characteristics of the region where the consumer is living, such as average price and income level, population density, average distance to hospitals, and whether there is an over- or undersupply of providers; and
- Characteristics of the contracts and financial incentives between plans and providers.

All European countries that apply risk equalization use age as a risk adjuster. This reflects the desired level of intergenerational solidarity and the desired way of paying health expenses over the life cycle in these countries. Nevertheless, age might (partly) be considered an N-type risk factor. Young people on average have relatively low health expenses and high expenses on housing, schooling and children, whereas for the elderly the opposite holds. The Affordable Care Act in the USA ('Obamacare') allows the insurers to differentiate their premium by a factor 1:3 for age for products sold via the so-called Exchanges, although the additional age-related variation in health expenses is compensated via risk equalization. In other words, age is partly an S-type risk factor and partly an N-type risk factor.

Region is often a disputable risk factor. It is likely that region captures differences in health status, which most likely are to be compensated. However, region also reflects differences in other risk factors, such as price level, oversupply, inefficiency, practice style, etc., which might not be compensated. The more health related risk factors are explicitly included in the equalization formula, the less will region reflect regional health differences and the more it will reflect regional differences in nonhealth factors.

If it is explicitly decided to adjust the equalization payments only for S-type risk factors and not for N-type risk factors, a logical consequence is that insurers are allowed to ask premiums from the consumers that are related to the N-type risk factors. If not, the insurers have incentives for selection based on these risk factors.

If the equalization payments are based on observed expenses, as is mostly the case in practice, the calculation of the risk-adjusted equalization payments could be as follows. Assume that $E(X)$ is the best estimate of the expected expenses in the next contract period for a person with risk characteristics X. An estimate of the acceptable cost level could then be $E(X)$ with the values of the N-type risk factors set at an acceptable level (e.g., the acceptable level of the price or supply of health care or the acceptable practice style). Because some components of the vector X have been fixed at some specific value, the acceptable cost level can be written as a function that only depends on the nonfixed values. Hence if $X = (Xs, Xn)$ and Xs has the S-type factors and Xn the N-type factors, then the acceptable cost level would be $A(Xs)$. The risk-adjusted equalization payment could then be a function of $A(Xs)$, for example, it could be $A(Xs)$ minus a fixed amount Y (or a certain percentage of $A(Xs)$, as in the USA Medicare). Negative equalization payments imply payments from the insurer to the subsidy fund. If it is assumed that the average premium (excluding surcharges for administration, selling costs, profits, etc.) equals the average predicted health expenses, the national average of the consumers' out-of-pocket premiums (i.e., premium minus equalization payment) equals Y. In countries such as Russia and Israel, $Y = 0$. In these countries the consumers do not pay any out-of-pocket premiums directly to their insurer. In countries such as Switzerland and Ireland, Y equals the average predicted per capita expenses. The Netherlands has an intermediate position, with Y equal to 45% of the average predicted per capita expenses.

Criteria for Risk Equalization

The application of risk equalization in practice is hindered because ideally the following criteria should be fulfilled:

1. Appropriateness of incentives: Insurers should have incentives for efficiency and health-improving activities, and no incentives for selection and for distorting information to be used for calculating the equalization payments.
2. Fairness: Ideally the risk-adjusted payments should only compensate for so-called acceptable costs, and depend only on so-called S-type risk factors. The payments should sufficiently compensate the insurers for their high-risk enrollees ('distributional fairness' and good predictive value), and should be sufficiently stable over time.
3. Feasibility: the required data should be routinely obtainable for all potential enrollees without undue expenditures or time. The data should be resistant to manipulation by the insurers and government should be able to control the correctness of the data. There should be no conflict with privacy and ideally the system should be acceptable to all parties involved. Information that is routinely collected, standardized, and comparable across different insurers and measures that are easily validated have greater feasibility than measures that require separate data collection, validation, and processing.

In practice most potential risk equalization models appear not to fully fulfill these criteria, resulting in complicated trade-offs.

Potential Risk Adjusters

Risk equalization research started in the late 1980s. The calculation of risk-adjusted equalization payments requires a good prediction of each individual's health expenses based on the individual's characteristics, which are called risk adjusters. Because it is clear that age and gender alone are insufficient adjusters, the research efforts in the past two decades focused on developing health adjusters.

The inappropriate incentives related to prior utilization as a risk adjuster ('rewarding high prior utilization') may be reduced by combining it with diagnostic information. Widely known classification systems are the ambulatory care group system and the diagnostic cost group (DCG) models. Diagnosis-based models begin by identifying a subset of all diagnoses that predict subsequent year resource use. The many codes are grouped into more aggregated groups based on clinical, cost, and incentive considerations. Diagnosis-based risk adjusters tend to do well in predictive accuracy and feasibility. Although these models outperform a model based on age and gender only, there still exist subgroups that are substantially undercompensated.

Health status information can also be derived from the prior use of prescription drugs. Lamers *et al.* (1999) developed the so-called pharmacy cost groups (PCGs). They classified drugs into different therapeutic classes and further classified them on the basis of empirically determined similarities in future costs. Although PCGs are good predictors of future health care costs, a point of attention is that if the additional subsidy for a PCG-classified enrollee (far) exceeds the costs of the prescribed drugs that form the basis for PCG-assignment, the insurer has an incentive to (stimulate the physician to) overprovide medication in order to ensure an increase in the future subsidies. To prevent perverse incentives, Lamers *et al.* (1999) used only 10% of all prescriptions to define the PCGs.

Disability and functional health status have been shown to be relatively good predictors of future expenditures, even after controlling for demographic factors and prior utilization. These indicators reflect someone's ability to perform various activities of daily living and the degree of infirmity, and seem to be an almost ideal adjuster.

There are different opinions about the usefulness of mortality as an additional risk adjuster (see e.g., Lubitz, 1987). Van Vliet and Lamers (1998) argued that mortality should not be used as a risk adjuster because most of the excess costs associated with the high costs of dying are unpredictable. Although cause-of-death information is theoretically attractive, practical concerns include reliability, validity, availability, manipulation, auditing, privacy of the data, and perverse incentives.

The only country with a competitive health insurance market that applies mortality as a risk adjuster is Belgium. In countries with a noncompetitive health insurance market it is not unusual to use mortality as a risk adjuster.

Do We Need Perfect Risk Equalization?

It is important to emphasize that in the case of premium rate restrictions the predictable profits and losses need not be reduced to zero. One should take into account an insurer's costs of selection and the (statistical) uncertainty about the net benefit of selection. A bad reputation resulting from selection activities such as keeping patients from the highest-quality care can be a high cost to an insurer. In addition, the information that is necessary for risk selection is not for free. So a 'perfect' risk equalization formula is not necessary. It should be refined to such an extent that insurers expect the costs of selection (including the cost of a bad reputation) to outweigh its benefits. By making the risk groups in the risk adjustment algorithm more homogeneous, the costs of selection increase although on average its profits fall. But it is still an unanswered question how much 'imperfection' is acceptable.

The European Experience with Risk Equalization: A Historical Perspective

The application of risk equalization in Europe started in the early 1990s, when several European countries started to radically reform their social health insurance system. In Belgium, Czech Republic, Germany, Israel, the Netherlands, Slovakia, and Switzerland the regulatory regime was changed such that the consumers have a guaranteed periodic choice among risk-bearing social health insurers, who are responsible for providing or purchasing health care for their enrolees. In some countries the social health insurers are called sickness funds (e.g., in Belgium, Germany, and Israel). In this article they will be indicated as '(health) insurers.'

Risk Adjusters and Ex-Post Cost-based Compensations

Before 2000 all these countries used predominantly demographic risk adjusters, in combination with community-rated out-of-pocket premiums. Some countries used disability and/or region as an additional risk adjuster(s). A disadvantage of predominantly demographic risk adjustment is that community-rated premiums create large predictable profits and losses for subgroups like the chronically sick, resulting in incentives for risk selection.

All European countries experience(d) severe implementation problems. Especially in the first years there was a serious lack of relevant data, in particular at the level of the individual enrollee. All in all the European experience indicates, in accordance with the experience in the USA that even the simplest risk equalization mechanisms are complex and that there are many start up 'surprise problems.'

Several countries used ex-post cost-based compensations as a complement to imperfect risk equalization, for example, Belgium and the Netherlands. In Israel the insurers receive a fixed payment for each person who is diagnosed with one of the following 'severe diseases': end stage renal failure requiring dialysis, Gaucher's disease, talasemia, hemophilia and acquired immune deficiency syndrome. Germany (until 2002) and Switzerland do not have any form of ex-post cost-based

compensations. Consequently, the incentives for selection in Germany and Switzerland are high.

Risk Selection

It is hard to give clear evidence of selection activities in practice. Even if insurers perform risk-segmenting activities, they may argue that it is not 'selection,' but normal commercial behavior because they are specialized in certain segments of the market. In addition, selection activities in the form of 'not investing in better care for unprofitable subgroups' are difficult to detect because it is not known what would have happened if insurers had put in place appropriate incentives. Therefore, rather than hard evidence of selection, the following anecdotal evidences of selection activities reported in the European countries are cited:

- Selective advertising/using the internet;
- Accessibility problems;
- Health questionnaires;
- Delayed reimbursements;
- Offering health insurance via life insurers who make specific selections based on health inquiries;
- Selectively terminating business in unprofitable regions, for example, by closing offices in high-cost areas;
- Opening clinics in healthy regions;
- Employer-related (group) health insurer;
- Via limited provider plans such as health maintenance organizations and preferred provider organizations;
- Offering high rebates in case of a deductible;
- Information to unprofitable enrollees that they have the right to change insurer;
- Turning away applicants on the telephone and ignoring inquiries and phone calls;
- Special bonuses for agents who are successful in getting rid of the most expensive cases by shunting them off to competitors; and
- Voluntary supplementary insurance.

Acceptable Costs

Belgium is the only country in which the distinction between S-type and N-type risk factors is a relevant policy issue in practice. It was decided that medical supply should not be included in the risk equalization system. Schokkaert and Van de Voorde (2003) illustrate the nontrivial impact of this political decision on the health insurers' results.

Although the Dutch government formally announced that the risk-equalization formula should only be based on age/gender/health, in practice the Dutch risk equalization model also contains risk factors such as region and being self-employed. These risk factors partly reflect health status (an S-type factor) and partly other risk factors (N-type factors). A correction for the biased weights in the current risk-equalization formula would have substantial financial consequences for the Dutch insurers.

Improvements of Risk Equalization

An effective way to prevent risk selection is to complement the demographic risk adjusters with health indicators. Then, (1) it is harder for the insurers to define who the preferred risks are; (2) on average the predictable profits/losses are less; and (3) there are less possibilities for insurers to select the preferred risks, than in the case that risk equalization is only based on age/gender. Since 2000 the risk equalization systems in several European countries have been improved by adding relevant health-based risk adjusters.

In the Netherlands the risk equalization model was extended with PCGs in 2002, with DCGs in 2004 and with an indicator of multi prior year high expenses in 2012. In Germany the incentives for risk selection were reduced by the implementation of an ex-post risk pooling for high costs insured in 2002, and by implementing a health adjuster in 2003 (yes/no being registered in an accredited disease management program). In 2009 a health-based adjuster was added that compensates for 80 severe diseases and costly and chronic diseases, based on diagnostic information and/or prescriptions. In Belgium risk adjusters based on inpatient diagnostic information and information about chronic conditions based on outpatient prescribed and reimbursed drugs were added in 2008. In Switzerland prior hospitalization has been included as a risk adjuster in 2012.

However, all these improvements in risk equalization are not necessarily a sufficient guarantee that selection will be reduced. Several arguments explain why selection may not be a major issue in the early stage of the implementation of a risk equalization mechanism in a competitive health insurance market, and why over time selection may increasingly become a problem. First, in the early stage many players, for example, consumers, health insurers, managers and providers of care, may be unfamiliar with the rules of the game. However, over time they will be better informed and can be expected to react to incentives for risk selection. Second, in the early stage the differences among health insurers with respect to benefits package, premiums, and contracted providers are relatively small. Over time they may increase. Third, most risk equalization systems have been implemented in the mandatory social health insurance system. Traditionally these health insurers are driven by social motives rather than by financial incentives. However, over time new insurers and increasing competition can make the market more incentive driven. As soon as one insurer starts with profitable selection, the others are forced to copy this strategy. Finally, one may argue that selection is not so much of a problem because doctors may be reluctant to discriminate among risks because of medical ethics. However, present ethics may change over time if the entire delivery system becomes more competitive.

How Good are the Risk-Equalization Formulas in Practice?

Currently the most sophisticated risk-equalization formulas can be found in the Netherlands, Belgium, Germany, and the USA-Medicare. How good are these formulas?

The results of an evaluation of the Dutch equalization formula 2007 indicates that this formula provides insufficient compensation for groups defined on health status, prior utilization, and prior expenses. These groups can be easily identified by the insurers. In case of an average premium, the average predictable losses per adult in these subgroups are

in the order of hundreds to thousands of Euros per person per year. For example, given an average community-rated premium the average predictable loss per adult for 21% of the population who report their health status as fair/poor, equals €541. The results also indicate predictable losses for groups of insured whose disease is included as a risk-adjuster in the equalization formula, for example, heart problems and cancer. Clearly not all of these patients fulfill the criteria to be classified as a patient eligible for a high equalization payment. It may be expected that other sophisticated risk equalization algorithms, such as the ones used in Belgium, Germany, and the USA-Medicare, yield similar results.

Lessons from the European Experience

In totality the European experience indicates, in accordance with the experience in the USA, that even the simplest risk equalization mechanisms are complex and that there are many start up 'surprise problems.'

In all countries the criterion 'appropriate incentives' did not appear to be a dominant one in choosing among different risk equalization models. On the contrary, redistributive effects among the insurers, feasibility (including acceptability) and fears for complexity were quite dominant criteria. In addition, one should not preclude, especially in the early days of risk equalization, an insufficient understanding of the problem. For example, the decision by the Swiss parliament in 1994 to limit the duration of the risk equalization model to a period of 13 years only can be easily countered. In autumn 2004 the Swiss parliament prolonged the formula for another 5 years, but voted once more against all propositions to improve the risk equalization formula. This decision reflects the compromise between the 49% arguing for further improvement and the 51% defending a deregulated liberal social health insurance.

Another country where risk equalization is a highly political issue is Ireland. Over a period of more than a decade there have been significant obstacles to the introduction of risk equalization because of political, legal, and implementation issues.

Another lesson is that even sophisticated risk equalization formulas currently in practice are not yet sufficiently refined and do not eliminate all incentives for risk selection that are caused by the community rating requirement.

Future Perspective of Risk Equalization in Europe

Given that current risk equalization schemes in most countries are far from perfect, the first priority should be further investment in improvement. Investment is needed both in better data and in research and development of better risk adjusters. In addition policymakers should seriously consider the use of ex-post cost-based compensation and reconsider the use of community rating. Policymakers must understand that risk selection is not inherent to the competitive insurance market, but is primarily the result of one possible form of regulation (i.e., community rating), and that alternative forms of regulation result in other outcomes.

Improving Risk Equalization

Current risk-equalization models in Europe can be improved by adding new health adjusters such as indicators of mental illness, indicators of disability and functional restrictions, multiyear DCGs rather than one-year DCGs, multiyear prior expenses and multiyear prior hospitalization and the enrollee's choice of a voluntary high deductible. In addition insurers might ex-post receive an ex-ante determined fixed amount for certain high-cost events (e.g., pregnancy) or diseases. New research efforts should in particular focus on individuals who are in the top 1% or top 4% of health care expenditure over a series of years because current risk equalization formulas perform worst for these groups.

The more risk equalization is improved, the more chronically ill people are likely to become preferred clients for efficient insurers, because the potential efficiency gains per person are higher for the chronically ill than for healthy persons. However, it is still questionable whether in practice a sufficiently refined risk equalization system is feasible. For example, approximately 6% of the Dutch population suffers from one or more of the 5000–8000 rare diseases for which the current formula does not compensate insurers and for which it is hard to find suitable risk adjusters. Because of the small number of people with each of these rare diseases the coefficients of the risk adjusters may change substantially from one year to the next, which conflicts with an essential precondition for ideal risk adjusters.

Improving Ex-Post Cost-Based Compensation

Ex-post cost-based compensation can be an effective complement to imperfect risk equalization, but also involve a selection-efficiency trade-off because they lower incentives for insurers to operate efficiently. However, the severity of the trade-off can be lowered by replacing existing compensation schemes with other forms of risk-sharing. Countries that currently have a uniform system in which insurers are retrospectively compensated for expenses above a threshold incurred by any enrollee (e.g., Germany), would be better off with a differentiated system in which a retrospective compensation is only given for individuals belonging to a small group of high risks determined in advance, for example, based on expenditures and hospitalization in the previous years. In this way it would be possible to increase insurers' financial risk without significantly increasing their incentives for risk selection.

A Better Understanding of the Regulatory Regime

Regulation of a competitive social health insurance market is a complex issue. Because many policymakers do not have sufficient understanding of the problem and the potential solutions society is often confronted with suboptimal regulatory regimes. An example is community rating. Most (if not all) policymakers confuse community rating as a goal and as a tool. Because their policy goal is that everybody in the community should pay more or less the same premium, they use mandatory community rating as a tool to achieve this goal. However, community rating creates incentives for risk

selection, which may result in the adverse effects outlined in Box 1. Therefore, although community rating has some important advantages (short-term affordability, transparency, and low transaction costs of organizing cross subsidies), it also has serious negative effects in the long term, particularly as a result of insurers' disincentives to provide good quality care to the chronically ill. Nevertheless, all European countries with a competitive social health insurance market require premiums to be community rated (although this most likely is in violation with the European regulation).

The major rationale for a competitive social health insurance market is to encourage insurers to be prudent purchasers of health care on behalf of their enrolees. Policymakers must understand that a condition sine qua non for achieving this aim is that insurers are adequately compensated for each enrolee – that is, they must receive a risk-based revenue related to each enrolee's predicted health care expenses, either from the enrolee in the form of a risk-rated premium or from a risk equalization scheme. Insurers will then focus on efficiency rather than on risk selection, and chronically ill people will become the most preferred clients for efficient insurers, rather than undesired predictable losses. This will in turn stimulate insurers to contract with providers who have the best reputation for high-quality well-coordinated care for chronically ill people.

An alternative for community rating is to allow risk-adjusted out-of-pocket premiums within a bandwidth, for example, a factor of two or four, in combination with subsidies for certain groups to improve affordability. Insurers are then free to charge risk-adjusted premiums provided their maximum premium does not exceed the minimum premium per product, for example, by a factor of two or four. With good risk equalization, the overwhelming majority of the premiums will be within the bandwidth. Policymakers might give this strategy a serious thought. Any information surplus the insurers might have would then be focused on premium differences rather than on selection. If the insurers are required to identify any risk factors they use for differentiation of their premiums, government could try to include the S-type risk factors in the equalization formula in subsequent years. (By definition government does not want to subsidize the N-type risk factors.) In this way the reduction of solidarity that results from the insurers' freedom to differentiate their premiums, may well be a short-term sacrifice to a long-term solution.

A Wider Application of Risk Equalization in Europe

So far most of the risk equalization literature in Europe focused on the social health insurance markets where consumers have a periodic choice of insurer. However, risk-adjusted payments can also be applied to noncompeting purchasers of care, for example, as risk-adjusted budgets for regions within a NHS such as in England, Italy, or Spain.

The recently proposed reforms in England to abolish the primary care trusts, to allocate approximately 75% of the NHS-budget to new general practitioner (GP)-consortia and to give the consumers a free choice of GP make the formula on which resources are allocated in the English NHS even more complicated. The formula must then be calculated at the individual consumer level, rather than at the small-area level. If one individual moves from one GP-consortium to another, this person's budget must follow the consumer. Therefore, the proposed health reforms provide England with a new challenge: how to prevent risk selection by the new GP-consortia? A first version of a person-based resource allocation formula for the NHS has already been developed.

England has a long tradition of research on resource allocation formulas. Therefore, the rich input of this English knowledge about resource allocation formulas, substantially enhances the risk equalization knowledge in Europe. A first issue is the concept of 'health.' In the risk equalization literature traditionally some crude proxies for health are used, such as DCGs, PCGs and other information derived from prior utilization, without much discussion about the validity of these indicators. In England, decades ago the discussion had already started about the concepts morbidity, need and demand, and about the difference between health status and need, or the various concepts of need, such as normative, felt, expressed, and comparative. Second, England has a long tradition of applying the supply of health care facilities as a so-called N-type risk factor in the allocation formula. This English experience may give interesting insights to countries as the Netherlands, Switzerland, and Germany, where supply of health care facilities is (still) considered an S-type risk factor. Third, researchers in England considered supply to be a function of health care needs and utilization, and applied econometric techniques to deal with this endogeneity problem. These insights will enrich the traditional risk equalization literature, where this endogeneity has been overlooked.

Provider Payment

Risk adjustment algorithms can also be used in the purchaser–provider relation. The rationale is that purchasers may give a risk-adjusted capitation payment to (a group of) providers of care to deliver a defined set of services to their enrollees and may try to share some of their financial risk with the providers of care. A simple form of capitation is the payment to GPs for the services they deliver themselves. If consumers have a choice among capitated providers, which mostly is the case, risk selection is an issue. It becomes more complicated if the providers' capitation payment also includes an ex-ante determined budget for several forms of follow-up care, as in the case of GP-fundholders. These follow-up costs may range from care prescribed by the GP, for example, prescription drugs, laboratory, and physiotherapy, to all follow-up costs (the so-called total-fundholder). In the latter case, in fact the GP functions as an insurer.

The functioning and effects of risk-adjusted capitations may strongly depend on the number of persons per capitated entity: for example, 2000 (a GP) or 2 000 000 (an insurer). In case of a small number of persons the conditions of the law of the large numbers are not sufficiently fulfilled, and the capitated provider may be confronted with a substantial financial risk because of large deviations from the statistically expected result. Forms of financial risk-sharing between the purchaser and the capitated provider of care can then be applied.

Risk-adjusted capitation payments and forms of financial risk-sharing ('bonuses') are essential elements of so-called pay-for-performance programmes. In these programmes it is crucial that the physicians' payments are sufficiently adjusted for the case-mix composition of their practices.

It is a great challenge to apply risk adjustment on the level of the physician and to analyze the consequences of the crucial differences and similarities between capitation payments for insurers and for primary care physicians.

Conclusion

Some conceptual issues for understanding the complexity of risk equalization have been discussed and an overview of risk-adjusted equalization payments in Europe has been provided. Most of the experience in Europe is with risk-adjusted payments to insurers in competitive health insurance markets. However, the relevance of risk adjustment for provider payment is increasing.

In practice, the risk equalization algorithms in all countries are imperfect and substantially undercompensate the high-risk enrollees. In addition, all European countries also implemented premium rate restrictions in the form of community rating. Consequently, the health insurers are confronted with incentives for risk selection, which may threaten affordability, efficiency, quality of care, and consumer satisfaction. There is evidence that risk selection is a serious issue in European countries. Some countries reduced the insurers' incentives for selection by giving them ex-post cost-based compensations. But these compensations also reduce their incentives for efficiency, resulting in a selection-efficiency trade-off. An alternative option that European countries may consider is to allow insurers to differentiate their premiums within a bandwidth, in combination with subsidies for certain groups to improve affordability.

The conclusion is that good risk equalization is an essential precondition for reaping the benefits of a competitive health insurance market. If insurers are confronted with substantial financial incentives to be irresponsive to the preferences of the chronically ill, the disadvantages of consumer choice of health insurer may outweigh its advantages. However, (also) the European experience indicates that in practice the implementation of even the simplest risk equalization scheme is very complex. This holds even more for the implementation of health-based risk equalization.

See also: Access and Health Insurance. Health Insurance in Developed Countries, History of. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in the United States, History of. Health Insurance Systems in Developed Countries, Comparisons of. Managed Care. Markets in Health Care. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Primary Care, Gatekeeping, and Incentives. Private Insurance System Concerns. Risk Adjustment as Mechanism Design. Risk Classification and Health Insurance. Risk Selection and Risk Adjustment

References

Lamers, L. M., van Vliet R. C. J. A. and van de Ven W. P. M. M. (1999). Pharmacy costs groups: A risk adjuster for capitation payments based on the use of prescribed drugs? *Report* (*in Dutch*) *of the Institute of Health Policy and Management*. Rotterdam: Erasmus University.

Lubitz, J. (1987). Health status adjustments for Medicare capitation. *Inquiry* **24**, 362–375.

Schokkaert, E. and Van de Voorde, C. (2003). Belgium: Risk adjustment and financial responsibility in a centralised system. *Health Policy* **65**, 5–19.

Van Vliet, R. C. J. A. and Lamers, L. M. (1998). The high costs of death: Should health plans get higher payments when members die? *Medical Care* **36**(10), 1451–1460.

Further Reading

Ash, A., Porell, F., Gruenberg, L., Sawitz, E. and Beiser, A. (1989). Adjusting medicare capitation payments using prior hospitalization data. *Health Care Financing Review* **10**(4), 17–29.

Bevan, G. (2009). The search for a proportionate care law by formula funding in the English NHS. *Financial Accountability and Management* **25**(1), 391–410.

Ellis, R. P., Pope, G. C., Iezzoni, L. I., et al. (1996). Diagnosis-based risk adjustment for Medicare capitation payments. *Health Care Financing Review* **12**(3), 101–128.

Lamers, L. M. and van Vliet, R. C. J. A. (1996). Multiyear diagnostic information from prior hospitalizations as a risk-adjuster for capitation payments. *Medical Care* **34**(6), 549–561.

Newhouse, J. P. (1996). Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature* **34**(3), 1236–1263.

PBRA Team (2009). *Developing a person-based resource allocation formula for allocations to general practices in England*. London: The Nuffield Trust.

Pope, G. C., Kautter, J., Ellis, R. P., et al. (2004). Risk adjustment of Medicare capitation payments using the CMS–HCC model. *Health Care Financing Review* **25**, 119–141.

Rice, N. and Smith, P. (2001). Capitation and risk adjustment in health care financing: An international progress report. *Milbank Quarterly* **79**(1), 81–113.

Stam, P. J. A., van Vliet, R. C. J. A. and van de Ven, W. P. M. M. (2010). A limited-sample benchmark approach to assess and improve the performance of risk equalization models. *Journal of Health Economics* **29**, 426–437.

Van Barneveld, E. M., Lamers, L. M., van Vliet, R. C. J. A. and van de Ven, W. P. M. M. (2000). Ignoring small predictable profits and losses: A new approach for measuring incentives for cream skimming. *Health Care Management Science* **3**, 131–140.

Van Barneveld, E. M., Lamers, L. M., van Vliet, R. C. J. A. and van de Ven, W. P. M. M. (2001). Risk sharing as a supplement to imperfect capitation: A trade-off between selection and efficiency. *Journal of Health Economics* **20**(2), 147–168.

Van de Ven, W. P. M. M., Beck, K., van de Voorde, C., Wasem, J. and Zmora, I. (2007). Risk adjustment and risk selection in Europe: Six years later. *Health Policy* **83**, 162–179.

Van de Ven, W. P. M. M. and Ellis, R. P. (2000). Risk adjustment in competitive health insurance markets. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, ch 14, pp. 755–845. Amsterdam: Elsevier.

Van de Ven, W. P. M. M., van Vliet, R. C. J. A., Schut, F. T. and van Barneveld, E. M. (2000). Access to coverage for high-risk consumers in a competitive individual health insurance market: Via premium rate restrictions or risk-adjusted premium subsidies? *Journal of Health Economics* **19**, 311–339.

Van Kleef, R., Beck, K., van de Ven, W. P. M. M. and van Vliet, R. C. J. A. (2008). Risk equalization and voluntary deductibles a complex interaction. *Journal of Health Economics* **27**, 427–443.

Weiner, J. P., Dobson, A., Maxwell, S., et al. (1996). Risk adjusted Medicare Capitation Rates Using Ambulatory and Inpatient Diagnoses. *Health Care Financing Review* **17**, 77–100.

Risk Selection and Risk Adjustment

RP Ellis and TJ Layton, Boston University, Boston, MA, USA

Glossary

Base period The period from which information is used to predict costs or other outcomes.
Community rating When insurance premiums are not allowed to vary across individuals.
Concurrent risk adjustment The use of variables measured in the prediction period to predict outcomes in the same period.
Conventional risk adjustment Risk adjustment models that focus solely on unbiasedness and statistical properties such as maximizing the models explanatory power (R^2).
Optimal risk adjustment Models of risk adjustment that incorporate behavioral objectives in setting plan premiums or other outcome targets, potentially allowing biased predictions.
Prediction period The period for which the regulator would like to predict an outcome.
Prospective risk adjustment The use of variables measured in a prior base period to predict outcomes in the prediction period.
Risk adjustment The use of information to explain variation in health-care spending or other outcomes such as resource utilization, mortality or health over a fixed interval of time, such as a quarter or year.
Risk selection When an individual's choice of insurance or a service is correlated with her cost (risk) to the insurer.
Service-level risk selection The act of distorting the level of services offered in an insurance contract in order to attract low risks; for example, the exclusion of diabetes specialists in an insurer's network to dissuade high-risk diabetics from enrolling in the plan.

Introduction

The problem of risk-based sorting often referred to as risk selection, and the use of risk adjustment to offset it are central concepts in health economics. After briefly defining risk selection and risk adjustment, this article provides an overview of the theoretical and empirical literatures that analyze these concepts. The issues covered here touch on numerous entries in this book, including health insurance, adverse selection, health plan competition, and death spirals, among others.

What is Risk Selection?

Risk selection occurs in health-care markets whenever consumers differ in expected cost (risk) that cannot be priced and make choices based on differences in risk, shifting the risk from the individual to the supplier. This choice can result in potentially inefficient or unfair sorting by average cost, quantity of visits, or quality. The most common reason for unpriced variation in risk is asymmetric information, in which consumers have private information about their health status, environment, or tastes for health care that insurers are unable to use when setting premiums. Incentives for risk selection can also be created even with full information when pricing is regulated, such as when regulators restrict the information that health plans are allowed to use when setting premiums or benefit features.

What is Risk Adjustment?

Although risk adjustment is defined in many ways, we offer one broad definition that includes almost all of the myriad ways the term is used: the use of information to explain variation in health-care spending, resource utilization, and health outcomes over a fixed interval of time, such as a quarter or year. Although it is not discussed, the term risk adjustment is also used in the health services research literature to refer to methods of explaining variation in a particular procedure or episode of treatment.

Theory of Risk Selection

Almost all theories of risk selection center on the choice of health insurance plans. In the classic Rothschild and Stiglitz model of risk selection, there are two types of consumers (high risk and low risk) and two states of the world (healthy and sick). The consumers differ in their probabilities of realizing the sick state, resulting in different expected costs in each potential state of the world. There is no moral hazard, so full insurance is optimal. Yet, under the assumptions of the model, a pooling equilibrium is either infeasible (when the low-risk types are unwilling to purchase the plan priced at a pooled premium) or inefficient. This model is recreated in graphical form in Figure 1. The two axes measure available consumption in each of the states of the world. If there is no insurance, available consumption is lower in the sick state due to health-care spending; hence, the initial endowment with no insurance is at a point such as E for both consumers. Because the two consumers differ in the probabilities of the two states of the world, their indifference curves between different levels of spending will diverge, with low-risk types having steeper indifference curves than high risks at every point. For risk-averse, utility-maximizing consumers, indifference curves between income in the two states of the world will be convex, and efficient consumption requires that insurance be provided until each type has equalized income in both states of the world (i.e., is on the 45° line). A possible outcome may be that

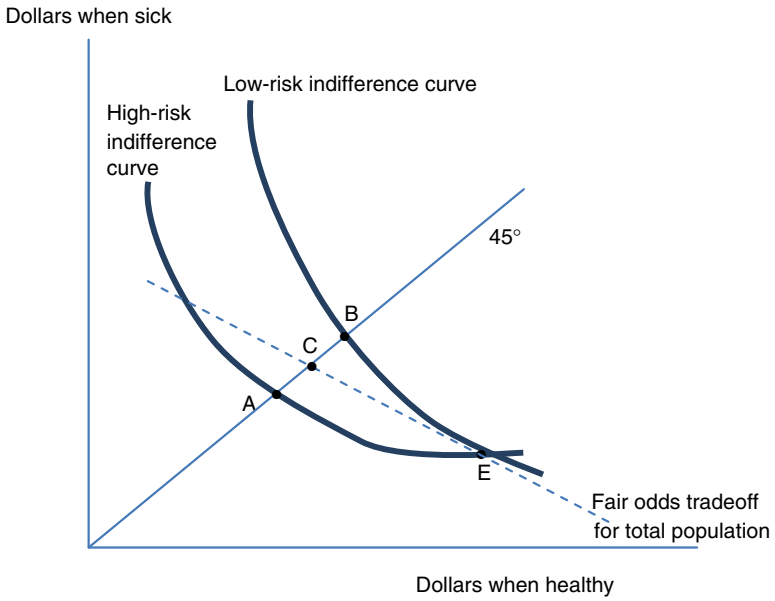


Figure 1 Indifference curves between consumption in two states of the world using the Rothschild and Stiglitz (1976) framework. Reproduced from Cutler, D. M. and Zeckhauser, R. J. (2000). The anatomy of health insurance. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics I*, pp. 563–637. Amsterdam: Elsevier.

the break-even pooled full insurance point in this framework is at a point such as *C*, which is preferred to no insurance by high-risk types but less preferred than no insurance by low risks, making it infeasible and inefficient. Moreover, even if a sponsor (the government or an employer) forces this option to be offered, the health plan will strongly prefer enrolling the low-risk types and may either distort plan offerings so as to be less attractive to high risks (high deductibles or cost sharing) or take costly efforts to avoid high risks.

Although the Rothschild–Stiglitz model is quite nice for describing risk selection among traditional indemnity health insurance plans, selection in the real world is more complex. In the US, managed care organizations (MCOs) such as Health Maintenance Organizations (HMOs), and Preferred Provider Organizations (PPOs) have captured a large share of the market for health insurance from the traditional indemnity plans. These plans often offer much lower cost sharing in return for much more tightly rationed health-care services. However, because consumer risk is still unpriced due to asymmetric information or regulation, the incentives for risk selection by profit-maximizing health plans are still quite strong, but the methods by which the selection occurs are likely to be quite different due to the low levels of cost sharing that are a hallmark of MCOs. The classic model for describing risk selection among managed care plans was formulated by Glazer and McGuire.

The Glazer and McGuire model of risk selection moves away from selection on the level of cost sharing offered by a plan toward a theory of 'service-level selection.' Because high-cost and low-cost individuals demand different services, MCOs can induce the high-cost individuals to avoid their plans by rationing the services demanded by these individuals more tightly than other plans. Likewise, they can attract low-cost individuals by rationing the services these individuals demand more loosely than other plans. For example, a profit-maximizing MCO may have incentives to ration care for a chronic condition like diabetes by including few or no diabetes specialists in its network. However, the MCO may want to provide easy access to acute services or alternative medicine like acupuncture or chiropractic services in order to attract the low risks. In equilibrium, MCOs offer less than the efficient quantity of some services and more than the efficient quantity of others.

Both types of models of risk selection describe inefficient equilibria due to a correlation between demand and unpriced risk. These inefficiencies lead to welfare losses. Note that the correlation does not have to be positive (high risks demand more, or adverse selection) to induce a welfare loss. If risk and demand are negatively correlated (high risks demand less, or advantageous selection), welfare losses still occur, but now the losses are due to plans offering 'too much' of something rather than too little. Also note that selection does not have to be limited to selection on cost sharing or to service-level selection. Selection can occur on any attribute including health plan quality (or service-level quality), through special offers such as gym membership discounts, etc. The key result, however, is that when unpriced risk is correlated with demand, inefficiencies and welfare losses are likely to occur.

Empirical Models of Risk Selection

Empirical models of risk selection are important for two reasons. First, they help us to determine where selection exists and on what characteristics selection occurs. If one knows where selection is a problem, one can implement solutions such as risk adjustment to fix the problem. If one knows what characteristics selection occurs on, one can use regulation to limit its effects. Second, they allow one to measure the welfare losses from selection. With a measure of welfare loss, one can

analyze trade-offs between inefficiencies caused by selection and inefficiencies caused by regulations intended to limit selection such as risk adjustment, reinsurance, and mandates.

A major difficulty involved with developing empirical models of risk selection is the confounding presence of moral hazard in health insurance. If there were no moral hazard, one could compare the average cost of individuals in Plan A with the average cost of individuals in Plan B and conclude that the plan with higher average cost is adversely selected. However, if Plan B has lower cost sharing or looser rationing of services and higher average cost, it is not clear whether the higher average cost is due to increased utilization due to the lower level of cost sharing (moral hazard) or to fundamentally higher cost individuals choosing Plan B because of the lower cost sharing (adverse selection). This problem can be solved with panel data and exogenous variation in health plan premiums, however. Essentially, moral hazard and adverse selection can be isolated by observing shifts in demand and corresponding shifts in average cost following a price change (see Einav and Finkelstein, 2011, for a graphical description of this method). Further complexity arises if adverse selection and moral hazard interact with one another. Individuals may choose a plan with lower cost sharing because they have a higher elasticity of demand rather than due to their higher risk (Einav *et al.*, 2013)

The method described above nicely allows for straightforward estimation of welfare losses. This method has produced estimates of welfare losses that are surprisingly small. These estimates are important because they allow researchers to use simulations to determine the welfare effects of various regulations such as incremental cost pricing, plan subsidies, or mandates. However, one strong assumption is necessary for the estimate of welfare loss to be complete: fixed contracts. The welfare losses being measured are really only those stemming from inefficient pricing of contracts. But the theoretical models of selection described above focus not just on pricing but also on the nature of the contracts themselves. The assumption of fixed contracts is likely valid in the context of employer provided insurance because the employer often chooses the plan parameters. However, in the context of a less regulated (or completely unregulated) market for insurance where insurers choose the majority of the parameters of the contracts they offer, the assumption of fixed contracts is likely to break down. In this unregulated environment, welfare losses occur not only through inefficient pricing of efficient contracts but through equilibria where only inefficient contracts are offered.

It is possible (and highly likely) that the welfare losses from distorted contracts are much larger than losses due to inefficient pricing. For example, coverage for mental health care in the US has been highly rationed in many health insurance plans because it attracts high risks. This is effectively a 'death spiral' that has occurred in a plan characteristic rather than of an entire plan. When contracts are not fixed, it is important for empirical models to be able to highlight the characteristics that selection occurs on because, as the Glazer and McGuire model of service-level selection points out, these are the characteristics that will be inefficiently rationed, and, thus, these are the services that regulators must focus on in order to achieve efficiency and minimize welfare losses. However, empirical models that can quantify the welfare loss due to these inefficient contracts are few in number due to the fact that these contracts are extremely complex due to the seemingly infinite number of parameters firms can vary (network size, cost-sharing parameters, in-network hospitals, etc.). Nevertheless, there is empirical evidence that inefficiencies such as service-level selection by MCOs exists, just no easy way to determine how much welfare is lost due to these inefficiencies.

Instead of trying to quantify welfare losses due to inefficient contracts, the empirical literature has sought to answer the question of why the welfare losses due to inefficient pricing are so low. This has resulted in interesting new theories and empirical evidence for interactions between selection and market frictions such as imperfect competition and switching costs and behavioral issues such as inertia and other mistakes.

The discussion of risk selection so far has focused primarily on the US setting where numerous, diverse health plans compete in multiple dimensions (premiums, benefit features, cost sharing, and selective contracting) with an important goal of attracting profitable enrollees. Similar structures of competition exist in Chile and Colombia. Several other countries also have multiple competing health plans (e.g., Belgium, Germany, Japan, Netherlands, Switzerland, and Israel), however benefit features, cost sharing, and premiums are regulated much more tightly in these countries, and selective contracting is relatively rare. Although selection incentives exist in these other competing health plan settings, the plans typically do not control providers and have relatively few tools available for influencing selection. Selection problems are typically even less of an issue in countries with a single social insurance plan (e.g., Canada, France, Denmark, Italy, Sweden, and the UK), although selection issues can still arise through competition among individual providers or geographically, where consumers get to choose among alternative local market areas. Selection concerns are also common when there are private complementary or supplementary insurance policies alongside of a single, publicly funded plan, as in Australia and Ireland.

Theory of Risk Adjustment

Remarkably, there is no unified or widely adopted theory of risk adjustment. Instead, there are models of risk selection that point to desirable features of risk adjustment models, and statistical models of risk adjustment that develop empirical risk adjustment models that satisfy various statistical properties (unbiasedness, minimum variance, robustness, and fair payment for subpopulations). One underlying reason why there is no unified theory of risk adjustment is that an important motivation for risk adjustment is usually equity, not just efficiency. With regard to efficiency-based arguments for risk adjustment, the appropriate risk adjustment model depends on the market and regulations in which competing health plans (or providers) operate. If premiums, cost sharing, and benefit plans are allowed to vary across consumer attributes that are observable to the health plan, then there will be no unpriced variation in costs or selection problems, and only fairness and equity concerns will remain. Once regulators

restrict premiums, cost sharing, and benefit coverage variation, risk adjustment is the classical tool for combatting risk selection.

Though there is no unified theory of risk adjustment, the literature has essentially assumed that the welfare loss from the distortions caused by selection are proportional to the sum of the squared differences between individuals' expected costs and the revenue a plan receives for those individuals. This assumption leads to the convenient result that the main goal of any risk adjustment system should be to minimize this sum of squared differences, or to maximize the fit of the payment system as measured by the R^2 statistic. Other theoretical models of risk adjustment have built on this assumption.

Glazer and McGuire were the first to develop theoretical models characterizing 'optimal risk adjustment,' which they distinguish from the existing statistical models that do 'conventional risk adjustment.' The central objectives of conventional risk adjustment are unbiasedness (paying each plan so that predicted costs equal actual revenue for each individual) and maximizing predictiveness (minimizing deviations between payments and expected costs). The essence of optimal risk adjustment is to allow biased risk adjustment models which optimally correct for identified incentive problems in health-care markets. Glazer and McGuire, (2000) choose to model the service distortion selection problem, in which competing health plans oversupply services that attract the healthy (e.g., acute care), and undersupply services that disproportionately attract the high cost, relatively sick (e.g., chronic care services). Since the signals used for risk adjustment are never perfect, even with conventional risk adjustment paying the expected costs it will be optimal for health plans to distort service offerings so as to attract the relatively healthy within a payment category, and deter the relatively sick. The solution Glazer and McGuire devise is to overpay on signals predicting a greater likelihood of being high cost, and underpay on signals predicting low cost, so as to undo the incentive to undertreat the high-cost enrollees. For example if only half of patients with asthma in a plan have their diagnoses recorded in the base period, and the incremental cost of the observed asthma patients is $500 higher than expected, then the plan should be paid twice this increment, or $1000 to compensate the plan for the under-reported patients with asthma. Conversely, one should pay less than the observed average cost for healthy signals in order to keep overall payments neutral. This twist in payments can in theory undo incentives to undertreat in capitated payment systems.

The service distortion problem that Glazer and McGuire model is a particular problem in the US, because many plans use selective contracting to increase or reduce the availability of specific types of services or providers, thereby influencing the attractiveness of their plan. Similar incentives and concerns arise in other countries, such as Australia and Ireland, where private insurance plans are allowed to choose the extent of coverage for services or copayments not covered generously by the public system. Other selection problems can require different optimal risk adjustment adjustments. For instance, in the US and several European countries (Germany and Belgium) there are concerns about intentional distortion of the signals used for paying competing health plans, or 'upcoding' the observed severity of patients.

Recent literature has explored risk adjustment in a setting where enrollee sorting on expected costs may not be as strong as sorting on the degree of risk aversion or other preferences. Two recent papers, (Bundorf *et al.*, 2012; and Glazer and McGuire, 2011) introduce the possibility that the demand for insurance is determined by both risk and taste and that there is not a perfect correlation between the two. This is especially relevant in the current environment in the US where an integrated HMO may be able to provide care for a chronically ill patient at a lower cost than a PPO but the chronically ill may prefer the PPO due to its wider selection of providers. Both papers show that in this environment consumers will have different incremental marginal costs, but the only way to get consumers to sort efficiently across plans is to charge each consumer her particular incremental marginal cost. Thus, if premiums are uniform across individuals, individuals may not sort efficiently across plans, even with perfect risk adjustment. Glazer and McGuire examine the market equilibria that occur under different regulatory arrangements, analyzing the tradeoff between efficiency and fairness. They show that if taste can be used as the basis of payment, both efficiency and fairness can be achieved using a tax. However, when taste cannot be used as a basis of payment (because it is not observed) and health status must be used instead, subsidies and taxes based on health status are required to achieve both efficiency and fairness. In other words, a uniform payment along with perfect risk adjustment is not enough.

Recent theoretical work is beginning to examine how to implement risk adjustment in the presence of imperfect community rating, which is to say that insurance plan premiums are allowed to vary within specified limits across certain individual attributes (such as age and smoking status). Further work is also examining how risk adjustment models can accommodate intentional benefit plan variation, such as is being allowed in the US health insurance exchanges where substantial variation in cost sharing is being permitted. This theoretical work is important because, as explained, the welfare loss from the distortions caused by selection incentives is proportional to the sum of squared differences between individuals' expected costs and the total revenues a plan received for those individuals. In the US, health insurance exchanges the total revenues can come from multiple sources: premiums, risk adjustment transfers, reinsurance payments, and risk corridor payments. It is clear that these sources of payments will interact with each other and those interactions need to be identified in order to determine how well they will fix the problems of risk selection and what other distortions they may cause.

Empirical Risk Adjustment Models

Early work in developing risk adjustment models focused on the statistical problem of maximizing the amount of variance in total spending that can be explained with available information (Ash *et al.*, 1989; Newhouse *et al.*, 1989). Even in this early work it was recognized that if lagged utilization or spending variables are used as explanatory variables, then the model is not only capturing the underlying illness burden, but also consumer taste for treatment, provider practice variation,

or differences in the underlying efficiency of treatment, which may lead to incentive problems. European risk adjustment implementation has been more precise than most US studies in distinguishing 'acceptable costs,' viewed as appropriate for risk adjustment, and 'unacceptable costs' which are viewed as ineligible for payment differentiation. Early work focused on self-reported measures from surveys that capture health status, however these measures are relatively expensive to gather and update, and not as highly predictive as insurance claims-based measures. In most modern risk adjustment models, diagnosis- and pharmacy-based information is used to predict spending. The extent to which each set of information is used in the models varies by country. The consensus view among risk adjusters and policy makers is that diagnoses and pharmacy signals, although not fully exogenous, are less endogenous than many other variables (such as health plan, provider type, access, taste, and consumer lifestyle), justifying their widespread use for risk adjustment.

From the onset, it has been recognized that health status information (whether self-reported, diagnoses, or pharmacy) from the base period can either be used to predict outcomes from the same period or the subsequent period (i.e., the future). The former is called concurrent (or sometimes retrospective) risk adjustment, whereas the latter is called prospective risk adjustment, and the two vary only in the prediction period. Most payment systems use prospective risk adjustment, due to concerns about endogeneity of the signals as well as the practical reason that it means that risk factors on which payments are based can be measured a year earlier than the spending being predicted. Concurrent models always have higher explanatory power than prospective models. For quality measurement or normalization of many other performance or outcome measures, a concurrent framework is widely used. Careful comparisons of predictive power from the two frameworks are provided in a US Society of Actuaries study authored by Winkelman and Mehmud, and in a series of studies conducted at York University in the UK.

A useful early contribution in the risk adjustment literature used fixed effects in panel data to calculate a 'lower bound on the upper bound' of what is potentially explainable at the individual level using time-invariant, prospective information. This method suggested that between 15% and 20% of the variance in spending was explainable using prospective variables. More recent studies suggest that the potentially achievable prospective R^2 is on the order of 25–35% of total health-care spending and varies with the population, year, and data quality.

To illustrate the importance of using more information than just age and gender to predict costs, consider **Figure 2**, which plots average 2009 covered health-care costs for each of the 65 one-year age cohorts in the US Truven MarketScan commercially insured claims and encounter data. **Figure 1** illustrates the importance of using relatively flexible specifications even for capturing age and sex adjustment of total health spending. The figure highlights that babies are disproportionately expensive, that women cost more than men through their childbearing years, and that in childhood males are slightly more expensive than females. These patterns are poorly captured by including a linear age term or even when using third or fourth degree polynomials of age. Most

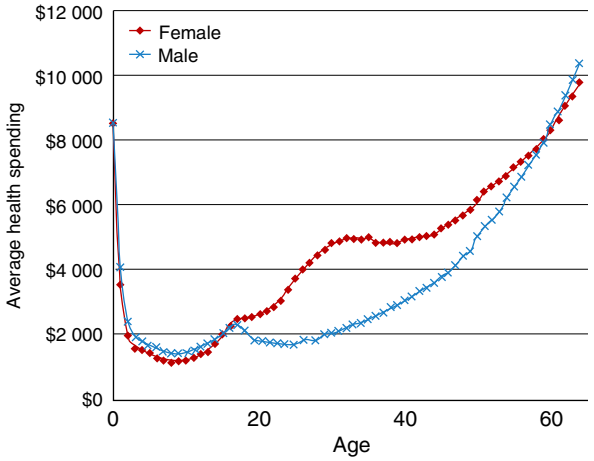


Figure 2 Actual spending by age and gender, 2009. Sample used is the US 2009 Truven MarketScan commercially insured claims and encounter data. All plan types and individual with a valid sex and age < 65 were included, although persons without pharmacy coverage were excluded. Each point plotted is the 1-year average total covered health spending per capita (medical plus pharmacy spending, including deductibles and copayments, but excluding dental and vision spending) for that 1 year age and gender group.

sophisticated risk adjustment models calibrated on large samples use 30 or more age–sex dummy variables to capture this nonlinear pattern.

Rather than only using (exogenous) age and gender, the most common approach used for risk adjustment is to use the rich information appearing on insurance claims as a proxy for individual health status. The most widespread information used is diagnoses, although pharmacy information is also common. Utilization measures (e.g., spending, hospitalizations, and counts of visits) are also highly predictive of future spending, although they contribute relatively modestly to the predictive power once a rich diagnostic model is used. Although claims-based information is only recorded when a visit to a health-care provider is made, and is potentially 'gameable' or amenable to manipulation, its strong predictive power and availability make it highly attractive.

Careful reviews of alternative risk adjustment models of total annual spending have been conducted in the US, Germany, and the UK, and are included in the recommended further readings section at the end of this entry. **Table 1** contains a few highlights of five diagnosis-based risk adjustment models used for payment by public insurance programs in the US (Medicare and Medicaid), as well as large numbers of private health plans. The interested reader can view further details at the references noted in the table.

A glimpse at the dimensions along which many risk adjustment models vary is summarized in **Table 2** from a Dixon *et al.* (2011) study using UK data. Looking first across the rows, age and gender alone only explain approximately 3–5% of total variation in spending at the individual level. Once diagnostic and prior utilization information are included in model (b), surprisingly little further variation is explained by including geographic variation (as captured by 152 geographical primary-care trust (PCT) dummies), 135 need

Table 1 Risk Adjustment models used for US public programs

Model feature	*Adjusted clinical groups (ACGs)*	*Chronic-illness disability Payment system (CDPS)*	*Clinical risk groups (CRGs)*	*Diagnostic cost groups (DCG)/ hierarchical condition categories (HCC)*	*Episode risk groups (ERGs)*
Background					
Model developer	Johns Hopkins	University of California, San Diego (UCSD)	3 M Health Information Systems	Verisk health (formerly DxCG)	Ingenix (formerly Symmetry)
Marketplace introduction	1992	1996	2000	1996	2001
Disease classification					
Additive/ categorical classification	Categorical	Additive	Categorical	Additive	Additive
Users: government programs	4 Medicaid	10 Medicaid	1 Medicaid	Medicare 1 Medicaid	1 Medicaid
Commercial (in 2009)	175	None	7	300 +	60
Prospective R^2:					
without truncation (%)	16.60	14.70	N/A	17.80	16.40
Truncated at $100 000(%)	21.80	20.80	N/A	24.90	24.40

Source: Reproduced from Weiner J. P., Starfield B. H., Steinwachs D. M. and Mumford L. M. (1991). Development and application of a population-oriented measure of ambulatory care case mix. *Med Care* **29**, 453–472; Kronick R. T., Dreyfus, T. and Zhou Z. (1996). Diagnostic risk adjustment for Medicaid: the disability payment system. *Health Care Fin Rev* **17**, 7–33; Averill R. F., Goldfield N. I., Eisenhandler J., et al. (1999). *Development and evaluation of clinical risk groups* (*CRGs*). Wallingford, CT: 3M Health Information Systems; Ash, A. S., Ellis, R. P., Pope, G. C., et al. (2000). Using diagnoses to describe populations and predict costs. *Health Care Fin Rev, Spring* **21**(3): 7–28; Symmetry Health Data Systems, Inc. (2001). *Episode risk groups: ERG user's guide*; Phoenix, AZ: Symmetry Health Data Systems, Inc; Prospective R^2's are from Winkelman and Mehmud (2007). Characterization of each system is from Florida Agency for Health Care Administration. (2009). Risk Adjustment Model Comparison. Available at: http://ahca.myflorida.com/Medicaid/quality_management/workgroups/managed_care/5_rar_model_comparison_050709.pdf (accessed 17.10.11).

Table 2 Results from the UK predicting FY2008 health spending per capita using prior 2 years of data

ID	*Explanatory variables in OLS models:*	*Number of parameters*	*Individual level R^2*		*Practice level R^2*
			Estimation Sample N = 5 206 651	*Validation Sample #1 N = 5 205 747*	*Validation Sample #2 N = 797*
a.	Age and gender only	38	0.0373	0.0366	0.3444
b.	Model (a) plus 152 diagnosis groups and 4 lagged utilization variables	194	0.2656	0.2610	0.7394
c.	Model (b) plus 151 geographic dummies	345	0.2659	0.2612	0.8046
d.	Model (c) plus 135 attributed need and 63 supply variables	543	0.2662	0.2615	0.8254
e.	Model (c) plus 7 attributed need and 3 supply variables	355	0.2671	0.2622	0.8254
f.	Age/gender, 152 diagnosis groups, 151 geographic dummies, 7 attributed need and 3 supply variables	351	0.1272	0.1229	0.7738

Notes: Diagnosis groups use only inpatient diagnoses from a two prior years. Utilization variables include inpatient episode count, outpatient visit count, dummy = 1 if any priority referral, and dummy = 1 if any outpatient visit; all measures are for prior two years. Estimation sample is a 10% random sample of the UK population. Validation Sample #1 is a different 10% random sample of the UK population drawn without replacement. Validation Sample #2 is a 100% sample of patients at 10% of primary-care practices. All results are from Dixon *et al.*, 2011, especially Table 7.4 and Appendix 13, Table 9. http://www.nuffieldtrust.org.uk/sites/files/nuffield/document/Developing_a_person-based_resource_allocation_formula_REPORT.pdf

variables (e.g., income, education, and prevalence of selected chronic conditions in the area), and 63 supply side variables (e.g., numbers of providers of various types and distances). Explanatory power at the individual level as measured by the R^2 differs only in the third or fourth decimal. The final row reveals that dropping the four prior utilization variables has a more significant effect on the model's predictive power, reducing the model's explanatory power by approximately half. Many would argue that the four lagged utilization variables are not only picking up health status heterogeneity,

but also patient and provider taste variation. (Key 'need' and supply side variables are still included in the model.)

Looking across the columns of Table 2 reveals that with 5 million observations in the estimation sample, there is no overfitting problem, even with more than 500 right-hand side explanatory variables. The final column shows that despite having only modest explanatory power at the individual level, where there is a great deal of individual patient randomness, the models do enormously better at the practice level where much of this randomness averages out. The third column sums up patients' actual and predicted spending to the level of 797 primary-care practices (averaging 6500 patients per practice) before using the conventional R^2 formula to calculate predictive power. The explained variation in spending at the practice level starts at 34% for the age–gender model, and increases to just more than 80% once geographic dummies are added in. Even the final model, which does not use the four utilization variables capturing patient and provider taste variation, explains 77% of practice-level variation in spending.

Risk adjustment has been used for more than three decades for the US Medicare Advantage (Part C) program, which offers diverse, competing private health plans to elderly and disabled individuals in the US as a voluntary alternative to conventional Medicare. The risk adjusted payments to health plans from 1985 to 1999 used only age, gender, Medicaid eligibility, institutional status (i.e., whether in a nursing home), and the county of residence of the enrollee to determine the payment amount. Since 2000 risk adjustment in the US Medicare program has used diagnostic information, initially using only inpatient diagnoses, but since 2004 diagnoses from outpatient clinician claims have also been used. After considering numerous alternative classification systems for diagnostic information, the Medicare program chose to implement the CMS Hierarchical Condition Category (CMS-HCC) classification system using 70 diagnostic groups for prediction. As of 2011, up to 86 HCCs are used, and the system is also used for Medicare Part D which includes prescription drug plans. Recent research has suggested that more sophisticated risk adjustment has led to less risk selection in the Medicare Advantage market, but insurance companies still have some ability to select low risks. They show that plans can still select on the individuals' risks, given risk adjustment. A more complete risk adjustment model that compensates plans for the average risk of as many targetable groups as possible might mitigate this problem.

In the UK, risk adjustment has been used for many years to allocate funds between geographically defined 'PCTs' using need, utilization, and health status variables, and done at the group level. More recent efforts in the UK have considered using individual information for risk adjusting payments not only to the geographically defined PCTs, but also to individual general practitioners. The main difficulty of using individual-level diagnostic information has been the process of obtaining this information from office-based physicians who are not required to record diagnoses as a condition of service payments, leading to exploration of models that use only inpatient diagnoses and counts of office and facility visits.

Risk adjustment models using a variety of adjusters are also used in all European countries with multiple, competing health plans, as well as in Chile and Colombia.

Econometric Issues

Risk adjustment models have been an active area for testing and developing new econometric methods. Early models used primarily linear models in part because the very large sample sizes and large number of explanatory variables made estimation of nonlinear models time-consuming if not infeasible. Since the 1990s and 2000s, there has been a surge of interest in building robust nonlinear models that are less sensitive to the outliers that are common in highly skewed expenditure data. The two-part log linear model used so widely in the Rand Health Insurance Experiment has been largely laid to rest by several studies that demonstrated the severe problems caused by uncorrected heterogeneity in such models. Among nonlinear models, Cox Proportional Hazard models, and Generalized Linear Models are the most widely used.

A central finding in the recent literature is that although nonlinear risk adjustment models may be superior for hypothesis testing, by creating test statistics for hypothesis tests that have well-behaved properties, the nonlinear models generally do worse than simple least squares models when used to predict sample and subsample means. Prediction of dependent variable means in levels in nonlinear models is seriously confounded by heteroskedasticity, which can be so multidimensional that it is very difficult to correct in medium-sized samples, and estimation of rich nonlinear models in mega samples needed to capture all of this heteroskedasticity are still hampered by the complexity of estimating precise models with hundreds (or even thousands) of parameters on multiple millions of observations. In sum, although nonlinear models can potentially produce better estimates, this improvement only comes through making various parametric assumptions. When these assumptions are not satisfied (as is probably often the case because they cannot be tested), simple least squares estimates are better because they do not require any parametric assumptions in order to be unbiased.

In recent years there has been a return of support for least squares models, as signaled by their use by researchers in Australia, Germany, the UK, and the US as well as for practical implementation in Belgium, Israel, Netherlands, and Switzerland. The preferred approach since 2000 for the US Medicare program has consistently been to use weighted least squares regressions of annualized spending on the risk adjusters, where annualized spending is actual spending divided by the fraction of the year a person is eligible, and this annualized amount is weighted by the fraction of the year a person is eligible to generate unbiased means. Such an approach replicates the mean exactly in disjoint groups, and is the only demonstrated approach that easily accommodates individuals with partial year eligibility. The mega-samples of multiple millions of observations, used to develop Figure 2 and Table 2 in this article, largely alleviate concerns about overfitting of outliers even with great skewness.

Future Directions in Risk Adjustment

Risk adjustment figures prominently in the US Affordable Care Act of 2010, notably in the proposals for establishing insurance exchanges to serve the individual and small group

insurance markets. To keep insurance affordable, premium subsidies will be offered by the government, and premium rate bands will limit premium variations across age and gender groups to be no more than three to one. It is readily seen from Figure 1 above that in the absence of regulation, plans would choose to charge 64-year-old males a premium that is approximately 10 times that of a 10-year-old male. Such regulated premiums can only be feasible if premium subsidies to plans are risk adjusted so that plans are paid for enrolling the aged and relatively unhealthy.

The US Department of Health and Human Services has developed new risk adjustment models for use in the insurance exchanges. The biggest changes between the new models and the CMS-HCC model are separate models for adults, children, and infants and the use of a concurrent rather than prospective framework. There are also different weights depending on the 'metal level' of the plan due to differing plan liability, which labels plans according to whether they are expected to cover 90% or more of health-care spending (platinum), 80% or more (gold), 70% or more (silver), or 60% or more (bronze). Because plan differences are explicit in terms of what is covered, selection incentives are strong, and aggressive risk adjustment is needed, which may explain the use of a concurrent rather than a prospective framework. The proposed concurrent models have R^2s from 0.289 to 0.360, much larger than the typical prospective model which typically has an R^2 approximately 0.12. The concurrent framework also accommodates the lack of previous diagnostic data for new enrollees. Although a concurrent framework solves this problem and substantially improves prediction, it also represents a payment system that begins to look similar to cost-based reimbursement. It will be interesting to see if it will bring some of the moral hazard problems (i.e., upcoding and increased utilization) that come with that type of payment scheme.

The exchanges have also introduced new questions about how risk adjustment interacts with other forms of plan payment. In the exchanges, plans will be paid through age-rated premiums and through risk adjustment, reinsurance, and risk corridor transfers, so the revenues a plan receives for an individual will be the sum of these payments, not just the risk adjustment payment. In the exchanges premiums can vary by age, and this age-based premium variation can lead to improvements in welfare by causing more efficient sorting of individuals to health plans. However, as risk adjustment and reinsurance payments compensate plans for age-based differences in cost, competition will cause the plans to vary premiums less and less. In the extreme (an extreme which is easily achievable) risk adjustment and reinsurance will fully compensate plans for age-based cost differences and premiums will not vary by age. This will lead to inefficient sorting. Hence, it is clear that when premium rating is allowed, there is a trade-off between minimizing selection incentives by maximizing the fit of a payment system and inefficient sorting caused by a lack of variation in premiums. It also remains to be seen how risk adjustment, premiums, and reinsurance will interact with taste heterogeneity.

New areas that are also receiving a great deal of attention are customized risk adjustment models that predict outcomes other than total spending. Predicting hospitalizations, length of stay, hospital resource use, readmissions, mortality, performance measures, and primary-care service needs are all examples of specific risk adjustment models that have been calibrated and are increasingly being used. Value-based payment is another example of a US reform that benefits from risk adjustment. With the new emphasis on detecting and rewarding good performance, risk adjustment is destined to see further expansion in use for these new outcomes globally.

Another important current area for risk adjustment in the US is in bundled payments to Accountable Care Organizations (ACOs), which are moderate-size health-care provider networks willing to receive a bundled payment in exchange for taking responsibility for providing all care to a panel of patients. Given the modest size of these panels, risk adjustment will be critical for ensuring that both healthy and sick enrollees are welcomed in the ACO. Comparing actual to risk-adjusted predictions of various performance outcomes within the ACO is also a key concept in these organizations.

A final important area for risk adjustment is in bundled payment for primary care, particularly as part of the Patient-Centered Medical Home (PCMH). In this CMS initiative, the Medicare program is encouraging primary-care providers to take responsibility for providing comprehensive primary care for patients from all payers (Medicare, Medicaid, and private) and offering increased primary care 'base payments' for the extra effort this will take (beyond what they will be reimbursed for via fee for service). These base payments will be partial capitation amounts, not fee based. Sizable bonus payments are also being considered to reward primary-care practices for achieving specified quality, cost, and patient satisfaction targets. If either the base payments or bonus payments are not risk adjusted, then primary-care practices could potentially act like insurance companies, striving to attract the healthy and avoid the relatively sick, undermining the potential of the PCMH initiative.

To date, risk adjustment models in the US have relied primarily on demographic and claims-based (usually diagnostic) information to adjust payments, utilization, and outcome measures. Occasionally self-reported information is used, although the relatively high cost of surveys and consumer input limit the widespread use of such information. A potentially huge source of information for the future are electronic health records, which capture not only what treatments are done, but also the results of various biometric and laboratory tests and imaging procedures. Health records will be challenging to use, but offer rich possibilities for improved prediction of diverse outcomes of key interest to researchers and policymakers.

See also: Health Insurance Systems in Developed Countries, Comparisons of. Long-Term Care Insurance. Managed Care. Modeling Cost and Expenditure for Healthcare. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Risk Adjustment as Mechanism Design. Risk Equalization and Risk Adjustment, the European Perspective. Sample Selection Bias in Health Econometric Models

References

Ash, A. S., Porell, F. W., Gruenberg, L., Sawitz, E. and Beiser, A. (1989). Adjusting Medicare capitation payments using prior hospitalization data. *Health Care Financing Review* **10**, 17–29.

Bundorf, K., Levin, J. and Mahoney, N. (2012). Pricing and welfare in health plan choice. *American Economic Review* **102**(7), 3214–3248.

Cutler, D. M. and Zeckhauser, R. J. (2000). The anatomy of health insurance. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics* I, pp. 563–637. Amsterdam: Elsevier.

Dixon, P., Dushieko M., Gravelle H., et al. (2011). Developing a person-based resource allocation formula for allocations to general practices in England. *Nuffield trust*. Available at: http://www.nuffieldtrust.org.uk/sites/files/nuffield/document/Developing_a_person-based_resource_allocation_formula_REPORT.pdf (accessed 15.07.11).

Einav, L. and Finkelstein, A. (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives* **25**, 115–138.

Einav, L., Finkelstein, A., Ryan, S., Schrimpf, P. and Cullen, M. (2013). Selection on moral hazard in health insurance. *American Economic Review* **103**, 178–219.

Florida Agency for Health Care Administration. (2009). *Risk Adjustment Model Comparison*. Available at: http://ahca.myflorida.com/Medicaid/quality_management/workgroups/managed_care/5_rar_model_comparison_050709.pdf (accessed 17.10.11).

Glazer, J. and McGuire, T. G. (2000). Optimal risk adjustment of health insurance premiums: An application to managed care. *American Economic Review* **90**, 1055–1071.

Newhouse, J. P., Manning, W. G., Keeler, E. B. and Sloss, E. M. (1989). Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Review* **10**, 41–54.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**, 629–649.

Winkelman, R. and Mehmud, S. (2007). *A comparative analysis of claims-based tools for health risk assessment*. Schaumburg, Ill: Society of Actuaries.

Further Reading

Ash, A. and Ellis, R. P. (2012). Risk-adjusted payment and performance assessment for primary care. *Medical Care* **50**, 643–653.

Breyer, F., Bundorf, M. K. and Pauly, M. V. (2012). Health care spending risk, health insurance, and payments to health plans. In Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds.) *Handbook of health economics* II, pp. 691–762. Amsterdam: Elsevier.

Ellis, R. P. (2008). Risk adjustment in health care markets: Concepts and applications. In Lu, M. and Jonnson, E. (eds.) *Paying for health care: New ideas for a changing society*, pp. 177–222. Germany: Wiley-VCH publishers Weinheim.

Ellis, R. P. and McGuire, T. G. (2007). Predictability and predictiveness in health care spending. *Journal of Health Economics* **26**, 25–48.

Iezzoni, L. I. (ed.) (2013). *Risk adjustment for measuring healthcare outcomes* (4th ed.). Ann Arbor, Michigan: Health Administration Press.

Pope, G. C., Kautter, J., Ellis, R. P., et al. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review* **25**, 119–141.

Rice, N. and Smith, P. (2001). Capitation and risk adjustment in health care financing: An international progress report. *Milbank Quarterly* **79**, 81–113.

Thomas, J. W., Grazier, K. L. and Ward, K. (2004). Comparing accuracy of risk-adjustment methodologies used in economic profiling of physicians. *Inquiry* **41**, 218–231.

Van de Ven, W. P. M. M., Beck, K., Buchner, F., et al. (2003). Risk adjustment and risk selection on the sickness fund insurance market in five European countries. *Health Policy* **65**, 75–98.

Van de Ven, W. P. M. M., Beck, K., Van de Voorde, C., Wasem, J. and Zmora, I. (2007). Risk adjustment and risk selection in Europe: 6 years later. *Health Policy* **83**, 162–179.

Van de Ven, W. P. M. M. and Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, pp. 782–845. Amsterdam: North-Holland.

Relevant Websites

http://www.nuffieldtrust.org.uk/sites/files/nuffield/document/Developing_a_person-based_resource_allocation_formula_REPORT.pdf
Nuffield Trust.

http://www.soa.org/files/research/projects/risk-assessmentc.pdf
Society of Actuaries.

http://sws.bu.edu/ellisrp/EllisPapers/2007_Ellis_Riskadjustment25.pdf
The Institute of Health Economics (IHE).

Sample Selection Bias in Health Econometric Models

JV Terza, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

Introduction

This article examines empirical models in health economics and health services research aimed at providing causal inference regarding the effect of a particular variable (the causal variable – X) and outcome of interest (Y). Such models are typically used to explain (predict) past (future) economic behavior, test an economic theory, or evaluate a past or prospective policy intervention. Common to all such applied contexts is the need to infer the effect of a counterfactual *ceteris paribus* exogenous change in X on Y, using statistical results obtained from survey data in which observed differences in X are neither *ceteris paribus* nor exogenous. The current article focuses on a particular survey context in which such lack of exogenous (and *ceteris paribus*) control in sampling can lead to bias in causal inference and prediction. In these cases, values of the outcome Y are not observable for all members of the relevant population and sampling preclusion is not random. Instead, it is governed by a systematic sample selection (SS) rule which is determined by both observable and unobservable factors. If the unobservable (and, therefore, uncontrollable) factors in the SS rule are common to (or correlated with) unobservable determinants of the outcome, then econometric methods that fail to take account of such correlation will likely produce biased estimates of causal effects. Selection bias will also plague predictions of the outcome obtained from such naive methods. For example, suppose a particular prescription (Rx) drug (henceforth, the drug) is under consideration for future deregulation and over-the-counter (OTC) sale, it would be of interest (e.g., to the producer of the drug) to know the deregulated (OTC) price elasticity of demand for the drug. In this example, Y is OTC consumption (demand) and X denotes drug price (or out-of-pocket (OOP) per unit payment for the drug). Until the drug is cleared for OTC sales (deregulated), only data on on Rx drug prices (or OOP payments) and Rx consumption can be obtained. Would-be OTC purchases cannot be observed because, before deregulation, the drug in question can only be purchased by prescription. In this case, the requirement that a prescription be obtained from a physician serves as a systematic selection rule precluding would-be OTC purchasers from consuming the drug. Suppose that

(a) physicians tend to prescribe the drug in question for the more severely ill,
(b) illness severity is unobserved and positively correlated with OTC demand for the drug, and
(c) price negatively affects physician prescribing behavior, then applying a method that ignores these facts (e.g. ordinary least squares (OLS)) to a sample of patients for whom the drug has been prescribed, and who are, therefore, consuming the drug, will likely produce a price elasticity estimate that understates the truth (i.e., is less negative than it should be).

As another example, suppose one seeks to predict the would-be utilization of a particular type of healthcare by currently uninsured individuals, if they were to become insured. One might consider OLS estimation of a health care utilization regression using a sample of insured individuals. The OLS results would then be used to predict utilization for the uninsured as if they were instead insured. Suppose, however, that the following are true:

(d) unobserved health status influences the probability of being insured (adverse selection or cream skimming)
(e) unobserved health status affects healthcare utilization, and
(f) the true predictor model differs between the insured and uninsured, then applying a prediction method (e.g., best linear prediction via OLS) that ignores these facts, to a sample of insured individuals, will likely be biased when used as a predictor of healthcare utilization for the uninsured if given coverage.

The remainder of the article is organized as follows. In the following section, a more formal discussion of SS bias is presented. In Section Using Control Functions to Correct for Sample Selection Bias, a commonly implemented remedy for such bias is discussed. The discussion therein begins with linear models. The role of instrumental variables (IVs) in this context is also discussed. More commonly, encountered (in health economics and health services research) nonlinear models and estimation methods are then considered. The final section summarizes and concludes.

Sample Selection Bias Because of Unobserved Confounders

At issue here is the presence of confounding variables which: (1) serve to mask the true causal effect of X on Y (TCE); (2) or bias predictions of Y that ignore them. Define a confounder as a variable that is correlated with both Y and: X; sample inclusion ($S=1$ if included, $S=0$ if not); or both. Confounders may be observable or unobservable (denoted C_o and C_u, respectively – in the current discussion both are assumed to be scalars (i.e., not vectors)). Here it is also assumed that there are no unobservable confounders for X (or for C_o) – that is, these variables are assumed to be exogenous. Observations on C_o can be obtained from the survey data, so its influence can be controlled in the modeling of Y and S and, therefore, in the estimation of the TCE. Clearly, C_o can be directly implemented in the prediction of Y. It may be assumed, however, that the correlation between C_u and Y, and between C_u and S cannot be ignored (e.g., unobserved factors that are correlated with OTC drug demand are also correlated with physician prescribing behavior). Moreover, one cannot directly control for C_u because it is unobservable. If left unaccounted for, the presence of C_u will likely cause bias in statistical inference regarding TCE and prediction. This happens because

 doi:10.1016/B978-0-12-375678-7.00908-1

estimation methods that ignore the presence of C_u will spuriously attribute to X (and C_o) observed differences in Y that are, in fact, because of C_u. Such bias is referred to as SS bias (or bias because of SS on unobservable confounders). SS bias can be formally characterized in a useful way. For simplicity of exposition, the true causal relationship between X and Y can be cast as linear and be written as

$$Y = X\beta + C_o\beta_o + C_u\beta_u + e \quad [1]$$

where β is the parameter that captures the TCE, β_o and β_u are parametric coefficients for the confounders, and e is the random error term (without loss of generality, it may be assumed that the Y intercept is zero). The SS rule determining the observability of Y can be modeled as

$$S = I(X\alpha + C_o\alpha_o + W\alpha_W + C_u > 0) \quad [2]$$

where the α's are parameters, W is an IV (i.e., an observable variable that is correlated with neither C_u nor Y – more on this later), and $I(A)$ denotes the indicator function whose value is 1 if condition A holds and 0 otherwise. In the naive approach to prediction and the estimation of the TCE (ignoring the presence of C_u), the ordinary least squares method (OLS) would be applied to

$$Y = Xb + C_o b_o + \varepsilon \quad [3]$$

using only observations for which values of Y are observed, where the b's are parameters and ε is the random error term. The parameter b is taken to represent the TCE and $Xb + C_o b_o$ is the relevant Y predictor. Correspondingly, $\hat{b}$ (the OLS estimate of b) estimates the TCE, and $X\hat{b} + C_o\hat{b}_o$ (with $\hat{b}_o$ being the OLS estimate of b_o) is the estimated predictor. It can be shown that OLS will produce unbiased estimates of b and b_o (here and henceforth, when unbiasedness is referred to it is done so in the context of large samples). It is also easy to show, however, that under general conditions

$$b = \beta + b_{X\lambda}\beta_u \quad [4]$$

and

$$b_o = \beta_o + b_{C_o\lambda}\beta_u \quad [5]$$

where $b_{X\lambda}$ is a measure of the correlation between X and λ (a function of $X\alpha + C_o\alpha_o + W\alpha_W$ – more on this later), and $b_{C_o\lambda}$ is similarly defined. As is clear from eqn [4], the bias of the naive OLS estimate of the TCE ($\hat{b}$) is $b_{X\lambda}\beta_u$. Moreover, it can be shown that the sign of $b_{x\lambda}$ is opposite that of the parameter α in eqn [2]. Consider the OTC drug demand example discussed in the introduction. Here, the TCE of price on OTC demand is β and under hypotheticals (a), (b), and (c), both β_u and $b_{X\lambda}$ are positive (the latter because α is negative). By the law of demand, β should be negative and, because $b_{X\lambda}\beta_u$ (the bias) is positive, the OLS estimate of the price effect on OTC demand will likely understate the true price effect (in absolute value). By a similar argument, in the insurance coverage and healthcare utilization example discussed earlier, hypotheticals (d), (e), and (f) imply that the bias of the OLS predictor $X\hat{b} + C_o\hat{b}_o$ is $X(b_{X\lambda}\beta_u) + C_o(b_{C_o\lambda}\beta_u)$.

An approach to estimation is needed that, unlike OLS applied to eqn [3], does not ignore the presence of, and potential SS bias because of, C_u. In the following section, methods that correct for selection bias through the inclusion of a control function which accounts for C_u are discussed. Such control functions also exploit sample variation in the IV (W) to eliminate SS bias because of correlation between C_u and S (more on this later).

Using Control Functions to Correct for Sample Selection Bias

As eqn [1] demonstrates, if C_u were observable then unbiased estimates of β and β_o could be obtained by applying simple OLS to (1) using the selected sample (i.e., the subsample with observable data on Y). As it turns out, if C_u (albeit unobservable) is assumed to follow a given probability distribution then, based on eqn [2], it can be shown that the following is true for the subset of the population with observable data on Y

$$Y = X\beta + C_o\beta_o + \lambda\beta_u + \nu \quad [6]$$

where λ is a function of $X\alpha + C_o\alpha_o + W\alpha_w$ and ν is the random error term possessing all of the requisite properties for unbiased regression estimation. This control function, which may be more explicitly stated as $\lambda(X\alpha + C_o\alpha_o + W\alpha_w)$, is the λ which is referred to in eqn [4] and eqn [5]. Its direct inclusion as a regressor in eqn [6] would serve to eliminate the SS bias plaguing regression estimation based on eqn [3] – made explicit in eqn [4] and eqn [5] for OLS. Strictly speaking, however, this is not feasible because λ involves the unknown parameters α, α_o, and α_W. Remainder of this section considers feasible linear and nonlinear estimators designed to circumvent the nonobservability of λ while producing unbiased estimates of the TCE of X and an accurate predictor for Y.

Unbiased Estimation of β and β_o in Linear Models

Despite the fact that λ is not directly observable, the parameters of linear models like eqn [1] can be estimated via the following two-stage method. First, estimate α, α_o, and α_W using the appropriate binary response model for eqn [2]. For example, if C_u is standard normally distributed, then estimates of the parameters of eqn [2] can be obtained by applying conventional probit analysis to a sample comprising observations with and without observable values of Y (i.e., both 'selected' and nonselected observations). The control function λ can then be estimated as $\lambda(X\hat{\alpha} + C_o\hat{\alpha}_o + W\hat{\alpha}_W)$, where $\hat{\alpha}$, $\hat{\alpha}_o$, and $\hat{\alpha}_W$ are the first-stage parameter estimates. In the second stage, unbiased estimates of β and β_o can be obtained by applying OLS to

$$Y = X\beta + C_o\beta_o + \hat{\lambda}\beta_u + \upsilon \quad [7]$$

using the subsample of observations for whom Y is observable ($S = 1$), where $\hat{\lambda}$ is the first-stage estimated value of the control function λ. If C_u is assumed to be standard normally distributed, then λ will have the familiar inverse Mill's ratio form and the two-stage estimator described here coincides with the Heckman-type SS model.

The inclusion of W (the IV) in eqn [2], and in the formulation of λ, warrants some discussion. Note that if the need

to control for the unobservable (C_u) in eqn [1] could be eliminated, the main source of selection bias would be neutralized and unbiased estimates of β and β_o could be obtained by applying OLS to

$$Y = X\beta + C_o\beta_o + e^* \quad [8]$$

where $e*$ is a random error term that fulfills the conditions for the unbiasedness of OLS. Note that it would not be required to control for C_u in eqn [1] if it were, indeed, NOT a confounder; in which case one could legitimately set β_u equal to zero. One way to break the confounding link between C_u and S would be to randomize the SS rule. Unfortunately, in applied health economics and health services research, as in other social sciences, explicit randomization (experimentation) is often prohibitively costly or ethically infeasible. A form of pseudorandomization is, however, possible in the context of survey (nonexperimental) data. If, for instance, a variable that is observed as one of the survey items is highly correlated with S but is correlated with neither Y nor C_u, then the sample variation (across observations) in the value of that variable can be viewed as providing variation in S that is not correlated with C_u – a kind of pseudorandomization for S. The IV W which was included in eqn [2] serves this purpose.

In the context of the OTC example discussed earlier, any variable that affects physician prescribing behavior, but is not correlated with OTC demand for the drug would be an IV candidate. For example, measures of individual physician overall preference for prescribing the drug have been used as IVs in similar empirical contexts. Likewise in the insurance coverage healthcare utilization prediction example, any observable variable that influences the likelihood of coverage that is not directly correlated with the type of healthcare usage in question can be used as an IV. For example, the existence and features of state-level government programs aimed at facilitating the acquisition of health insurance coverage have been used for this purpose.

It should be noted here that the inclusion of W in eqn [2] (and by implication in λ) is not required for the technical legitimacy, feasibility, or unbiasedness of the two-stage estimator described earlier. Notwithstanding this fact, applications of the two-stage estimator that do not include an IV – so-called identification solely via functional form – are generally viewed as lacking.

Unbiased Estimation in Nonlinear Models

The linear model (as specified in eqn [1]) does not conform to most empirical contexts in health economics. In most applied settings, the range of the outcome is limited in a way that makes a nonlinear specification more sensible. For example, the researcher is often interested in estimating the TCE of X on whether or not an individual will engage in a specified health-related behavior. In this case, the outcome of interest is binary so that a nonlinear specification of the true causal model would likely be more appropriate. In the OTC drug demand model discussed earlier, the outcome of interest (drug consumption) is nonnegative. An exponential regression specification of the true causal model is more in line with this feature of the data than is the linear specification in eqn [1]. Another common example of inherent nonlinearity in health economics and health services research, is in the modeling of healthcare expenditure or utilization (E/U). It is typical to observe a large proportion of zero values for the E/U outcome. In this and similar empirical contexts the two-part model (2PM) has been widely implemented. The 2PM allows the process governing observation at zero (e.g., whether or not the individual uses the healthcare service) to systematically differ from that which determines nonzero observations (e.g., the amount the individual uses (or spends on) the service conditional on at least some use). The former can be described as the hurdle component of the model, and the latter is often called the levels part of the model. Both of these components are nonlinear – binary response model for the hurdle; nonnegative regression for E/U levels given some utilization.

To accommodate these and other cases, the generic nonlinear version of the true causal model in eqn [1] can be written as

$$Y = \mu(X, C_o, C_u; \theta) + e \quad [9]$$

where $\mu(X,C_o,C_u;\ \theta)$ is known except for the parameter vector θ. It is very often assumed that $\mu(X,C_o,C_u;\ \theta) = M(X\beta + C_o\beta_o + C_u\beta_u)$, where $M(\ )$ is a known function and $\theta = [\beta\ \beta_o\ \beta_u]$. In this linear index form the true causal models corresponding to binary and nonnegative outcomes are commonly written, respectively, as

$$Y = F(X\beta + C_o\beta_o + C_u\beta_u) + e \quad (Y = \{0,1\}) \quad [10]$$

and

$$Y = \exp(X\beta + C_o\beta_o + C_u\beta_u) + e \quad (Y \geq 0) \quad [11]$$

where $F(\ )$ is a function whose domain is unit interval. It is to be noted here that for the generic nonlinear model characterized by eqn [9] the TCE is not embodied in any particular parameter (e.g., β) as in the linear models defined by eqn [1]. Instead, the TCE will be a nonlinear function of all parameters (θ) and all of the right-hand side variables (X, C_o, C_u) of the model. Moreover, the exact form of the TCE in nonlinear settings will differ depending on the researcher's policy relevant analytic objective(s). These issues will not, however, be discussed here and are the subject of other articles of this encyclopedia. The current discussion focuses on estimation of the vector of parameters θ.

The nonlinear generality of eqn [9] brings with it considerable, though not insurmountable, complications in the formulation of the nonlinear analog to eqn [6]. In the generic nonlinear model, although the SS rule is still defined as in eqn [1], the relevant control function is implicit and does not have a closed form, as did $\lambda(X\alpha + C_o\alpha_o + W\alpha_w)$ in the linear case. In light of this, accounting for the presence of C_u in eqn [9] does not involve a simple substitution of the control function, as was the case in moving from eqn [1] to eqn [6]. With these issues in mind, the nonlinear analog to eqn [6] can be written as

$$Y = \mu^*(X, C_o, C_u, W;\ \alpha^*, \theta) + v \quad [12]$$

where $\mu^*(X, C_o, C_u, W; \alpha, \theta)$ is a known function derived from eqn [9] and $\alpha^* = [\alpha\ \alpha_o\ \alpha_w]$. Unbiased estimates of the

parameters of eqn [12] can be obtained via the following two-stage protocol. First, estimate α^* as in the linear case. In the second stage, estimate of θ by applying the nonlinear least squares method to the following version of eqn [12].

$$Y = \mu^*(X, C_o, C_u, W; \hat{\alpha}^*, \theta) + \upsilon \quad [13]$$

where $\hat{\alpha}^* = [\hat{\alpha} \ \hat{\alpha}_o \ \hat{\alpha}_W]$ is the first stage estimate of α^*. It should be noted that the only case in which eqn [12] has a closed form is when it is derived from eqn [11] – the exponential regression model. In that case, eqn [13] can be written

$$Y = \exp(X\beta + C_o\beta_o)\hat{\lambda}^* + \upsilon \quad [14]$$

where the estimated control function $\hat{\lambda}^*$ is a function of $X\hat{\alpha} + C_o\hat{\alpha}_o + W\hat{\alpha}_W$.

Summary

Often sample inclusion is not random. Unobservable determinants of sample inclusion may also influence the outcome of interest. Naive regression estimates that ignore such latent correlation are subject to a kind of endogeneity bias – so-called SS bias. SS bias in regression parameter estimation is also manifested in corresponding causal inference and prediction. In this article the general circumstances in which SS bias is likely to be a problem is detailed and examples are given. Most empirical studies that confront potential SS bias are cast in a linear framework and implement a relatively simple two-stage method to correct for it. This method, and also the sources and implications of SS bias in linear models, are discussed in detail. Linear models and methods are not, however, compatible with most empirical contexts in health economics and health services research which often involve outcomes that are qualitative or otherwise limited in range. For this reason, a general nonlinear framework for modeling potential SS is also discussed, and a recently developed two-stage estimation approach (details of which can be found in the references for Further Reading) is outlined.

See also: Instrumental Variables: Methods. Modeling Cost and Expenditure for Healthcare. Models for Count Data. Models for Discrete/Ordered Outcomes and Choice Models

Further Reading

Gronau, R. (1974). Wage comparisons – A selectivity bias. *Journal of Political Economy* **82**, 1119–1143.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* **42**, 679–694.

Heckman, J. (1976). The Common structure of statistical models of truncation sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica* **48**, 1815–1820.

Ray, S. C., Berk, R. A. and Bielby, W. T. (1980). Correcting sample selection bias for bivariate logistic distribution of disturbances. In: *Proceedings of the Business and Economics Section of the American Statistical Association*, pp. 456–459. Alexandria, Virgina: American Statistical Association.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* **84**, 129–154.

Terza, J. V. (2009). Parametric nonlinear regression with endogenous switching. *Econometric Reviews* **28**, 555–580.

Terza, J. V. and Tsai, W. (2006). Censored probit estimation with correlation near the boundary: A useful reparameterization. *Review of Applied Economics* **2**, 1–12.

Searching and Reviewing Nonclinical Evidence for Economic Evaluation

S Paisley, University of Sheffield, Sheffield, South Yorkshire, UK

Glossary

Bibliographic databases Electronic database of bibliographic information organized, indexed, and searchable by subject descriptors, authors, title, and abstract keywords and other fields. Medline is an example of a bibliographic database.

Citation pearl growing or snowballing A search method using as a starting point a single, known, index source of information. This source is used to find related items (e.g., items citing and cited by the index source) and sources with similar characteristics (e.g., by the same author or containing similar subject terms).

High yield patch or rich patch Any source containing a high proportion of relevant information (e.g., clinical experts, existing economic evaluations).

Indirect retrieval or secondary retrieval Retrieval of information on one topic while searching for information on another topic.

Investigative searching A search method whereby information is identified in a piecemeal fashion rather than using a single, predefined search query. Sources of relevant information are used as leads or links to further relevant sources.

Precision A measure of the quality of a search. It is the proportion of items identified by a search strategy that is relevant.

Proximal cues A source of information that prompts a search for other similar or related, potentially relevant information.

Reference sources Source of information accepted or used on the grounds of its authority in the context of decision-making (e.g., drug formulary, classification of disease, and clinical guideline).

Routine data sources Source of information compiled primarily for administrative rather than research purposes (e.g., prescribing rates).

Search filters A predefined search strategy aimed at restricting search results to studies with specific methodological or clinical characteristics.

Sensitivity A measure of the quality of a search. It is the proportion of relevant items identified by a search strategy from all the relevant items that exist.

Outline

Economic evaluations in the form of decision-analytic models draw on many different types of secondary evidence in addition to costs and effects. The additional types of information used include natural history, epidemiology, quality of life weights (utilities), adverse events, resource use, and activity data. This information is drawn from different sources which can be research and nonresearch based.

Methods for identifying and reviewing evidence on randomized controlled trials (RCTs) of clinical effectiveness are well established. Some methods, many of which have not been validated, have been developed for other types of research-based information. Little guidance is available on identifying and assessing nonresearch based sources of information. On the whole, searching and reviewing methods have been developed in the context of systematic reviews and do not necessarily take into account factors specific to the task of developing decision-analytic models. This article considers how evidence is used to inform decision-analytic models and the implications of this for the methods by which evidence is identified and assessed.

The article considers the following:

- The types of information used to inform decision-analytic models for cost-effectiveness analysis.
- How to search for and to locate these types of evidence.
- How to select and to review evidence to inform a model.
- Factors specific to searching and reviewing in the context of decision-analytic modelling.

Introduction

Decision-analytic cost-effectiveness models aim to inform resource allocation decisions in health care. Evidence is assessed within a framework that reflects the complexity of the decision problem. A broad range of evidence, in addition to evidence on costs and effects, is required to support this approach. Important additional types of information include epidemiology, quality of life weights, natural history, and resource use. This information is drawn from different types of study design including experimental and observational research and from nonresearch based sources including routinely collected data, administrative databases, and experts.

A range of information retrieval methods is required to identify the diversity of information used to inform a decision-analytic models. Methods for identifying and reviewing evidence of clinical effectiveness from RCTs have been developed by organizations such as the Cochrane Collaboration. To a lesser degree, methods for the identification and review of other types of study design, such as observational studies, have been developed, although these have generally not been subject to validation and are not so well established. Very little guidance exists to support the systematic identification and assessment of nonresearch-based sources, including, for

 doi:10.1016/B978-0-12-375678-7.01403-6

example, sources of routinely collected data such as prescribing rates. Such sources can be difficult to locate and to access, and the navigation and interrogation of their nonstandard, often complex format, can be challenging.

On the whole, searching and reviewing methods have been developed in the context of systematic reviews of clinical effectiveness. As such they have been designed to identify and to assess experimental evidence with high internal validity in order to address single, focussed questions on the effects of treatment. Systematic review search methods require extensive searching in an attempt to identify all studies that match the characteristics of the focussed review question (typically defined by the populations, interventions, comparators, and outcomes of interest). The quality of a search is defined according to its sensitivity (that is, the extent to which it has identified all studies that match the review question). The systematic review search approach has become the benchmark approach to searching in all types of health technology assessment (HTA), including decision-analytic modelling, and the concept of sensitivity has become the defining characteristic of a high quality search. Such methods, however, do not necessarily take into account factors specific to the task of developing decision-analytic models such as addressing issues of complexity, drawing on a wide range of different types of evidence and assessing effectiveness in the absence of direct or good quality evidence. This article explores how evidence is used in models and the implications of this for the methods by which evidence is identified and assessed.

Types of Information and Forms of Evidence: The Classification of Evidence

Types of Information

The need for a range of information to inform model parameter estimates is well understood. Information on clinical effect size, baseline risk of clinical events, costs, resource use, and quality of life weights have been identified as the five most common data elements required in the population of decision-analytic models.

Information is also used to support a number of modelling activities in addition to parameter estimation. An analysis of the sources of evidence cited in the reporting of models identified a range of different types of information, drawn from different information sources and used for a variety of modelling activities (see Table 1). These included, in addition to model parameters, the definition of the model structure, the overall design and scope of the model, and various analytical activities and modelling methods.

The types of information used in models can be grouped into five broad categories. The first category relates to the condition or disease area of interest and includes information on natural history, epidemiology, and prognosis. Typical uses of such information include the specification of the model structure through the definition of the disease pathways and of health states within the pathway. Information is also used to provide estimates of baseline population characteristics and baseline risks of clinical events. Some evidence on costs, resource use, and utilities will also be included here.

Further information is required to inform the specification of the available treatment options and the management of the condition or disease. A range of information including clinical practice guidelines and policy documents, expert advice, prescribing rates, and other activity data can be used to inform the definition of management options including relevant comparators, treatment strategies and procedures, and management options at various stages of the disease pathway.

Category three relates to the costs and effects of the clinical interventions of interest. The types of information required include effectiveness evidence, comprising clinical outcomes, adverse effects, and quality of life weights. Adverse effects and quality of life weights often constitute additional, separate requirements where trials provide no or insufficient evidence on these outcomes. As such, searches focussing on these types of evidence might be undertaken in addition to searches for RCT evidence. Other types include information on the cost and resource use implications of delivering the interventions.

Some models might take into account factors that affect the uptake of treatment and that might impact on costs or effects outside the controlled environment of an experimental setting. Potentially relevant information includes patient

Table 1 Classification of evidence used in models of cost-effectiveness by type, source, and use of evidence

Types of information	*Types of source of evidence*	*Uses of evidence within model*
Adverse effects	Evidence synthesis	Design and specification of model framework
Compliance	Expert judgment	Model validation
Current practice	Methodological theory and empirical evidence	Modelling and analytical approach
Epidemiology	Observational research	Population of model parameters
Modelling methods	RCT (clinical and economic)	Sensitivity and uncertainty analysis
Natural history	Reference sources	
Patient preferences	Routine data sources	
Prescribing rates		
Prognosis		
Resource use		
Results and methods from other models		
Clinical outcomes		
Unit costs		
Utilities		

preferences, adherence, or acceptance of treatment as a function of the mode of delivery of an intervention such as homed-based versus hospital-based delivery, or oral versus infusional delivery.

The final category provides analytical rather than clinically related information. This is used to support choices relating to the modelling approach and analytical methods. It can include methodological standards and guidelines, empirical, theoretical methodological research, and modelling approaches used in existing economic analyses.

Formats of Evidence

The range of information used in decision-analytic models is drawn from a number of different forms of evidence. Examples are listed in Table 1. Although it is difficult to devise a definitive classification of the different evidence formats, it is possible to identify several study designs and formats which can be categorized broadly as research-based and nonresearch based evidence.

Research-based sources take the form of a number of different study designs, reflecting the different types of information used in models. These include evidence syntheses, including meta-analyses, systematic reviews, and cost-effectiveness models, RCTs, primary economic evaluations and observational study designs in the form of cross-sectional surveys and longitudinal cohort studies, and theoretical and empirically based methodological studies.

The use of nonresearch based sources reflects the need for models to address real world issues and to take account of the context of the decision. Such sources include expert opinion, both published and in the form of expert advice, routinely collected data and 'reference sources.' Routine data sources cover, for example, national disease registers (although some disease registers will be classed as observational research), life tables, and health service activity data. The category of sources referred to as 'reference sources' describes standard sources often providing generic information, including drug formularies or disease classifications, or sources that have some inferred authority due to consensus on their reliability or relevance to the context of the decision. Examples of the latter might include policy guidance and practice guidelines.

To some extent it is possible to associate the different types of information with the different forms of evidence. For example, it is well established that evidence of clinical effectiveness should ideally be taken from RCTs. Members of the Cochrane and Campbell Economics Methods Group (C-CEMG) have devised a useful hierarchy of evidence for the most important model data inputs, although it should be borne in mind that this does not cover all the types of information used to inform a model.

For those outcomes where there might be insufficient trial evidence, quality of life weights might be taken from cross-sectional observational studies, and evidence of adverse effects can be found in observational cohort studies and post-marketing surveillance data. Epidemiological information can be drawn from longitudinal studies, including long-term routinely collected data. Observational studies can provide information on factors impacting on the uptake of the intervention. Information on the latter might also be found in qualitative studies. This is useful where an issue is considered sufficiently important to warrant some form of discussion but where quantitative data for incorporation in a model are not available.

Evidence to inform the specification and estimation of costs and resource use can be difficult to categorize. In terms of the specification of a cost analysis (i.e., the identification and definition of relevant cost and resource groups) observational costing studies, previous economic analyses, expert opinion, and documentation of how a condition is managed, such as clinical guidelines, are useful. In terms of data with which to populate a model, routinely collected data such as prescribing costs and reference sources such as drug formularies and widely accepted compilations of unit costs are preferable. Where an overall estimate of the costs and resources associated with being in a particular health state are required (e.g., the cost of managing a stroke), it is necessary to use an existing cost analysis, for example, from a previous economic analysis. It is important, however, that consideration is given to the reliability of such a summary estimate.

The type of evidence used to inform the specification for the management of the condition of interest is also difficult to categorize. The definition of disease management and the specification of current practice are complex information needs requiring information on, for example, treatment options, disease management pathways, and clinical decision-making rules or policies. This is unlikely to be satisfied using a single source of information, particularly when a model is required to reflect the variations in disease management in different countries. To represent current practice it is necessary to draw together a range of different types of information to form an overall picture or description of how a disease or condition is managed. The sources from where this information can be found might include, but will not be restricted to, treatment and practice guidelines, activity data, and expert opinion.

In addition to the different types of evidence used to inform a model, it is also important to consider the scope of evidence, within each type, that will be relevant to the model. For example, it is likely that the scope of relevant clinical effectiveness evidence will not be restricted to that which relates to the intervention(s) of interest. In the absence of direct, head-to-head trials, evidence relating to the comparator(s) of interest should also be sought in order that indirect comparisons can be undertaken. This scope of evidence will be further extended where mixed treatment comparisons are undertaken. Here, a network of evidence, including trials of treatments that are not formal comparators, is required. The scope of relevant evidence will be further determined by the breadth or scope of the framework within which a decision problem is analyzed. Although the purpose of some models is to assess the cost-effectiveness of an intervention or interventions at a specific point in the disease pathway, this is done in the context of the whole of the disease pathway. Therefore, the effectiveness and cost-effectiveness of an intervention may need to be considered in terms of its impact on the whole of the disease pathway. For example, the effectiveness of a diagnostic intervention will be considered not just in terms of its diagnostic accuracy but on the extent to which it

impacts on the longer term management of a condition and, ultimately on the clinical endpoints relevant to the condition, including survival and quality of life. As such, evidence on, for example, treatment effects, quality of life weights, costs, and resource use is required not just for the intervention of interest at the point of the decision but for all important management options and health states at all stages of the disease pathway. This is an important consideration in defining the scope and range of searches to be undertaken.

Searching for and Locating Evidence

Models have multiple information needs that cannot be satisfied by a single search query. A series of focussed, iterative search activities should underpin the development of a model. To identify the diversity of evidence required, it is necessary to access a range of different types of information resources and to adopt a number of different information retrieval techniques.

Search methods for the retrieval of RCT evidence of clinical effectiveness are well documented by organizations such as the Cochrane Collaboration and are not covered in depth here. Bibliographic databases remain important sources, particularly for the identification of research-based evidence. However, given the diversity of information required, the totality of the evidence base used to inform any single model will be fragmented. In particular information from nonresearch based sources is scattered and the location and format of these sources can make retrieval difficult.

General biomedical bibliographic databases such as Medline and Embase provide access to a substantial volume of published information and can be interrogated using subject-specific keyword search techniques. Search filters, which are predefined combinations of keywords designed to identify specific study designs or types of information, are useful in identifying the individual types of information required for a model. For example, a number of filters exist for the retrieval of RCTs and cost information. Filters exist for many types of information. Most have been designed pragmatically and few have been validated. Nonetheless, filters are widely used to improve the relevance of search strategies. The Information Specialists' Sub-Group of Inter Technology Assessment Consortium, the UK academic network undertaking HTA for the National Institute for Health and Clinical Excellence (NICE), have developed an extensive resource of critically appraised search filters. In terms of non-RCT evidence, the Campbell and Cochrane Economic Methods Work Group provides advice on searching for economic evaluations and on quality of life weights. Research on the retrieval of evidence on adverse events is currently being undertaken at the University of York in the UK. These sources provide advice both on searching general biomedical databases and on specialist resources particular to specific topic areas.

Some useful specialist databases and resources exist for the retrieval of research-based non-RCT evidence. The Centre for Reviews and Dissemination at the University of York provides access to the National Health Service Economic Evaluation Database (NHS EED) and the HTA database, with the latter including the register of projects and publications of the member organizations of International Network of Agencies for Health Technology Assessment (INAHTA). The TUFTS Cost-Effectiveness Analysis Registry extracts utilities data from systematically identified cost-utility analyses. The Health Technology International (HTAi) Vortal is an extensive searchable repository of resources relating to the field of HTA.

Routine data and reference sources are disparate and cannot easily be brought together as a single reliable resource that can be interrogated using a uniform search approach. Compilations of resources have not been developed to any great degree, although some of the sources aforementioned can provide some coverage. Part of the value of this type of information is its relevance to the decision-making process in terms of geographical context or relevance to a specific decision-making authority. For example, the British National Formulary, the Office for National Statistics and NHS Reference Costs constitute three highly relevant sources for informing model parameter estimates in England and Wales, but do not carry the same authority in other decision-making jurisdictions. It would, therefore, be difficult to develop an international, generic, and comprehensive resource. Organizations such as the World Health Organization and the Organization for Economic Cooperation and Development can provide access to useful country-specific information. International resources such as HTAi and the International Society of Pharmacoeconomics and Outcomes Research (ISPOR) also provide useful starting points. At a national or more local level, harnessing knowledge and skills in order to provide access to compilations of information sources relevant to specific, local decision-making is an important area for development.

In the absence of such resources, the systematic identification of routine and reference sources of information cannot rely on the traditional keyword searching approach commonly associated with systematic review search methods. An alternative approach to retrieving information is to follow systematic lines of enquiry. This investigative search approach is similar in principle to citation pearl growing and snowballing references, used as supplementary search techniques in systematic reviews. Investigative searching does not attempt to identify everything in one search attempt using a single, catch-all, search query. Rather, the objective is to retrieve information in a systematic, auditable, and piecemeal fashion. This is done by identifying one or more relevant starting point(s), to search and to use these to identify relevant leads which can be followed up systematically and iteratively until sufficient information has been identified. A search starting point could be a known publication, advice from an expert or highly focussed exploratory searches using a bibliographic database (e.g., Medline) or an internet search engine (e.g., Google). The repeated following up of leads, known as proximal cues, in the form of new relevant keywords, references and names of authors and organizations will form a network of sources for further investigation. An illustration of a brief investigative search is given in **Figure 1**.

In terms of obtaining information from experts, the potential for bias is high and it is important to demonstrate a transparent and systematic approach to gathering this type of information. Formal methods of eliciting information from

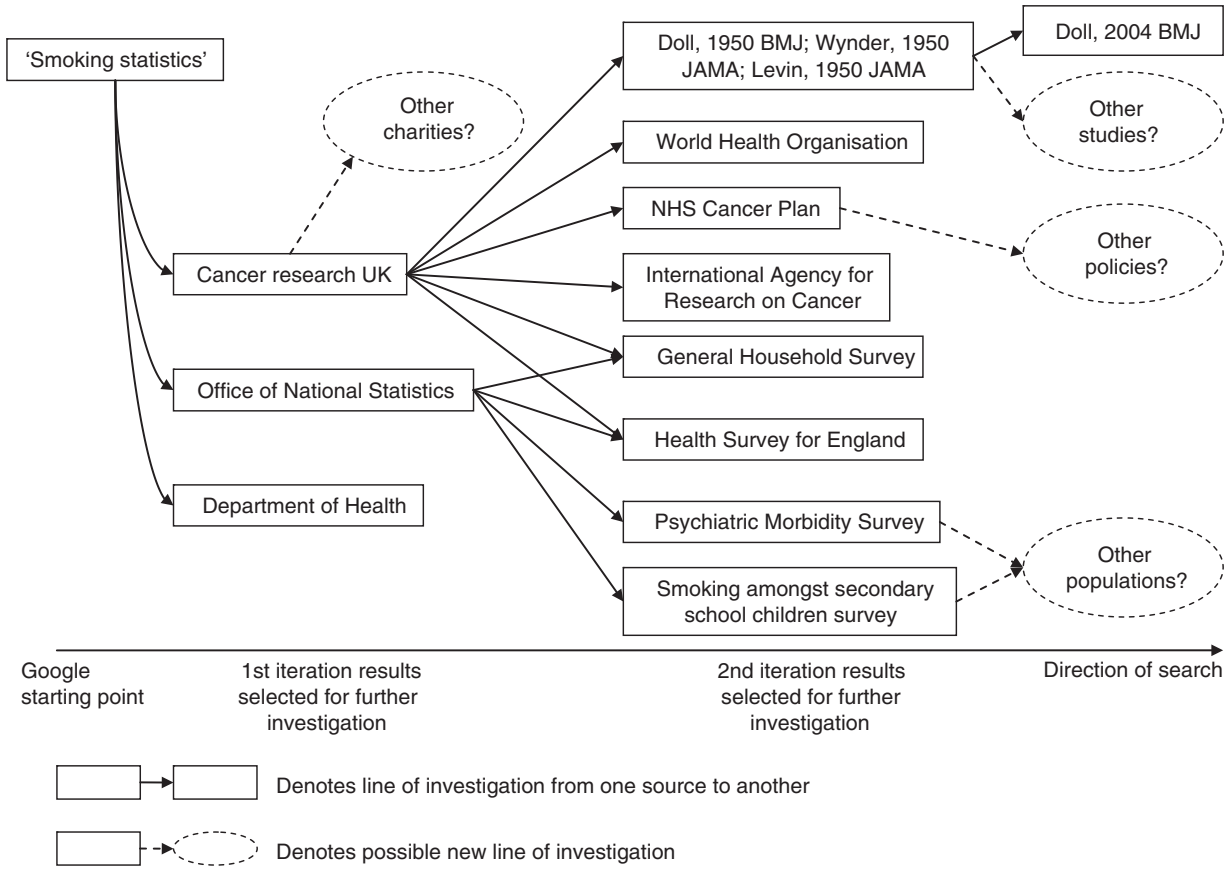


Figure 1 Example of investigative search.

experts include Bayesian elicitation methods, qualitative operational research techniques, and consensus methods. Other forms of good practice include the use of a range of experts in order to capture variation in opinion and in clinical practice. Precautions should also be taken to ensure that the full range of clinical expertise relevant to the decision problem is represented.

How to Search Efficiently

Searches using a systematic review search approach typically attempt to achieve a high-level sensitivity. That is, they are designed to maximize the retrieval of evidence that matches a prespecified and focussed topic. This often necessitates the retrieval of a large volume of irrelevant information in order to ensure that relevant evidence is not missed by the search. Exhaustive searching, aimed at high sensitivity, is considered difficult in the context of modelling. Limited time and resources are frequently cited as constraints when undertaking searches for models. The process of model development generates multiple, interrelated information needs that have, to some extent, to be managed simultaneously and iteratively rather than sequentially. This is very different from the single focussed question scope of typical systematic review searches.

To progress the process of model development it is necessary to assimilate a broad range of information as efficiently as possible. Using techniques that focus on precision, rather than sensitivity, might provide a more efficient approach to searching. Precision is defined as the extent to which a search retrieves only relevant information and avoids the retrieval of irrelevant information. Search techniques that focus on precision can provide a means of maximizing the rate of return of relevant information. The objective of such techniques is to front-load the search process by attempting to capture as much relevant information as early as possible, and to assess the diminishing returns of subsequent, broader, search iterations. Such techniques do not preclude further iterations of extensive searching where a more in-depth approach is required. All the techniques described aim to maximize precision. They are suggested with the caveat that when used on their own without subsequent broader iterative searching there is an increased risk of missing potentially relevant information.

A search process aimed at maximizing the rate of return of relevant information and at minimizing the opportunity cost of managing irrelevant information can adopt widely used information retrieval techniques. Restricting searches to specific fields within bibliographic databases is a commonly recognized technique aimed at maximizing precision. For example, searching for relevant terms within the title of journal articles should minimize the retrieval of irrelevant information. Depending on the nature and amount of information retrieved, a judgment can be made as to whether to extend the search across other fields, such as the abstract, with

a view to increasing sensitivity. A similar technique can be adopted when using search filters to restrict search results to specific study designs. Search filters can be designed to maximize either the sensitivity or precision of this restriction. The choice of high precision filters, sometimes referred to as one-line filters, can be used to maximize relevance. The Hedges project at McMaster University in Canada has developed and tested a set of filters, including one-line filters.

A highly pragmatic approach is to restrict the number of sources or databases searched. If a decision is made to extend the search, the results of the first iteration can be used to select and follow up highly relevant lines of investigation, for example, by carrying out focussed searches of newly identified keywords or by following up key authors.

Existing cost-effectiveness models in the same disease area may be important sources of information, and can be used to gain an understanding of the disease area, to identify possible modelling approaches and to identify possible evidence sources with which to populate a model. This can be described as using a 'rich patch' or 'high yield patch' of information whereby one source of information (i.e., the existing economic evaluation) is drawn on to satisfy multiple information needs. The use of high yield patches can provide useful shortcuts or can help cover a lot of ground quickly in terms of gaining an understanding of a decision problem. However, it is important that consideration is given to the limitations or reliability of a potentially rich source and to ensure that the weaknesses of existing economic analyses are not simply being replicated.

It has already been stated that information needs do not arise sequentially but that multiple information needs might be identified at the outset of the modelling process or might arise simultaneously during the course of developing the model. A useful way of handling multiple information needs is to consider information retrieval as a process of information gathering alongside a more directed process of searching. The pursuit of one information need might retrieve information relevant to a second or third information need. The yield of this secondary or indirect retrieval can be saved and added to the yield of a later more directed retrieval process. For example, a search for costs might also retrieve relevant quality of life information. This would constitute secondary or indirect information retrieval. It could be retained and added to the yield or results of a later search focussing specifically on quality of life.

Sufficient Searching

There is some debate as to what constitutes sufficient searching in the context of decision-analytic modelling. In particular, the need to achieve a high level of sensitivity when undertaking searches is open to question both on practical and theoretical grounds. The NICE has developed useful principles for the searching and reviewing of evidence for decision-analytic models. In terms of searching, this advice states that "the processes by which potentially relevant sources are identified should be systematic, transparent and justified such that it is clear that sources of evidence have not been identified serendipitously, opportunistically or preferentially." To uphold this principle it is important to consider the factors that might influence decisions to stop searching and that might inform judgments as to whether a sufficient search process has been undertaken.

This section summarizes some of the factors that could contribute to a definition of sufficient searching in the context of modelling and that might be used to support a judgment that sufficient searching has been undertaken.

In practical terms, it is often argued that there is not sufficient time or resource to undertake exhaustive, systematic review-type searching for every information need generated by the model. This supports the need for efficient methods such as those described in the previous section. It also requires pragmatic decisions on when to stop searching. Such decisions should be transparent in order that users of a model can judge the perceived acceptability or limitations of the scope of the searches undertaken.

Models are usually required to inform decision-making in the absence of ideal evidence. A decision to stop searching might be driven by an absence or lack of relevant evidence. If a relatively systematic search process exploring a number of different search options has retrieved no relevant evidence, it could be considered acceptable to assume that further extensive searching would not be of value.

One-way sensitivity analysis can be used to explore the implications of decisions to stop searching and of using a range of alternative sources of evidence on the outputs of a model. This type of analysis is widely regarded as a useful means of supporting judgments underpinning the identification and selection of evidence and is stated as a requirement in the Methods of Appraisal issued by NICE.

An extension of this idea would be to undertake some form of value of information analysis to understand the impact of uncertainty in the model and ultimately on the decision-making process. The process of bringing together, within one framework, multiple and diverse sources of evidence bring with it unavoidable uncertainty that cannot fully be understood or removed by comprehensive searching on every information need within the model. On theoretical grounds, therefore, it could be argued that exhaustive searching would not fully inform an understanding of uncertainty. Value of information analysis might be useful in assessing the value of undertaking further searching for more evidence and in determining where search resources should be focussed. It could act as a device to prioritize areas for further rounds of searching during the course of a modelling project or to develop research recommendations for more in-depth searching or reviewing on specific topics to inform future models. It is important to note, however, that the usefulness of this approach is dependent on timing and the extent to which the priorities for searching are revisited during the course of a project. A particular parameter which appears unimportant during the early stages of model development may become more important as the model is further refined over time.

Reviewing and Selecting Evidence

Reviewing techniques used in systematic reviews are, in principle, applicable in the context of reviewing and selecting

evidence for decision-analytic models. However, due to certain factors, gold standard systematic review methods such as those of the Cochrane Collaboration, are not directly transferable and consideration has to be given to adapting these useful and well established methods for modelling. These factors include time and resource constraints, the need to balance quality of evidence with relevance to the context of the decision and the frequent need to accommodate a lack of available, relevant evidence.

Time and Resource Implications

As in the case of searching for information, reviewing the breadth and diversity of evidence used in models is typically constrained by limited time and resources. The value of in-depth reviewing for every information need in the model, such as the use of strict inclusion criteria, extensive quality assessment procedures, and independent reviewing by two reviewers, is also open to question. Various pragmatic rapid review methods can be applied including reduced levels of data extraction, quality assessment, and reporting. In addition, important aspects of the model can be identified and prioritized as requiring a greater proportion of the available reviewing resource.

Assessing the Quality of Evidence

The systematic assessment of the quality of studies considered for inclusion in a systematic review is one of the many mechanisms aimed at checking for and minimizing the risk of bias. Quality assessment tools, sometimes in the form of checklists, exist for many different types of study design including RCTs, observational studies, and economic evaluations. The tools provide a series of questions on the conduct and design of individual studies allowing the systematic consideration of the strengths and weaknesses in terms of reliability, validity, and relevance. Many different checklists exist. Organizations such as the Centre for Reviews and Dissemination at the University of York provide access to a range of checklists. The Cochrane Collaboration, however, recommends against the use of checklists arguing that this leads to an oversimplification of the quality assessment process. The Collaboration has developed the Cochrane Risk of Bias tool, which can be applied to both RCTs and nonrandomized studies.

Given the diversity of information used to inform models it is highly likely that quality assessment tools do not exist for every type of evidence used. In such circumstances it might be possible to generate a number of questions, possibly based on existing tools, to guide the process of assessment. Members of the Cochrane and Campbell Economic Methods Working Group suggest a number of issues for consideration in the assessment of evidence on quality of life weights. The absence of standards for quality assessment is a particular problem in the assessment of routine data sources and reference sources. The quality of commonly used sources, such as national statistical collections, classifications of disease, and drug formularies might be regarded as sufficiently authoritative to be accepted without in-depth quality assessment. The ISPOR has created a Task Force on Real World Data to consider issues, including quality, relating to the use in models of routinely collected data and data not collected in conventional RCTs.

Two tools have been developed that can be used to support the process of quality assessment specifically in the context of decision-analytic models for cost-effectiveness. The tool devised by members of the Cochrane and Campbell Economic Methods Working Group is a series of hierarchies of evidence for five important model data inputs: clinical effect sizes, baseline clinical data, resource use, unit costs, and quality of life weights. Although the hierarchies do not allow the assessment of the quality of each individual study, they form a useful tool for some form of assessment between studies. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system provides an 'economic evidence profile'; a framework and criteria for rating the quality of evidence collected from all potential sources relating to all components that may be used to populate model parameters, including research-based and nonresearch-based sources.

Assessing the Relevance of Evidence

The relevance of evidence to the context of the decision is a crucial consideration in assessing evidence to inform decision-analytic models. This is an important distinction in the assessment of evidence for models compared with assessment for systematic reviews. Although standard quality assessment tools used in systematic reviews often include questions relating to relevance, the focus, in reality, is on the assessment of scientific rigor and internal validity. In its consideration of the role of routine data sources the ISPOR Real World Data Task Force emphasizes that 'context matters greatly' in determining the value of available sources. The C-CEMG hierarchies and the GRADE system both place emphasis on factors of relevance directly alongside the assessment of scientific quality. In assessing the value or usefulness of nonclinical sources of evidence, the trade off between contextual relevance and internal validity is often a particularly important consideration.

The tension between relevance and scientific quality is highlighted by the role of expert judgment as a source of evidence. As a basis for parameter estimates, expert judgment is placed at the bottom of the evidence hierarchy, as in systematic reviews. However, the role of expert judgment in interpreting the available evidence and in assessing the face validity or credibility of a model as an acceptable representation of the decision problem is recognized as an important source in the validation of the model.

Selecting Evidence for Incorporation in the Model

The assessment of quality and relevance plays an important part in the selection of evidence for use in a model. Although some initial form of eligibility criteria might be used to select references from a list of search results, this does not take the form of strict, predefined inclusion and exclusion criteria. This may be for a number of reasons. The incorporation of nonclinical evidence in a model tends not to involve the synthesis of data from a number of studies to generate a single pooled estimate. Rather, a range of estimates may be derived or selected from a number of possible options all of which might

be relevant for different reasons. In addition, there is often an absence of 'ideal' evidence. Although a clinical effectiveness review can remain inconclusive due to there being no available evidence, this is not an option for decision-analytic models which have to support a decision-making process regardless of the available evidence base. In doing so, the purpose of a decision model is to identify what might be, on average, the best option and to quantify the uncertainty surrounding the decisions. In the absence of ideal evidence the application of more flexible selection criteria allows the identification of the best available evidence from which the best or closest match can be judged and selected.

Therefore, evidence is not defined as being eligible or ineligible. Rather, the various characteristics of individual sources can be seen as offering different attributes by which eligibility might be judged. These attributes may be different to those of another source; that is, sources will be judged as being useful for different reasons. The quality-relevance tradeoff is an example of this.

The application of more flexible selection criteria allows the identification of a relatively varied set of potentially relevant or candidate evidence from which final selections can be made. The selection of evidence for incorporation in a model is made through a process of weighing up the attributes of each source against each other. The assessment of quality and relevance, sometimes supported by some level of data extraction to aid comparisons between studies, supports this process. This is different to the process of quality assessment and data extraction in standard systematic reviews which take place after evidence has been selected for inclusion. Moreover, the selection of one source of evidence over another does not necessarily lead to the exclusion of evidence as the implications of selecting alternative sources can be explored through sensitivity analysis.

The use in a standard systematic reviews of predetermined, strict selection criteria is another mechanism aimed at minimizing the risk bias in the review. The more flexible approach used in modelling incorporates choices based on weighing up and trading off the relative merits of alternative sources of evidence. This avoids the need to label evidence as being relevant or not relevant, allows the inclusion and exploration of a range of sources of evidence and, in the absence of ideal evidence, permits the incorporation of the best available evidence. In adopting this approach it is important that the process of selecting evidence is transparent, that the choices made are justified and that the extent and impact of uncertainty caused by possible bias resulting from the selection process is accounted for through sensitivity analysis.

A number of procedures can be used to systematize the process of selection. The use of quality assessment and data extraction can help systematize the choices being made. The reporting of the process, including the tabulation of key characteristics of the candidate sources of evidence, can improve transparency. The process of selection can be done through systematic discussion within the modelling project team in order that joint decisions are made. Finally, sensitivity analysis can assess the impact of uncertainty generated by the process of selection, including exploration of alternative sources not used in the base case analysis.

Conclusion

Decision-analytic models assess cost-effectiveness within a complex analytic framework. A broad range of information, in addition to evidence on costs and effects, and including both research- and nonresearch-based information, is required to inform this approach.

Information retrieval and reviewing methods, developed for the conduct of systematic reviews of clinical effectiveness, can be used for the identification and assessment of evidence used to inform decision-analytic models. To make the best use of these methods it is necessary to adapt them to the specific requirements of the task of developing decision-analytic models. This includes employing a range of information retrieval techniques in order to access and exploit diverse formats of evidence and adapting quality assessment and data extraction procedures to support systematic and transparent judgments in the selection of the best and most relevant available evidence.

Decision-analytic models permit an evidence-based approach to decision-making that takes account of the complexities and factors of specific relevance to the context of the decision problem. The collective grouping and interplay of diverse sources of evidence brings with it inevitable uncertainty. This cannot wholly be addressed by minimizing the risk of introducing forms of bias that might be associated with individual sources of evidence. Searching and reviewing methods used in the context of developing a model can provide an efficient means of capturing and understanding a broad range of information and can be used to address systematically and transparently the limitations of the best available relevant evidence such that a full account of the uncertainty associated with that evidence base can be supported.

See also: Observational Studies in Economic Evaluation. Problem Structuring for Health Economic Model Development. Synthesizing Clinical Evidence for Economic Evaluation

Further Reading

Centre for Reviews and Dissemination (2008). *Systematic reviews: CRD's guidance for undertaking reviews in health care*, 3rd ed. York: CRD, University of York.

Cooper, N., Coyle, D., Abrams, K., Mugford, M. and Sutton, A. (2005). Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *Journal of Health Services Research & Policy* **10**, 245–250.

Cooper, N. J., Sutton, A. J., Ades, A. E., Paisley, S. and Jones, D. R. (2007). Use of evidence in economic decision models: Practical issues and methodological challenges. *Health Economics* **16**, 1277–1286.

Egger, M., Davey, G. and Altman, D. G. (eds.) (2001). *Systematic reviews in health care: Meta-analysis in context*, 2nd ed. London: BMJ.

Garrison, L. P., Neumann, P. J., Erickson, P., Marshall, D. and Mullins, D. (2007). Using real-world data for coverage and payment decisions: The ISPOR real-world data task force report. *Value in Health* **10**, 326–335.

Glanville, J. and Paisley, S. (2010). Identifying economic evaluations for health technology assessment. *International Journal of Technology Assessment in Health Care* **26**, 436–440.

Golder, S. and Loke, Y. (2010). Sources of information on adverse effects: A systematic review. *Health Information and Libraries Journal* **27**, 176–190.

Golder, S., Glanville, J. and Ginnelly, L. (2005). Populating decision-analytic models: The feasibility and efficiency of database searching for individual parameters. *International Journal of Technology Assessment in Health Care* **21**, 305–311.

Higgins, J. P. T. and Green, S. (eds.) (2011). Cochrane handbook for systematic reviews of interventions: Version 5.1.0 (updated March 2011). Available at: www.cochrane-handbook.org (accessed 07.07.11).

Kaltenthaler, E., Tappenden, P., Paisley, S. and Squires, H. (2011). Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models. *Nice DSU technical Support Document: 13*. Sheffield: NICE DSU.

Paisley, S. (2010). Classification of evidence in decision-analytic models of cost-effectiveness: A content analysis of published reports. *International Journal of Technology Assessment in Health Care* **26**, 458–462.

Shemilt, I., Mugford, M., Vale, L., Marsh, K. and Donaldson, C. (eds.) (2010). *Evidence-based decisions and economics: Health care, social welfare, education and criminal justice*, 2nd ed. London: Wiley-Blackwell.

Weinstein, M. C., O'Brien, B., Hornberger, J., et al. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices-modeling studies. *Value in Health* **6**, 9–17.

Relevant Websites

http://bnf.org/
BNF (British National Formulary) British drug formulary proving cost information (example reference source).

http://www.c-cemg.org/
C-CEMG (Cochrane Campbell Economics Methods Group) Evidence synthesis methods for combining economics and systematic reviews.

http://www.cochrane.org/
Cochrane Collaboration Source of CDSR (Cochrane Database of Systematic Reviews), CENTRAL (randomised controlled trials register), Cochrane Handbook (systematic review methods manual).

http://www.york.ac.uk/inst/crd/index.htm
CRD (Centre for Reviews and Dissemination, University of York, UK) Source of specialist systematic review and HTA databases, systematic review methods manual, health economics resource lists.

http://www.gradeworkinggroup.org/
GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group GRADE quality assessment tool.

http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx
HIRU (Health Information Research Unit), McMaster University, Canada Hedges project search filter resource.

http://www.htai.org/
HTAi (Health Technology Assessment international) Generic source of HTA information, including HTA methods, HTAi Vortal.

http://www.who.int/classifications/icd/en/
ICD (International Classification of Diseases) International standard diagnostic classification (example reference source).

http://www.inahta.org/
INAHTA (International Network of Agencies for Health Technology Assessments) Generic source of HTA information, including HTA methods.

http://www.ispor.org/
ISPOR (International Society for Pharmacoeconomics and Outcomes Research) Generic source of HTA information, including HTA methods, modelling good practice guidelines, ISPOR Real World Data Task Force.

http://www.york.ac.uk/inst/crd/intertasc/
ISSG (InterTASC Information Specialists' Sub-Group) ISSG search filter resource.

http://www.guideline.gov/
National Guideline Clearing House –US database of clinical practice guidelines (example reference source).

https://www.gov.uk/government/publications/nhs-reference-costs-financial-year-2011-to-2012
NHS Reference Costs UK source of costs data (example routine data source).

http://www.nicedsu.org.uk/
NICE DSU (National Institute for Health and Clinical Excellence Decision Support Unit, University of Sheffield, UK) TSD series (Technical Support Documents) proving methodological guidance for submission to NICE of decision-analytic models.

http://www.oecd.org/
OECD (Organisation for Economic Co-operation and Development) International source of statistics (example routine data source).

http://www.statistics.gov.uk/default.asp
ONS (Office for National Statistics) UK statistical collections (example routine data source).

http://www.ncbi.nlm.nih.gov/pubmed/
PubMed – US National Library of Medicine open access Medline database (example biomedical bibliographic database).

https://research.tufts-nemc.org/cear4/default.aspx
Tufts CEA (Cost-Effectiveness Analysis) Registry Database of cost-utility analyses (example specialist database).

http://www.who.int/en/
WHO (World Health Organisation) – International source of statistics (example routine data source).

Sex Work and Risky Sex in Developing Countries

M Shah, University of California, Los Angeles, CA, USA

Introduction

The literature on the economics of commercial sex in developing countries has been a burgeoning area of recent growth for various reasons. First, unprotected commercial sex is a major human immunodeficiency virus (HIV) transmission vector. Sex markets play an integral role in the spread of sexually transmitted infections (STIs) including HIV/acquired immunodeficiency syndrome (AIDS). Each day 20 000 people become infected with HIV (UNAIDS, 2002), and many of these new infections occur in the developing world. In developing countries, sex workers play a central role in the spread of HIV and other STIs as they generally have higher infection rates and more sexual partners relative to the general population. **Table 1** displays HIV prevalence for adults, pregnant women, and sex workers across the developing world. In almost every country, sex workers have significantly higher HIV prevalence than the general population. In countries like Kenya, 85.5% of sex workers are HIV positive relative to 8.3% for the general adult population. Although condoms are an effective defense against infection, and despite the large amounts of financial aid channeled into educating sex workers about the importance of condom use, many sex workers are risking infection by not using condoms. Research in economics has sought to understand this issue because HIV/AIDS and other STIs have potentially devastating implications for economic development.

Table 1 HIV prevalence (per hundred) among adults, pregnant women, and sex workers

Country	*Adult*	*Pregnant women*	*Sex workers*
Benin	1.2	0.4	53.3
Burkina Faso	6.7	12.0	60.4
Cameroon	3.0	1.9	21.2
Congo, DR	3.7	4.6	30.3
Congo, Republic	7.2	7.1	49.2
Ivory Coast	6.8	11.6	67.6
Ethiopia	2.5	4.9	67.5
Ghana	2.3	2.2	30.8
Kenya	8.3	13.7	85.5
Malawi	13.6	32.8	78.0
Mali	1.3	3.5	55.5
Nigeria	2.2	3.8	22.5
Rwanda	7.2	25.3	87.9
South Africa	3.2	10.4	3.2
Uganda	14.5	21.2	86.0
Zimbabwe	17.4	35.2	86.0
Argentina	0.4	2.8	4.2
Brazil	0.7	1.7	11.2
Dominican Republic	1.0	2.8	7.0
El Salvador	0.6	0.0	2.0
Haiti	4.4	8.4	41.9
Honduras	1.6	1.0	20.5
Jamaica	0.9	0.7	24.6
Cambodia	1.9	3.2	43.0
China	0.0	0.0	0.3
India	0.4	0.3	51.0
Indonesia	0.05	0.0	0.3
Myanmar	1.5	1.3	18.2
Nepal	0.05	0.0	0.9
Thailand	2.1	2.4	18.8
Vietnam	0.07	0.0	0.24

Source: Reproduced from World Bank (1999). *Confronting AIDS: Public Priorities in a Global Epidemic*. New York: Oxford University Press.

Second, the sex market is a large source of employment for many women in poor countries, which has both microeconomic and macroeconomic implications. Vandepitte *et al.* (2006) estimate the percentage of females who make their living from sex work to be 12% in Diego-Suarez, Madagascar; 4.3% in Ouagadougou, Burkina Faso; 2.8% in Phnom Penh, Cambodia; 1.4% in Jakarta, Indonesia; and 7.4% in Belize. Obviously these are significant numbers and constitute a large number of women earning their living by performing potentially risky work. In addition, these numbers are increasing as more women enter the sex market due to lack of outside labor market opportunities in most developing countries. On a macroscale, the amount of revenue associated with the sex sector is considerable. For example, revenue from the Indonesian sex sector was estimated at between US$1.2 and 3.3 billion, or between 0.8% and 2.4% of the country's gross domestic product (Lim, 1998). In Thailand, close to US$300 million is transferred annually from urban sex workers to rural areas in the form of remittances (Lim, 1998). Therefore, both the financial turnover and the sheer number of women involved in the sex industry are strong motivators for further economics research.

Lastly, from an economics perspective, very little is known about the commercial sex sector, as demonstrated by the numerous policy failures. For example, many interventions aimed at teaching sex workers the alternate skills necessary to get them out of the sex industry have failed. The simplest explanation for this failure is that sex work pays well and most alternative low-skilled jobs do not, so getting women out of the sex industry is incredibly difficult. As another example, despite all the international funding that has been spent on educating sex workers about the risk of diseases such as HIV/AIDS, high rates of risky behavior (i.e., noncondom sex) are still observed in many developing countries.

This article investigates the following questions in the context of developing countries using recent studies from the economics literature:

1. Why might women enter the sex market?
2. Why do female sex workers engage in noncondom use?

To properly set the stage, this article begins with a brief description of sex workers' sociodemographics and transaction characteristics from three different countries: Mexico, Ecuador, and Kenya.

Table 2 Sex worker summary statistics

	Mexico, 2001	*Ecuador, 2003*	*Kenya, 2005–06*
Age	27.82	27.9	28.43
Age of first compensated sex	21.79	23.6	18.67
Very attractive (=1)	0.21	0.27	N/A
Has children (=1)	0.62	0.86	0.76
Can read and write (=1)	0.84	0.92	0.88
Had HIV test (=1)	0.89	0.85	0.60
Had sexually transmitted infections (STIs)/vaginal[a] problems (=1)	0.17	0.52	0.34
Civil status			
Single (=1)	0.41	0.47	0.44
Married/civil union/cohabitating (=1)	0.22	0.48	0.13
Divorced/separated/widowed (=1)	0.38	0.06	0.43
Used condom (=1)	0.91	0.88	0.82
Average transaction price[b]	44.75	7.20	6.40
Number of women	1029	2902	192

[a]A self-reported STI problem in last 3 months for Kenya, ever for Ecuador, and last year for Mexico.

[b]Denotes price in 2003 USD.

Note: This table reports the mean characteristics for sex workers from Mexico (Reproduced from Gertler, P., Shah, M., and Stefano, B. (2005). Risky business: The market for unprotected commercial sex. *Journal of Political Economy*, University of Chicago Press **113**(3), 518–550), Ecuador (Reproduced from Arunachalam, R. and Shah, M. (2008). Prostitutes and brides? *American Economic Review Papers and Proceedings* **98**(2), 516–522; Arunachalam and Shah, (2013) data), and Kenya (Reproduced from Robinson, J. and Yeh, E. (2011). Transactional sex as a response to risk in Western Kenya. *American Economic Journal: Applied Economics* **3**(1), 35–64).

Sex Worker Characteristics

Table 2 presents summary statistics of samples of sex workers in Mexico, Ecuador, and Kenya. The Mexican statistics are constructed from the dataset used in Gertler *et al.* (2005), the Ecuadoran statistics from the dataset used in Arunachalam and Shah (2008, 2013), and the Kenyan summary statistics from data used by Robinson and Yeh (2011). Please see the papers for more detailed information on sampling techniques and data collection activities. Interestingly, sex workers across all these three countries appear to be quite similar in terms of their sociodemographics. They are on average 28 years old, and the vast majority can read and write. Almost 50% are currently married or in civil union partnerships or were married and now are divorced, and more than 60% have children.

The majority of these women have taken HIV tests (89% in Mexico, 85% in Ecuador, and 60% in Kenya). In the past year, 17% of sex workers in Mexico self-reported having STI-related problems, 52% of women in Ecuador reported having an STI problem in the past, and 34% of women in Kenya reported STI-related problems in the past 3 months. Although HIV/AIDS rates tend to be much lower in Mexico and Ecuador (approximately 1% among sex workers), STI rates are relatively high, especially compared to the general adult population in these countries. High STI rates in a population are a risk factor for a potential HIV/AIDS epidemic because untreated STIs facilitate easier transmission of the HIV virus. Condom use rates range on average from 80% to 90% in the last three sexual transactions. The average transaction price varied from approximately US$45 in Mexico to US$7 in Ecuador to US$6.50 in Kenya.

Why Might Women Enter the Sex Market?

Understanding why women enter the sex market is crucial, especially if effective policies related to the sex market are to be implemented. There is obviously no one single reason that women and girls become prostitutes. Some have hypothesized that women enter because sex work pays well (Edlund and Korn, 2002), or because they face economic shocks in a world of poverty (Robinson and Yeh, 2012), or due to lack of outside options in the labor market, and of course there are those who enter due to force, kidnaping, and/or trafficking. Issues related to trafficking have not been discussed too much in the economics literature. This is likely due to the illegal, hidden nature of this market, which makes it difficult to collect good microdata, as well as the fact that many models in economics tend to implicitly assume some semblance of free choice. Without the intention of understating the importance of trafficking and dangers faced by many girls and women who enter the sex market via this channel, this article will not cover this area.

Edlund and Korn (2002) introduce a puzzling stylized fact that prostitution is "low-skilled, labor intensive, female, and well-paid." There is no other occupation like it that is female dominated and pays so well (although it is low skilled), and the authors offer a provocative explanation for this puzzle: sex workers draw a compensating differential due to the foregone opportunity to sell their fertility in the marriage market. Edlund and Korn (2002) not only provide the first formal model of occupational choice involving prostitution but also draw an intriguing link between the labor market and the marriage market that holds for only one occupation. When a woman chooses to become a sex worker, she relinquishes the compensation she would otherwise receive in marriage, because taboos prevent prostitutes from marrying. Thus, even in settings where prostitution is legal, it must draw an earnings premium. Beyond drawing considerable media attention, the richness of the Edlund–Korn model has made it the starting point for economists' discussions of sex work (e.g., Giusta *et al.*, 2004). An especially attractive feature of the paper is that it generates a number of testable predictions.

Arunachalam and Shah (2008) test the Edlund–Korn model. They utilize two large-sample datasets on sex workers, collected in Ecuador and Mexico, which they match to national labor survey data in the respective countries. They corroborate the existence of a sizable earnings premium for sex work, but fail to find support for the marriage-based explanation for this premium. Sex workers are actually more likely to be married than nonsex workers at younger ages when the earnings premium for sex work is highest. Furthermore, they find that the premium to male sex work is even larger than that for women. They hypothesize that the earnings premium would be better explained as a compensating differential, akin to that observed in other risky professions.

Although Robinson and Yeh (2012) agree that that sex work pays much better than other available jobs especially in poor countries and that this level difference in average income is clearly important, another key consideration is the variability of consumption. In Africa, as in much of the developing world, shocks are quite common and formal safety nets are often missing. In addition, insurance through informal systems of gifts and loans is rarely, if ever, complete (Townsend, 1994). Robinson and Yeh (2012) show that women enter the transactional sex market in western Kenya because clients send transfers in response to negative income shocks. These women develop relationships with regular clients who then become the primary source of interperson insurance that women receive. For example, transfers from regulars increase by 67–71% on the days around one's own illness and by 125% on the days around the death of a friend or relative.

Why Do Sex Workers Engage in Noncondom Use?

Much of the health policy literature argues that in many cases sex workers engage in unprotected sex because they are uninformed of the risks and that they would protect themselves if they fully understood the risks (World Bank, 1999; Lau *et al.*, 2002). In the cases in which sex workers are aware of the risk, others hypothesize that noncondom use occurs because condoms are either very expensive or not available at all (Negroni *et al.*, 2002), implying there are serious supply-side constraints. Alternatively, others have argued that sex workers are forced to have unprotected sex (Karim *et al.*, 1995; World Bank, 1999), and because they face physical or economic threats, they engage in noncondom use.

Although ignorance does exist and the forced exploitation of sex workers does occur, another possible explanation is that sex workers are willing to risk infection by not using condoms with clients if they are adequately compensated. Indeed, economic theory has long posited the general principle of compensating wage differentials (Rosen, 1986), and a number of authors have documented wage differentials that compensate for risky work activities in other labor sectors such as mining, police work, and firefighting (Viscusi, 1992; Siebert and Wei, 1998). Although there is anecdotal evidence that sex workers charge more for sex without a condom (Ahlburg and Jensen, 1998), there was little empirical evidence testing this claim before research from the economics community. In addition, it has been widely documented that men have strong preferences for noncondom sex. Therefore, if men are willing to pay more for sex without a condom, then sex workers might simply respond to these market incentives.

Understanding why sex workers do not use condoms is critical for the development of policy that is effective in increasing condom use and consequently in reducing the transmission of STIs, including HIV. The usual policy recommendations are to intervene on the supply side (World Bank, 1999). These policies include (1) educating sex workers about the risks, (2) increasing access to inexpensive condoms, (3) reducing environmental barriers to condom use by working with gatekeepers such as brothel owners and the police, and/or (4) empowering sex workers by improving their negotiating skills and fostering self-help organizations. Additionally, governments are urged to implement and enforce laws against human trafficking, rape, assault, and indentured servitude. However, if some clients are willing to pay substantially larger sums for unprotected sex, supply-side interventions alone are less likely to sufficiently reduce unprotected commercial sex. Even knowledgeable sex workers with condoms, who are free to turn down clients, might be willing to supply unprotected sex if the price is right. In this case, complementary interventions on the client side that reduce the demand for unprotected sex are also necessary in order to increase condom use.

Gertler *et al.* (2005) investigate whether sex workers are rationally responding to market incentives by testing whether sex workers charge more to take the risk of providing unprotected services. However, selection is an issue due to both sex worker heterogeneity as well as client sorting. For example, in terms of sex worker heterogeneity, better educated sex workers might charge higher prices and also be more likely to use condoms. Similarly, better educated, wealthier clients who value condom use and have a higher willingness to pay may select these sex workers. This will create a positive correlation between price and condom use, which is not necessarily causal or related to compensation for taking a risk. To control for the endogeneity of condom use, they collect information on the last 3–4 transactions for each sex worker to create a panel dataset. They estimate a model with sex worker fixed effects to control for bias from both unobserved sex worker heterogeneity and client selection where the dependent variable is log price. Additionally, they control for client characteristics using sex worker reports of clients' looks, wealth, cleanliness, and risk preferences.

Gertler *et al.* (2005) begin by constructing a simple bargaining model of commercial sex that has a number of empirically testable predictions. The model predicts that a condom will not be used when the client's maximum willingness to pay not to use a condom is greater than the minimum the sex worker is willing to accept to take the risk. Surprisingly, however, the model also predicts that when the client is worried about the risk of infection from unprotected sex, he may be charged more for using a condom than for unprotected sex. Similarly, when the sex worker prefers not to use a condom, the client is given a discount for not using a condom. The price differential between protected and unprotected sex is a weighted average of the maximum the client is willing to pay for not using a condom and the minimum the sex worker is willing to accept to take the risk of infection by not using a condom. The weights are a function of the relative bargaining power of the client and the sex worker.

Table 3 Sex worker risk premium for noncondom sex

Risk premium for noncondom sex (%)	*Noncondom sex (%)*	*Country*	*Source of risk premium*
13	10	Mexico	Gertler *et al.* (2005)
12	12	Ecuador	Arunachalam and Shah (2013)
66	47.2	India	Rao *et al.* (2003) Dandona *et al.* (2006)
9.3	17	Kenya	Robinson and Yeh (2011)

Note: This table reports both risk premium for noncondom sex and noncondom sex rates by country from various studies.

Gertler *et al.* (2005) test these predictions using a panel dataset collected in 2003 from the Mexican states of Michoacan and Morelos. They find that Mexican sex workers receive a 23% premium for unprotected sex from clients who requested not to use a condom, and this premium jumps to 46% if the sex worker is considered very attractive. They also find that clients who requested condom use paid 9% more for protected sex, and sex workers who requested not to use a condom gave clients a 20% discount. The results are completely consistent with the theoretical predictions.

Studies in other developing countries have found similar results with male clients paying a premium for noncondom use in India (Rao *et al.*, 2003), Ecuador (Arunachalam and Shah, 2013), and Kenya (Robinson and Yeh, 2011). **Table 3** summarizes the main results from all these studies. Arunachalam and Shah (2013) and Robinson and Yeh (2011) use a similar sex worker fixed effects empirical strategy as Gertler *et al.* (2005).

Interestingly, Arunachalam and Shah (2013) show that the premium men pay for noncondom sex in Ecuador is approximately 12% but that this premium increases in locations with higher disease rates. A one percentage point increase in the local STI rate increases the premium for noncondom sex by 33%. This is the first paper in this literature to have biological STI outcomes for sex workers. Therefore, the authors are able to identify this source of the risk premium as directly linked to STI rates. These results suggest that market forces may curb the self-limiting nature of STI epidemics exacerbating the spread of disease. To a greater extent than other epidemics, economists argue that the spread of STIs is shaped by individuals' behavioral responses. For example, with an increase in awareness of the risk of contracting disease, individuals substitute away from risky sex toward abstinence (Kremer, 1996), toward protected sex (Ahituv *et al.*, 1996; Dupas, 2011), or away from sex with men toward sex with women (Francis, 2008). Viewing risky sex much like other commodities in the market, economists anticipate that demand declines as the expected cost increases (Posner, 1992). Hence, economists tend to see behavioral responses to STI prevalence as generating a self-limiting incentive effect of epidemics (Geoffard and Philipson, 1996). Evidence from the commercial sex sector, however, suggests that market forces may dampen the self-limiting feature of STI epidemics be cause sex workers draw a premium for engaging in risky unprotected sex.

Robinson and Yeh (2011) use a panel dataset constructed from 192 self-reported diaries of women who provide transactional sex in Busia, Kenya. They find that women who engage in transactional sex increase their supply of risky, better compensated sex to cope with unexpected health shocks, particularly when another household member is ill. More specifically, they find that women are 3.1% more likely to see a client, 21.2% more likely to have anal sex, and 19.1% more likely to engage in unprotected sex on days in which another household member (typically a child) falls ill. These behavioral responses obviously entail significant health risks for these women and their partners, and suggest that women are unable to cope with risk through other consumption smoothing mechanisms. This is an extremely critical issue in a place like Busia, where the estimated HIV prevalence was 9.8% in 2004 (CBS, 2004).

In India, Rao *et al.* (2003) use 2003 data from Songachi, Kolkata's oldest and best established red-light district, with more than 4000 sex workers working in 370 brothels that service approximately 20 000 clients a day. Rao *et al.* (2003) use an instrumental variable strategy to estimate the premium for noncondom sex and correct for unobserved heterogeneity. They use exposure to an HIV/AIDS intervention targeted at sex workers as the instrument because this intervention is highly correlated with a sex worker's propensity to use condoms, but the authors claim it is uncorrelated with a sex worker's income. They estimate that when sex workers use condoms, they earn between 66% and 79% less.

The loss in earnings associated with condom use clearly represents a major disincentive to safe sex. The problem comes from the fact that clients, most of whom do not want to use condoms, are able to exploit competition among sex workers. If a sex worker insists on having sex with a condom, the client can simply go to the next brothel or another sex worker in the same location where he will find a sex worker who is more willing. The solution to this can come either from educating clients about safe sex or creating an agreement among sex workers to collectively agree to refuse condom-free sex (Rao and Shah, 2012).

Conclusion

The scope for research on the sex sector remains large given how little economics work has been done in this area to date. For example, still very little is known about how regulating this market might affect public health outcomes. Different countries apply distinct regulatory techniques to their own sex markets. In some countries sex work is illegal, in other countries it is decriminalized, and in other places it is legal. However, very little is known about how these different regulatory procedures affect things like public health outcomes or the welfare of women involved in the sex industry.

Gertler and Shah (2011) show that regulating the sex market in Ecuador can have some unintended consequences. They find that increasing enforcement of regulation in the street sector significantly decreases STIs. However, increasing enforcement in the brothel sector increases the probability that a sex worker will be infected with an STI. This is because increasing enforcement in the street shifts sex workers on the margin from the more risky street into the less risky brothels, thereby increasing street prices and reducing the overall number of street clients. As a result, overall infection rates fall. In contrast, increasing enforcement in the brothel sector can exacerbate public health problems by inducing some unlicensed brothel sex workers into the riskier street sector. This example and this entire article illustrate that sex workers clearly respond to economic incentives, which is an important lesson for policy implementation.

Savings behavior is another area calling for future research. Unlike most other professions, sex workers earn more at younger ages and their income decreases with age and experience (Arunachalam and Shah, 2008). Because of this inverted earnings to age profile, saving at younger ages should be a priority for these women. However, data show that young sex workers do not save much of their earnings. Whether this is due to lack of access to banks, lack of demand on their part, or some other reason is still unclear. This question is a much needed topic for future research.

See also: HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Infectious Disease Externalities

References

Ahituv, A., Hotz, V. J. and Philipson, Tomas (1996). The responsiveness of the demand for condoms to the local prevalence of AIDS. *Journal of Human Resources* **31**(4), 869–897.

Ahlburg, D. and Jensen, E. (1998). The economics of the commercial sex industry. In Ainsworth, M., Fransen, L. and Over, M. (eds.) *Confronting AIDS: Evidence from the developing world*, pp. 147–173. Brussels: European Commission (for World Bank).

Arunachalam, R. and Shah, M. (2008). Prostitutes and brides? *American Economic Review Papers and Proceedings* **98**(2), 516–522.

Arunachalam, R. and Shah, M. (2013). Compensated for life: Sex work and disease risk. *Journal of Human Resources* **48**, 345–369.

Central Bureau of Statistics (CBS), Ministry of Health, and ORC Macro (2004). *Kenya demographic and health survey 2003*. Calverton, MD: CBS, MOH, and ORC Macro.

Dandona, R., Dandona, L., Anil Kumar, G., et al., and the ASCI FPP Study Team. (2006). Demography and sex work characteristics of female sex workers in India. *BMC International Health and Human Rights* **6**, 5.

Dupas, P. (2011). Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics* **3**(1), 1–36.

Edlund, L. and Evelyn, K. (2002). A theory of prostitution. *Journal of Political Economy* **110**(1), 181–214.

Francis, A. M. (2008). The economics of sexuality: The effect of HIV/AIDS on homosexual behavior in the United States. *Journal of Health Economics* **27**, 675–689.

Geoffard, P.-Y. and Philipson, T. (1996). Rational epidemics and their public control. *International Economic Review* **37**(3), 603–624.

Gertler, P. and Shah, M. (2011). Sex work and infection: What's law enforcement got to do with it? *Journal of Law and Economics* **54**, 811–840.

Gertler, P., Shah, M. and Stefano, B. (2005). Risky business: The market for unprotected commercial sex. *Journal of Political Economy* **113**(3), 518–550.

Giusta, M. D., Di Tommaso, M. L. and Strøm, S. (2004). *Another theory of prostitution. Economics and management discussion papers*. Berks, UK: Henley Business School, Reading University.

Kremer, M. (1996). Integrating behavioral choice into epidemiological models of AIDS. *Quarterly Journal of Economics* **111**(2), 549–573.

Karim, Q. A., Abdool Karim, S. S., Kate, S. and Martin, Z. (1995). Reducing the risk of HIV infection among South African sex workers: Socioeconomic and gender barriers. *American Journal of Public Health* **85**, 1521–1525.

Lau, J. T. F., Tsui, H. Y., Siah, P. C. and Zhang, K. L. (2002). A Study on female sex workers in southern China (Shenzhen): HIV related knowledge, condom use and STD history. *AIDS Care* **14**, 219–233.

Lim, L. L. (1998). *The sex sector: The economics and social bases of prostitution in Southeast Asia*. Geneva: International Labor Organization.

Negroni, M., Bassett Hileman, S., Vargas, G., et al. (2002). Reaching Mobile Populations for AIDS Prevention in Southern Mexico Border Towns. In Poblaciones moviles y VIH/SIDA en Centroamerica, Mexico, y Estados Unidos, Papers presented at the XIV International Conference on AIDS, Barcelona (July). Cuernavaca, Mexico: Instituto de Nacional Salud Publica.

Posner, R. A. (1992). *Sex and reason*. Cambridge, MA: Harvard University Press.

Rao, V., Gupta, I., Lokshin, M. and Jana, S. (2003). Sex workers and the cost of safe sex: The compensating differential for condom use among Calcutta prostitutes. *Journal of Development Economics* **71**, 585–603.

Rao, V. and Shah, M. (2012). Sex work. In Basu, K. (ed.) *Oxford companion to economics in India*. Delhi: Oxford University Press.

Robinson, J. and Yeh, E. (2011). Transactional sex as a response to risk in western Kenya. *American Economic Journal: Applied Economics* **3**(1), 35–64.

Robinson, J. and Yeh, E. (2012). Risk-coping through sexual networks: Evidence from client transfers in Kenya. *Journal of Human Resources* **47**(1), 107–145.

Rosen, S. (1986). The theory of equalizing differences. In Ashenfelter, O. and Layard, R. (eds.) *Handbook of labor economics*, vol. 1. Amsterdam, North-Holland: Elsevier.

Siebert, W. S. and Wei, X. (1998). Wage compensation for job risks: The case of Hong Kong. *Asian Economic Journal* **12**, 171–181.

Townsend, R. (1994). Risk and insurance in village India. *Econometrica* **62**(3), 539–591.

UNAIDS (2002). *Report on the Global HIV/AIDS Epidemic*. Geneva: UNAIDS.

Vandepitte, J., Lyerla, R., Dallabetta, G., et al. (2006). Estimates of the number of female sex workers in different regions of the world. *British Medical Journal* **82**, 18–25.

Viscusi, W. K. (1992). *Evidence on the value of life: Case studies from the labor market. Fatal tradeoffs: Public and private responsibilities for risk*. New York: Oxford University Press.

World Bank (1999). *Confronting AIDS: Public priorities in a global epidemic*. New York: Oxford University Press.

Smoking, Economics of

FA Sloan, Duke University, Durham, NC, USA
SP Shah, John Hopkins University School of Medicine, Baltimore, MD, USA

Glossary

Activities of daily living limitations Limitations in basic activities in which individuals must engage in their daily lives, such as bathing, feed oneself, toileting.
Behavioral economics A branch of economics that incorporates psychological insights into economics.
Commitment devices Various methods individuals use to be sure that they follow through on prior decisions.
Compensatory behavior Behavioral response to a regulated outcome, for example, an increase in the tax on cigarettes increases the amount of the cigarette that is smoked.
Endogenous Arising from within a system, for example, in a system describing health and health behaviors, health may be endogenous to smoking and conversely.
Excise tax A tax on the manufacture or sale of a product.
Forward-looking An individual who considers future benefits and costs in making a decision.
Marginal utility of consumption The gain in utility from an increase in one unit of a (composite) good.
Marginal utility of wealth The gain in utility from an increase in one dollar of wealth.
Over-the-counter Term often used for drugs that are sold without a prescription.
Price elasticity or elasticity of demand Minus the percentage change in the quantity demanded divided by the percentage change in the product's price.
Randomized controlled trial An experiment in which subjects are randomly assigned to a treatment or control group.
Reverse causation A dependent variable influences an explanatory variable and conversely.
Risk averse A willingness to pay positive amounts to reduce risk while not changing the expected outcome, i.e., tastes when the certainty equivalent (the utility review obtaining a given amount for sure) exceeds the expected utility of a gamble, i.e., the sum of products of the probability of each outcome and the value associated with the outcome.
Risk tolerant Not risk averse.
Shifting The increase in price of a good following an increase in the tax imposed on the good.
State Attorneys General The state attorney general in each of the 50 US states and territories is the chief legal advisor to the state government and the state's chief law enforcement officer.
Utility function A function that represents tastes by ranking consumption bundles.

What is the Economics of Smoking and Why Should One Care?

Economists divide markets into demand and supply. Demand reflects individual preferences, relative prices, income, and other factors. Supply differs according to market structure; prices depend on whether the market is competitive or not and the extent of government intervention in the market. On the demand side, decisions about whether or not to smoke are referred to as the extensive margin. Conditional on smoking, the frequency, and intensity of smoking refers to the intensive margin. The decision to smoke is particularly interesting given the addictive properties of cigarettes and the delayed consequences of smoking. Given such consequences, perceptions of risk of future adverse effects, and the discount rate applied to the future cost of smoking have a key role in decision-making. Additionally, government interventions prompt individuals to take externalities into account in their demand decisions.

On the supply side, much emphasis is placed on the cigarette industry's advertising practices. Governments seek to offset such promotion by publicizing the adverse effects of smoking. In the USA, the Master Settlement Agreement (MSA) limits the industry's ability to promote its products and the settlement has increased production costs in the USA which, coupled with substantial increases in excise taxes on cigarettes, has further increased cigarette prices.

The economics of smoking is worthy of attention for several reasons. First, smoking is among the most harmful behaviors to personal health and longevity. Considering the combined cost of internalities and externalities, the societal cost per pack of cigarettes was US$37 in 2004. The main internalities relate to adverse health effects to the smoker. This raises a question about why someone would engage in such behavior. Second, the topic raises issues of fundamental importance to economics, including whether or not people are rational in their decision-making, the accuracy of risk perceptions, preference heterogeneity (e.g. in risk and time preference), and the role of addiction in explaining consumption patterns over time. Third, from a public policy perspective, the track record includes some successes, but effects of many public policies are below initial expectations. Even after decades of policy intervention, in the USA more than 20% of adults continue to smoke. Fourth, even though the share of the US population that smokes has declined, in other countries, especially in lower-income countries, this share continues to increase.

This review will mainly focus on recent economic literature and focus on the USA, where most studies have been conducted. The authors do not discuss important research conducted outside economics.

 doi:10.1016/B978-0-12-375678-7.00313-8

The Section Alternative Frameworks provides an overview of theoretical frameworks by economists and direct empirical tests of the implications of these models. In the Section Direct Empirical Tests of the Alternative Frameworks, other empirical tests of these frameworks are turned to. In the Section Other Empirical Evidence to Explain Continued Smoking, the empirical evidence on continuing smoking is reviewed. Sections Price: Effect of Cigarette Taxes and Effects of Other Cigarette Demand Determinants review findings on price and other determinants of cigarette demand. In Sections Effects of Smoking on Longevity and Health and Other Effects of Smoking, the authors summarize empirical evidence on the consequences of smoking on health, wages, labor force participation, and expenditures on personal health care services. The Section Effects of Policy Interventions evaluates evidence on demand- and supply-side public policies that have the goal of reducing cigarette consumption.

Alternative Frameworks

At the risk of oversimplification, empirical analysis of the economics of smoking is guided by three alternative frameworks: rational addiction, imperfectly rational addiction, and irrational addiction.

Rational Addiction

At least at first glance, addiction seems to be the last place one would find rational behavior. By 'rational,' economists mean that agents use all available information to weigh benefits and costs before making a decision. 'Available' does not mean that the agent has perfect information because the cost of search is nonzero and not trivial. For decisions with consequences over many periods, the rational agent considers downstream and current effects, which are discounted to present value. These implicit calculations reflect subjective beliefs about probabilities of various outcomes and utilities associated with each outcome with choices based on relative expected utilities.

Smoking is addictive and addiction has the special property that consumption in one period affects future marginal utility and, hence, future consumption of the good. The seminal contribution to the theory of addiction using a rational framework is Becker and Murphy (1988). In this model, the agent is a rational, forward-looking individual with stable preferences. Becker and Murphy give precision to the concept of addictive behavior as consisting of these properties: Tolerance – higher levels of past consumption reduce the marginal utility of consumption in the current period; withdrawal – there is disutility associated with cessation in consumption of the good; reinforcement – higher consumption of the addictive good yields higher marginal utility which leads to higher consumption of the good.

In their model, the rational agent maximizes finite lifetime utility with consumption of the addictive good, a composite nonaddictive good, and the stock of the addictive good (how addicted the agent is). The model implies that consumption of the nonaddictive good is set where marginal utility of consumption of the good equals the discounted marginal utility of wealth – a standard result. For the addictive good, the marginal utility of consumption equals the discounted marginal cost of consuming a unit of the good. Hence the agent anticipates adverse future health effects of consuming the addictive good now as well as future prices of the good.

Imperfectly Rational Addiction

In an imperfectly rational addiction framework, some assumptions underlying the rational addiction model are relaxed: consistency of time preferences, and accuracy of risk perceptions about the harms potentially caused by addictive behaviors or the probability that consuming addictive goods will lead to the person becoming addicted. Under hyperbolic discounting the person's discount rate, which applies to both gains and losses, increases as the time for implementing the change becomes closer. The discount rate is constant and exponential (the usual assumption about discounting in economics) for decisions occurring in the future, but is higher for decisions occurring in the present. Thus, if the person considers quitting smoking in 10 years, they use exponential discounting. If the issue is quitting smoking today, they use a higher rate. This has the effect of reducing the present value of the benefit of quitting, i.e., reducing the future cost of adverse health effects. To the extent that the expected benefit is lower, there is less incentive to quit. Because the discount rates depend on the timing of the decision, this behavior is said to be 'time inconsistent.'

Various studies have used this approach and made many modifications to the basic theory, for example, by distinguishing between sophisticated and naïve hyperbolic discounters. The sophisticates know that they will be time inconsistent, so they employ self-control devices, for example, they may not have cigarettes around the house. Also, people may favor smoking bans or higher excise cigarettes as methods of self-control. Others, the naïfs, do not realize that they have time-inconsistent preferences, and so do not consider that at the later time they will discount the benefits of quitting highly, and so not quit early.

Irrational, Cue-Triggered Addiction

In an irrational framework, individuals do not base decisions on objective comparisons of costs and benefits of specific choices, but rather are swayed by emotions. Addictive goods' consumption may result from visceral urges provoked by external cues, which lead to impulsive consumption that would not have occurred in the absence of these cues. More generally, decision-making often follows pattern matching rather than an explicit weighting of benefits and costs. Behaviors that cues elicit are part of the pattern matching process.

Characteristic 'irrational' assumptions are that consumption among addicts is considered by the addicts themselves as a mistake; that environmental cues based on past experiences trigger consumption; and that addicts understand and manage their susceptibility, i.e., they are 'sophisticates'. In one model (Bernheim and Rangel, 2004), a person can be in either a 'cold' or a 'hot' state. When cold, the person is rational. However, when exposed to certain environmental

stimuli, for example, places where the individual has often smoked in the past, the person switches to the hot state in which emotions supersede rational decision-making. The person can control the probability of receiving an environmental stimulus by altering their lifestyle activity. However, the utility from a lifestyle that provides a high probability of a stimulus (e.g. attending parties) is higher than the utility from another activity (e.g. in the extreme, rehabilitation).

Direct Empirical Tests of the Alternative Frameworks

Rational Addiction

Chaloupka (1991) used data on individuals from the second National Health and Nutrition Examination Survey to test the rational addiction model. He regressed daily cigarette consumption on cigarette prices, past, present, and future, and lagged and future cigarette consumption. The price parameter estimates were not statistically significant, providing weak support for the rational addiction model.

Becker *et al*. (1994) used cross sectional data on US states to test the rational addiction model. A key implication of the model is that people are forward-looking in making decisions about consumption of an addictive good. They found that, as cigarette prices rise, current consumption of the good falls, which is what the rational addiction model predicts.

Auld and Grootendorst (2004), using annual aggregate data on milk and the rational addiction model as the underlying conceptual framework, found that milk is the most addictive of all commodities evaluated, including cigarettes! The literature contains other critiques of the models and tests as is, perhaps, to be expected given the relative youth of the topic and its difficult 'fit' with standard economics.

Imperfectly Rational Model

Even if people are not rational and forward-looking, they may be simply forward-looking. Gruber and Köszegi (2001) tested whether cigarette consumption was negatively related to announced cigarette excise tax increases. They found that it was, which they interpreted as evidence for the notion that people are forward-looking but not necessarily time-consistent in their preferences.

There is some empirical evidence on hyperbolic discounting by smokers in particular. Indicators of hyperbolic discounting include: the use of commitment devices, discount rates that vary according to the time horizon with short-terms exceeding longer term ones, and actual behavior not matching stated plans. Odum *et al*. (2002) evaluated discounting on the part of small samples of current, former, and never smokers using two nonlinear decay models, one for exponential and the other for hyperbolic discounting. They found the hyperbolic model provided a better fit between the two. Current smokers discount health gains and losses more than never smokers do using both exponential and hyperbolic functional forms. Several other studies, however, have produced conflicting results, and strong empirical support for a key role of hyperbolic discounting in the smoking decision is not yet available.

Irrational Mode

There is a large literature on the relationship between advertising and smoking. Two reviews are in the Further Reading section. However, the vast majority of studies have used aggregate data, which do permit a direct test of the effect of cues on smoking. There is some evidence that teen exposure to TV advertising has a positive and significant impact on the probability of smoking. However, most direct testing of the effects of cues has been conducted by noneconomists.

Other Empirical Evidence to Explain Continued Smoking

Overview

A fundamental reason why some people smoke and others do not may simply be that their preferences differ. Also, persons who believe that they have a lower life expectancy may be more prone to smoke. Persons smoke because they underestimate the probability of adverse consequences of smoking. Now each of these hypotheses will be examined.

Heterogeneity in Preferences

There is empirical evidence that preferences differ between smokers and others. For example, it seems that willingness to pay to be in good health is considerably higher for never smokers than for current smokers, implying that one reason people continue to smoke is that they value being in good health less.

Smokers select riskier jobs but receive lower risk-adjusted compensation than nonsmokers. This seemingly anomalous result can be predicted from a model in which employers' offers and workers' utility depend on wages and the probability of injury on the job. Smokers and nonsmokers appear to be segmented labor groups having distinctive preferences and distinctive labor market curves. Smokers are more risk tolerant and more impatient. Moreover, heavy smokers tend to be more impatient and less risk averse than never smokers whereas former smokers are more patient and risk averse than never smokers.

Biased Risk Perceptions

Youth risk perceptions are particularly relevant for the initiation of smoking as almost all initiation occurs before age 22, whereas adult risk perceptions are important for what they tell us about cessation. A common assumption appears to be that youths start to smoke because they are overoptimistic about life outcomes.

The empirical evidence reveals a more complex picture. Youths are extremely pessimistic about the probabilities of lung cancer due to smoking or, indeed, dying for any reason by age 20 (Sloan and Platt, 2011).

Overall, for adults, comparing subjective with objective risk, subjective beliefs are quite close on average to their objective counterparts. However, there are differences by smoking status. Even though subjective beliefs about the probability

of dying tend to be higher for current smokers than for never smokers, current smokers tend to be relatively optimistic and never smokers relatively pessimistic in assessments of their own mortality. However, risk perceptions seem not to be important in the decision of adults aged 50–70 to continue smoking. The evidence on relative optimism and pessimism is consistent with a more general finding that low-risk groups tend to overestimate and high-risk groups tend to underestimate their mortality risks. Overestimation of risk reduces the probability that a person will be a current smoker, a result found in several studies from various countries.

Price: Effect of Cigarette Taxes

Another explanation of continued smoking is that although real cigarette prices have risen, cigarette prices may remain too low to deter much smoking. There is a basic distinction between the extensive margin and the intensive margin. The former refers to a decision to smoke or not to smoke and the latter refers to the amount, conditional on whether the person smokes at all.

Extensive Margin

At the extensive margin, the tax responsiveness of youths and adults depends on two separate types of behaviors – initiation and cessation. A higher cigarette tax affects initiation decisions for youths and cessation decisions for adults. The general consensus is that the price of cigarettes in the USA is negatively related to smoking participation. However, the magnitude of these negative price elasticities and whether youth smoking is more sensitive to price than adult smoking has been the center of much recent controversy.

Studies alternatively measure price responsiveness with data on cigarette prices or excise taxes. The effect of a given increase in the excise tax on prices depends on the amount of shifting that occurs. The amount of shifting of an increase in a state excise tax on retail prices of cigarettes might be less near a state border if state B does not follow state A's excise tax increase.

Conventional wisdom suggests that youth cigarette consumption is highly sensitive to price and is greater than price sensitivity for adults. This is not borne out by empirical studies that control for other determinants, or else only weakly. There is little evidence that higher taxes prevent smoking initiation in adolescence but some that taxes influence decisions regarding cessation and at the intensive margin. Initiation decisions may, after all, be driven by noneconomic and unmeasured determinants such as peer acceptance and many studies have not included direct measures of 'smoking sentiment' amongst peer groups or in localities, which may be correlated with taxes and prices.

Carpenter and Cook (2008) used repeated cross-section data from 1991–2005 and controlled for antismoking sentiment, finding that price elasticities for smoking participation range from -0.23 to -0.56. For adults, the general consensus has been that price elasticities for adults fall within the -0.3 to -0.5 range. Higher taxes appear to reduce smoking participation by older adults, especially for those who are less educated and from low-income households. Associated with a $1 increase in the excise tax are participation elasticities ranging from -0.29 to -0.31 for the 45–59 age group, just above -0.2 for persons aged 60–64. These elasticities appear to be lower in other countries (e.g. Russia and China). The price elasticity of smoking varies with other demographic factors, including education and gender. Among Irish women, for example, cigarette taxes have the greatest negative effect on initiation for women with intermediate levels of education. For cessation, cigarette taxes have the greatest effect for women with the lowest level of education. However, the pathway through which educational attainment affects the propensity to smoke remains unidentified. Although there is no established theoretical reason that would explain differences in smoking by gender, historically men have had higher rates of smoking than women but the gender gap has narrowed in recent decades. Research findings on the impact of cigarette prices on smoking participation by gender are mixed. It seems that women are nearly twice as responsive to cigarette taxes as men.

Intensive Margin

At the intensive margin, an increase in the cigarette tax results in a decrease in the number of cigarettes consumed daily; however, one study showed that smokers often compensate either by extracting more tar and nicotine from each cigarette (Adda and Cornaglia, 2006) or by shifting to a cigarette brand with more tar and nicotine content (Farrelly *et al.*, 2004) with the result that tar and nicotine consumption is not reduced and for one group (18–20) appears to have increased. More recently, Abrevaya and Puzzello (2012) reexamined Adda and Cornaglia's (2006) evidence on the compensatory behavior of smokers who, facing higher taxes, reduced cigarette consumption although maintaining their cotinine (a biomarker for nicotine) levels. They used (1) appropriate clustered standard errors, (2) a larger sample from the same years and survey than the data in Adda and Cornaglia's (2006) analysis, (3) cigarette-prices instead of and in addition to cigarette-taxes, and (4) sampling weights. Abrevaya and Puzzello (2012) found that the Adda and Cornaglia (2006) results were not robust. They find little empirical support for compensatory behavior found in subsamples of smokers. Stehr (2007) reported elasticities of intensity for adult men and women of -0.09 and -0.12, respectively, which implies that most of the effect of an increase in the excise tax is from a reduction in the fraction of adults who smoke.

Effect of Cigarette Prices on Smuggling

Cigarette price differentials cause changes in buying patterns, both across geographic areas, such as US states, and across selling modes, such as internet sales versus sales from bricks and mortar stores. Two studies focus on tax avoidance through internet purchases of cigarettes. Lower cigarette prices for cigarettes obtained over the internet have two major potential effects. First, they may increase aggregate cigarette consumption. Second, they may shift purchases from other retailers to internet vendors. Internet penetration increases the

negative effect of an excise tax increase on taxable cigarette packs per capita sold.

Effects of Other Cigarette Demand Determinants

Educational Attainment

Two important stylized facts are pertinent for describing the relationship between educational attainment and smoking. First, more highly educated persons tend to be healthier than others and, second, on average more highly educated persons smoke less. Although the associations are indisputable, there is controversy about causation and, if there is causality, the magnitude of the effect. Moreover, the relative importance is not clear of the various pathways through which educational attainment influences smoking.

Although the correlation between education, health behaviors, and health is well established, the notion that additional schooling affects health behaviors and causes health improvements is not. Reverse causality from health in general, and health behaviors in particular, to years of schooling completed, and/or the presence of third factors correlated with schooling and smoking but omitted from the analysis may cause educational attainment, health behaviors, and health to vary in the same direction. In high-income countries, but not necessarily in other countries, there is little reason to expect reverse causation from health to years of schooling because generally the temporal lag between the completion of formal education and the time at which health declines is several decades.

Possible omitted third variables is a more likely source of bias. Among these variables are: native ability, time preference, including at the time schooling choices are made, genetic factors, poor health in early life which may be positively correlated with poor health in adulthood (and which may also affect educational attainment in early, low income in early life which similarly may affect educational attainment and, independently, health in later life (Jayachandran and Lleras-Muney, 2009).

One's ability to make valid causal inferences depends on the quality of the identification strategy. Various strategies have been used. For example, de Walque (2010) constructed panels based on smoking histories, finding that among women, college education has a negative influence on the probability of smoking and more educated persons' smoking responded more quickly to diffusion of information on smoking. He offered explanations for differences by gender. Other identification strategies were based on availability of high school and college openings, draft avoidance during the Vietnam War and abolition of secondary school fees. In general, studies find causal effects, but there are exceptions. Most empirical evidence on effects of educational attainment comes from the USA and other high-income countries. However, empirical evidence from other countries supports the finding of a negative effect of educational attainment on smoking.

Peers

Peer effects have different potential roles depending on the stage of the life cycle. For adolescents, peers may encourage smoking. In the past, in adult life, smoking may have been seen as useful in promoting business or social interactions. Also, spouses may influence a person's smoking patterns. Isolating peer effects is difficult in particular because choice of peers is endogenous. People who enjoy being around smokers are likely to associate with smokers and conversely for persons who suffer from being around smokers. The conventional wisdom is that peer effects are important in influencing adolescents to start smoking.

Powell *et al.* (2005) analyzed peer effects on smoking. Peer effects were measured as behavior of all students in a school less the student in question. The instrumental variables were characteristics of other students in the school. They found that peer effects have an important role in influencing individual adolescents to smoke. In particular, moving from a school in which no students smoke to one in which a quarter of the students smoke increases the probability that a youth smokes by 14.5% points. Although the specific estimates vary, virtually all studies in the USA and elsewhere, have found positive effects of peer group smoking.

One type of peer effect reflects interactions of siblings and spouses within a household, where the expectation is that the effect would be strong. The relationship between decisions about smoking of spouses may reflect several underlying influences: correlation due to matching in the marriage market; bargaining within marriage; and social learning.

Health Shocks

Surviving a major health shock, such as a heart attack, may affect continued smoking for at least two reasons. First, the shock may reveal inherited susceptibility, i.e., new information about the personal effects of a life style including smoking. Second, the person may seek to forestall further health damage to self. However, the health shock may serve to increase unhealthy behaviors if it leads to thinking that they will gain little or no benefit from cessation given that they do not have much time to live. The onset of smoking-related health shocks other than cancer, nonsmoking related health shocks, onset of activities of daily living limitations, and onset of fair/poor health all increase the probability of smoking cessation. Smoking-related health shocks may generate new information on the effects of smoking to the individual. But the nonsmoking-related health shocks must be affecting smoking though another mechanism.

Health shocks to spouses as well as their smoking behavior may influence an individual's smoking decision. Among never smokers, but not for current and former smokers, spousal smoking has a negative effect on a person's longevity expectations. Spousal smoking-related health shocks also reduce longevity expectations of never smokers for reasons that are not understood because they have no effect on such expectations for current smokers. This might occur through three channels: Consumption externalities – one spouse's welfare affects the other spouse's welfare; altruism – one spouse reduces smoking in response to the other spouse's bad health; and learning about risks of smoking from the health experience of the other spouse. There is some evidence suggestive that consumption externalities are at work.

Stress

Even though there is a substantial amount of noneconomic literature on the effect of stress on smoking, the economic literature on the topic is quite limited, and only some results support an effect of stress on smoking. Job-related stress seems to increase smoking at the extensive margin, but the relationship is not generally statistically significant at the intensive margin. Death of a parent within the previous 2 years increases the probability of continuing to smoke. Without fixed effects, being separated, divorced, widowed in the past 2 years leads to continued smoking, but with fixed effects included, these relationships disappear.

Effects of Smoking on Longevity and Health

The effect of smoking on health is well documented in the epidemiologic literature. Economists have made some contributions as well showing that smoking is very harmful to personal health. The authors review three types of economic studies here – effects of smoking on: the smoker's health, nonsmoker's health, and neonate's health.

Between one third and three quarters of excess US veteran deaths are attributed to heart disease and lung cancer caused my military-subsized smoking. Smoking significantly contributes to inequality by income in predicted mortality.

A mother's decision to smoke involves balancing the utility of smoking against the disutility of adversely affecting the health of her child. Agee and Crocker (2007) assumed that a mother has three arguments in her utility function – her consumption, her health, and her child's health. Using data from the National Maternal and Infant Health Survey, they found that a mother values the health of her child more than her own health. The authors conclude that antismoking messages should mention health benefits to children from mothers not smoking.

Smoking is particularly harmful when done by pregnant women. The effect of smoking on birth weight is in the range of negative 100–150 g.

Compared to the 1950s when epidemiological studies on the health harms of smoking were just beginning to appear, people have become far more knowledgeable about the adverse effects of smoking. Thus, women who smoke during pregnancy in recent years are a much more select group than women who smoked while pregnant in the mid-twentieth century. Using data from three sources, the National Child Development Study, the British Cohort Study, and the Millennium Cohort Study, approximately half of the reported effects of smoking on the probability of low birth weight is due to unobserved maternal characteristics that are correlated with prenatal smoking.

Other Effects of Smoking

Wages

Empirical evidence indicates that smokers have lower wages than nonsmokers. The issue is not whether or not there is a difference in wage rates by smoking status but rather whether or not the relationship from smoking to wages is causal. In particular, smoking may be systematically related to omitted factors that also affect worker productivity.

Expected lifetime contributions to Social Security are US$3800 lower for smokers than never smokers among men and approximately US$200 lower among women (US$2000). A study using Canadian General Social Survey data to determine the effect of smoking on wages found a loss in earnings due to daily smoking of 24%. One reason why Social Security contributions are lower for smokers is that they may have been out of the labor force for longer periods. For example, smoking affects labor market participation indirectly through its effect on diabetes (through its effect on blood sugar levels and insulin resistance) and cardiovascular disease onset.

Expenditures on Personal Health Care

Although smokers tend to be sicker on average, they also live shorter lives. The fewer years of exposure to health expenditures may offset higher expenditures smokers incur per year that they are alive. However, it is probably correct to conclude that, overall, smoking raises expenditures per smoker, though by a trivial amount. Approaches which evaluate smoking-related expenditures at a point in time rather than over lifetimes yield much higher estimates of smoking cost (Sloan *et al.*, 2004).

Effects of Policy Interventions

Advertising

Cigarettes are one of the most advertised and promoted products in the USA, at least historically, in spite of the fact that government-imposed limits on such advertising also have a long history, beginning with a ban on advertising of cigarettes on television and radio implemented in 1970. The focus here is on effects of cigarette advertising bans and of promotion of smoking cessation products.

Heckman *et al.* (2008) criticize much past research that infers advertising has a causal influence on smoking initiation. Specifically referring to the Cochrane Review studies, but applying more generally as well, the authors have three specific criticisms. First, the studies do not develop adequate models on which to base the empirical analysis. Second, they do not adequately account for endogeneity of advertising exposure. Third, there may be insufficient variation in advertising exposure to generate statistical differences in responses.

Overall, evidence on the effects of advertising bans is mixed. Advertising restrictions mainly influence the smoking rate through their impact on concentration of the cigarette product market. With fewer sellers, people smoke less. The effects of both limited and comprehensive bans seem to be greater in developing countries than in developed countries.

Another strategy to reduce cigarette consumption is to promote products that reduce smoking behavior. Avery *et al.* (2007) investigated a policy in which the Food and Drug Administration (FDA) allowed all cessation products to be sold over-the-counter (OTC) immediately. They projected that

this policy would increase advertising of smoking cessation products by 80%. In a second simulation, they assessed the effect of offering each product available OTC a year earlier. They found that this change would increase advertising of smoking cessation products by 9%. As the number of competitors in this market increases initially, the effect on advertising is positive. However, the effect of adding competitors diminishes with entry of new sellers and eventually becomes negative.

Smoking Bans

Smoking bans may have desirable intentions in terms of reducing smoking rates and environmental tobacco smoke, but whether or not in fact they do reduce them is an empirical question. The evidence from the literature on the effectiveness of smoking bans is mixed.

Local laws restricting workplace smoking in Ontario, Canada reduced environmental tobacco smoke for blue-collar workers but not for other workers. Smoking bans in Norway have made smokers more considerate of others in areas where smoking is not banned but it has not been established that the smoking bans cause changes in social norms.

Poutvaara and Siemers (2008) theorized about the role of social norms in smoking. If the social norm is that smokers may smoke at will they argued there would be too much smoking – nonsmokers are hesitant to ask smokers not to smoke in their presence even though the disutility of the other person's smoking exceeds the utility gain from the social interaction. In this type of situation, smoking bans may represent a second-best policy. Others have found that bans reduce smoking among persons who frequent bars and restaurants but not among the population overall.

There is also research on unintended adverse side effects of smoking bans. One such effect may be that people travel further to get to locations where they can both smoke and consume alcoholic beverages, which may increase the prevalence of drunk driving and some evidence supports this view. Another perverse effect may arise in the form of increasing exposure of nonsmoking family members because bans in public places increase smoking at home.

Master Settlement Agreement

The MSA reached between 46 state attorney generals and the four major cigarette manufacturers in November 1998 represents the largest single public intervention in tobacco control in US history (Sloan and Chepke, 2011). The MSA settled numerous lawsuits filed by individual states against cigarette manufacturers, which alleged that the manufactures had promoted smoking, thereby increasing the smoking rates in the USA, which in turn increased medical costs incurred by the states, mainly through Medicaid.

The MSA included an assessment on cigarette companies that resulted in a substantial increase in cigarette price. The price increase and MSA antismoking provisions reduced smoking rates by 13% for persons aged 18–20 and 65+ and by 5% for others. For the first 15 months after MSA implementation, the effect of the MSA is estimated to have reduced prenatal smoking by 2.4%. The MSA appears to have led states to increase the excise tax on cigarettes, presumably because the MSA weakened cigarette manufacturers' political power in opposing such increases.

Another explanation for the increase in state cigarette excise taxes post-MSA is that publicity immediately before and after the MSA was implemented affected voter preferences about the cigarette companies and issues related to tobacco control policies with the result that voters were more favorable to such policies after the MSA was implemented.

Although the MSA imposed costs and various restrictions on large cigarette manufacturers, it also reduced the uncertainty about outcomes of legal disputes that existed before MSA implementation. The MSA led to a decrease in manufacturers' cost of capital by at least 2.2%. Overall, the MSA has been a mixed success. Cigarette price increases can be accomplished more efficiently by excise tax increases, which do not involve the high legal expense of litigation. Moreover, the MSA was a 'cash cow' for states rather than a source of revenue dedicated to public programs to reduce smoking and improve population health. Of course, there is no guarantee that additional excise tax revenue from cigarettes would not be spent in the same way.

Behavioral Economics Solutions

As indicated in the Section Alternative Frameworks, some aspects of smoking are consistent with predictions of behavioral economics. There is a limited amount of research on policy interventions whose designs reflect insights of behavioral economics. An experimental product (CARES) offers smokers an opportunity to invest in a savings account in which they deposit funds for 6 months. After this, if they pass a urine test for cotinine and nicotine, they are returned the money. Otherwise, the money goes to charity. Although the results suggest that the program was effective in inducing successful quitting, the possibility remains that participants were more motivated to quit. Further, as the authors acknowledge, only a minority of smokers successfully quit.

Volpp *et al.* (2009) evaluated a randomized controlled trial of a smoking cessation intervention at one firm. 878 employees were randomly assigned to two groups: (1) a group that only received information about benefits of smoking cessation; (2) the other group received this information plus financial incentives to stop smoking. There was an immediate payment of US$100 for completing the education program, US$250 for cessation within 6 months, and another US$400 for abstaining from smoking for an additional 6 months. An insight of behavioral economics is that immediate payments and implementation of self-control devices are important motivators for behavioral change. The latter group had significantly higher rates of smoking cessation than did the control group – 14.7% versus 5% quitting after 9–12 months following enrollment, which was somewhat lower than the quit rate at 6 months.

In spite of the amount of economic research that has been conducted, several controversies remain, including the magnitude of effects of cigarette prices and taxes on smoking initiation and youth smoking more generally and the causal

effect of educational attainment on smoking. The theory of rational addiction made an important contribution in showing that smoking and other addictive behaviors may reflect an explicit choice by informed individuals. A body of empirical research supports some of the model's predictions.

However, the introduction of psychological concepts into economics by behavioral economics has helped explain important stylized facts about smoking decisions that are not well explained by the fully rational framework. Empirical tests of the imperfectly rational and irrational frameworks are still in their infancy. Furthermore, it is one thing to state that people derive utility from such behaviors as smoking; but what really motivates people to smoke in a point in time, given that they know that smoking is very bad for personal health? Certainly virtually everyone knows that smoking is costly in the long run, but many people smoke and continue to smoke.

In the end, what makes smoking interesting to study is not only its relevance to human health, but also that understanding smoking behavior requires us to draw on a variety of disciplines as well as fields within economics.

See also: Addiction. Alcohol. Education and Health: Disentangling Causal Relationships from Associations. Education and Health. Illegal Drug Use, Health Effects of. Medical Decision Making and Demand. Multiattribute Utility Instruments and Their Use

References

Abrevaya, J. and Puzzello, L. (2012). Taxes, Cigarette consumption, and smoking intensity: Comment. *American Economic Review* **102**(4), 1751–1763.

Adda, J. and Cornaglia, F. (2006). Taxes, cigarette consumption, and smoking intensity. *American Economic Review* **96**(4), 1013–1028.

Agee, M. D. and Crocker, T. D. (2007). Children's health benefits of reducing environmental tobacco smoke exposure: Evidence from parents who smoke. *Empirical Economics* **32**(1), 217–237.

Auld, M. C. and Grootendorst, P. (2004). An empirical analysis of milk addiction. *Journal of Health Economics* **23**(6), 1117–1133.

Avery, R., Kenkel, D., Lillard, D. and Mathios, A. (2007). Regulating advertisements: The case of smoking cessation products. *Journal of Regulatory Economics* **31**(2), 185–208.

Becker, G. S., Grossman, M. and Murphy, K. M. (1994). An empirical analysis of cigarette addiction. *American Economic Review* **84**(3), 396–418.

Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy* **96**(4), 675–700.

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review* **94**(5), 1558–1590.

Carpenter, C. and Cook, P. J. (2008). Cigarette taxes and youth smoking: New evidence from national, state, and local youth risk behavior surveys. *Journal of Health Economics* **27**(2), 287–299.

Chaloupka, F. (1991). Rational addictive behavior and cigarette smoking. *Journal of Political Economy* **99**(4), 722–742.

Farrelly, M. C., Nimsch, C. T., Hyland, A. and Cummings, M. (2004). The effects of higher cigarette prices on tar and nicotine consumption in a cohort of adult smokers. *Health Economics* **13**(1), 49–58.

Gruber, J. and Köszegi, B. (2001). Is addiction "rational"? Theory and evidence. *Quarterly Journal of Economics* **116**(4), 1261–1303.

Heckman, J. J., Flyer, F. and Loughlin, C. (2008). An assessment of causal inference in smoking initiation research and a framework for future research. *Economic Inquiry* **46**(1), 37–44.

Jayachandran, S. and Lleras-Muney, A. (2009). Life expectancy and human capital investments: Evidence from maternal mortality declines. *Quarterly Journal of Economics* **124**(1), 349–397.

Odum, A. L., Madden, G. J. and Bickel, W. K. (2002). Discounting of delayed health gains and losses by current, never- and ex-smokers of cigarettes. *Nicotine & Tobacco Research* **4**(3), 295–303.

Poutvaara, P. and Siemers, L. H. R. (2008). Smoking and social interaction. *Journal of Health Economics* **27**(6), 1503–1515.

Powell, L. M., Tauras, J. A. and Ross, H. (2005). The importance of peer effects, cigarette prices and tobacco control policies for youth smoking behavior. *Journal of Health Economics* **24**(5), 950–968.

Sloan, F. and Chepke, L. (2011). Litigation, settlement, and the public welfare: Lessons learned from the. *Widener Law Review* **17**, 159–226.

Sloan, F. and Platt, A. (2011). Information, risk perceptions, and smoking choices of youth. *Journal of Risk and Uncertainty* **42**(2), 161–193.

Sloan, F. A., Ostermann, J., Conover, C., Taylor, Jr., D. H. and Picone, G. (2004). *The price of smoking*. Cambridge, MA: MIT Press.

Stehr, M. (2007). The effect of cigarette taxes on smoking among men and women. *Health Economics* **16**(12), 1333–1343.

Volpp, K. G., Troxel, A. B., Pauly, M. V., et al. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *New England Journal of Medicine* **360**(7), 699–709.

de Walque, D. (2010). Education, information, and smoking decisions evidence from smoking histories in the United States, 1940–2000. *Journal of Human Resources* **45**(3), 682–717.

Social Health Insurance – Theory and Evidence

F Breyer, University of Konstanz, Konstanz, Germany

The Concept of Social Health Insurance

Unlike private health insurance (PHI), 'social' health insurance (SHI) is characterized by three distinguishing features:

- Compulsory membership, at least for the great majority of the population.
- Community rating, i.e., premiums unrelated to individual risk.
- Open enrollment, i.e., even if the insurance market is competitively structured, an applicant cannot be denied coverage by an insurer.

The theory of SHI has two distinct branches. In the normative branch, the main questions are: (1) What are the efficiency and equity reasons for introducing and maintaining such a compulsory institution in a market economy; (2) How should SHI be designed so as to optimally attain the respective goals, both with respect to the benefits covered and to the way how it is financed; and (3) How should a competitive market for SHI be regulated to achieve its targets? The positive branch explains why SHI exists in most democracies and why certain observable features are characteristic.

Normative Theories of Social Health Insurance

Possible Efficiency Reasons for Social Health Insurance

According to neoclassical welfare economics, the violation of one or more assumptions of the First Welfare Theorem (Mas-Colell *et al.*, 1995) in a particular market is necessary (but by no means sufficient) (for sufficiency, it must be shown that government can create an institution which is suitable to bring about a better allocation) to justify government intervention in this market on efficiency grounds. In the case of SHI, the assumptions typically alluded to are market transparency (which may be violated in the cases discussed in the Sections Asymmetric information in the health insurance market and Insurance against reclassification risk) and independent preferences (for violations see Sections Altruism and free riding and Externalities from medical care).

Asymmetric information in the health insurance market

If competitive insurance markets are best described by the assumptions of the Rothschild and Stiglitz (RS) model, then asymmetric information of the 'hidden information' type (i.e., absence of transparency) may lead to an inefficient market equilibrium. If the insured has more precise information on his individual risk distribution than the insurer, the only possible RS equilibrium is a separating one (a separating equilibrium is an equilibrium in which each risk type is offered a contract which is not bought by other risk types.) in which only the highest risk types are offered complete coverage at actuarially fair premiums. Lower risks obtain more favorable terms but are rationed in terms of coverage. They would prefer to have more but this would make their contract attractive to unfavorable risks. Relative to such an equilibrium, SHI which forces all individuals into a pooling contract with partial coverage can achieve a Pareto improvement: high risks are made better off because they pay lower premiums for the mandated part of their coverage, whereas low risks benefit from improved total (social plus private) coverage (Newhouse, 1996).

However, several objections can be raised against this defense of SHI: first, it is unclear to what extent asymmetric information on health risks is really a problem as medical examinations can be used and are used to determine the risk of an insured. Second, the assumptions of the RS model are highly unrealistic: firms can neither cross-subsidize one insurance plan by another nor anticipate their competitors' reaction to their own market entry. If these possibilities are taken into account, then the inefficiency of the competitive market equilibrium vanishes and so does the justification of SHI.

Insurance against reclassification risk

In contrast to the adverse selection by Rothschild and Stiglitz (1976) model mentioned in Section Asymmetric information in the health insurance market, private health insurers charge different premiums according to health status. As this may change over time in an unpredictable way, individuals face the risk of uncertain premiums ('reclassification risk'). As before, the existence of a problem does not per se justify government intervention. Indeed, private insurers may cover this risk in two ways. First, they can offer 'guaranteed renewable contracts' that provide a premium guarantee to individuals in exchange for a prepayment (Pauly *et al.*, 1995). Second, Cochrane (1995) proposed to insure premium risk by a separate insurance that pays an indemnity to individuals who become a high risk ('premium insurance'). However, both of these market solutions suffer from problems. Guaranteed renewable contracts lock consumers in, which means that they will be unable to switch to another insurer in case they are dissatisfied with the service. (Thus, the case for government intervention ultimately rests on the assumption of nontransparent product quality.) Premium insurance contracts are likely to be incomplete because it is difficult to define the risk type with sufficient precision. Thus, SHI with community rating may be the only appropriate solution for the problem of reclassification risk.

Altruism and free riding

Altruistic rich members of society may be willing to subsidize the provision of healthcare to the poor if they are more interested in the health than in the subjective well-being of the poor – a case of interdependent preferences. Private charity is not suitable to achieve an efficient allocation as donations to the poor, whether in cash or in kind, have a public-good

 doi:10.1016/B978-0-12-375678-7.00907-X

characteristic because they increase the utility not only of the donor but also of other altruistic members of society. The free rider problem of potential donors could be solved either by a tax-financed National Health Service or a specific system with free healthcare for the poor (such as Medicaid) or an SHI with compulsory membership and contributions according to the ability to pay.

Externalities from medical care

Besides the 'psychological' externalities described in Section Altruism and free riding, there are physical externalities involved in some types of medical care, in particular in the treatment and isolation of patients with contagious diseases as well as in vaccination services. However, given the limited extent of infectious diseases, it is questionable if these effects are a sufficient rationale for SHI or if there is a weaker interference in free markets such as the subsidization of vaccines.

Optimal taxation when health and income are correlated

A related justification of SHI is derived from the theory of optimal taxation. If abilities cannot be observed by tax authorities, the extent to which income taxation can be used for redistribution from the high skilled to the low skilled is limited because the high skilled can always 'mimic' the low skilled by reducing their labor supply. However, if there is a negative correlation between ability and the risk of illness, a mandatory SHI with community rating implicitly redistributes between the ability groups in the desired fashion and thus improves social welfare. It must be emphasized, however, that this justification departs from Paretian welfare economics by postulating a specific redistributive goal.

Social Health Insurance and Equity

A further and perhaps the most compelling justification, also known as the 'principle of solidarity' relates to the achievement of equality of opportunity: people differ in their health risk already at birth, and some indicators of risk are readily observable. Moreover, with the rapid progress of genetic diagnostics and the spread of tests during pregnancy, the ability to measure individual health risks of newborns will become more and more pronounced. In PHI, these differences in risk immediately translate into differences in premiums so that those persons who are endowed by nature with a lower stock of 'health capital,' and are thus already disadvantaged, have to pay a higher price for the same coverage on top of this. Behind the veil of ignorance, one would desire at least an equalization of the monetary costs of illness.

There are in principle two ways to achieve solidarity in health insurance. First, PHI premiums can be subsidized for those who would have to pay excessive contributions. The transfer could be on a current basis or a lump sum, equal to the estimated present value of future excess premiums over the whole expected lifespan of beneficiaries. Both have the important advantage of permitting full competition in PHI (or SHI), including insurers acquiring information about true risk. Besides means testing and the need to define a benchmark contract to determine the amount of the subsidy, the second variant has the disadvantage of shifting the risk of longevity to beneficiaries. The second alternative is a compulsory SHI scheme with open enrollment and community rating that prevents differences in health risk from being translated into differences in contributions but, if combined with a competitive structure, induces cream skimming and therefore requires risk adjustment schemes (RAS; see Section Competition in Social Health Insurance) as a secondary neutralizing regulation.

The Design of the Benefit Package of Social Health Insurance

The efficiency reasons given above for the existence of SHI with compulsory membership can be convincing only if the design of the SHI contract is in some sense 'optimal' from the point of view of some 'representative' consumer. The most important design feature is the depth of coverage or, more precisely, the use of copayment provisions in SHI design. What are the main reasons justifying deviations from full coverage?

Administrative costs

Copayment provisions can be called for to keep administrative costs low such as costs of handling claims. For this reason, and assuming expected utility maximization on the part of consumers, it is optimal to exclude partially or entirely expenditures on healthcare items that occur frequently but in limited amounts such as minor medications. More specifically, if administrative costs are proportional to the expected volume of health expenditures, a feature of the optimal insurance contract is a fixed deductible, which serves to equalize marginal utility of disposable income in all insured states of the world. Only in the absence of administrative costs would the optimal deductible be zero. However, in some SHI systems such as the German one, doctors and hospitals are paid directly by the sickness funds and the payment is only weakly related to the volume of services provided so that the handling of individual claims from the insured person by the sickness fund is not necessary and thus this reason is irrelevant.

Noninsurable loss

Illness typically involves not only monetary costs but also nonmonetary losses such as pain and suffering. Optimal health insurance equalizes marginal utility of wealth in all states of nature but this is not equivalent to full coverage if there are complementarities between nonmonetary and monetary losses. In particular, if marginal utility of wealth is lower in case of illness than in good health (e.g., due to reduced ability to enjoy expensive types of consumption), optimal health insurance does not fully reimburse the monetary loss. Although some papers find that marginal utility of consumption is higher in case of sickness, the bulk of the evidence points in the opposite direction, thus supporting the use of copayments for this reason.

Ex ante *moral hazard*

If the insurer cannot observe preventive effort on the part of the insured, a high degree of coverage reduces the incentive for prevention. Hence, there is a trade-off between risk spreading through insurance and maintaining incentives to keep the risk

of illness low. This trade-off leads to a premium function which is convex in the degree of coverage, such that full coverage should be particularly expensive. In SHI such a premium function is nowhere observed, although it could be easily administered because consumers cannot circumvent the convex schedule by purchasing many insurance contracts with limited coverage and low premiums. One reason may be that this type of moral hazard is small due to the nonmonetary costs of illness. Moreover, empirical evidence suggests that people with health insurance live healthier lifestyles, although this may be due to a selection effect. Finally, a system of taxes on harmful and subsidies on healthy consumption goods may be a better alternative (Arnott and Stiglitz, 1986).

A different reasoning applies for prevention through medical services, whose costs can themselves be included in an insurance contract although their occurrence is not a random event. Ellis and Manning (2007) showed that it is efficient to include at least partial reimbursement for preventive services in SHI coverage to align privately optimal demand for these services with the social optimum, which considers the effect of prevention on the premium. In particular, the coverage rate for preventive services should be higher the lower the coinsurance is on treatment (so insurance for treatment and prevention are complementary) and the more risk-averse consumers are.

Ex post *moral hazard*

If the insurer could observe the health status of the insured, the optimal type of health insurance would provide indemnity payments, i.e., the insurance payment would not depend on the insured's healthcare expenditure. With asymmetric information, however, linking reimbursement to expenditure is inevitable. Still, copayment provisions are needed to fend against overconsumption of medical care. The optimal copayment rate is higher the more price elastic is the demand for the particular type of medical services. Empirical evidence, for example, from the research and development health insurance study, shows that there is a small, albeit statistically significant, price elasticity of demand for most medical services.

Competition in Social Health Insurance

In an unregulated PHI market, high risks pay higher premiums for the same level of coverage than do low risks. Community rating in SHI prevents this, making 'cream skimming' attractive, which in turn runs counter to the aim of open enrollment.

Risk selection can take different forms. Health insurers perform direct risk selection if they influence directly who signs a contract. For example, insurers may 'lose' the contract form handed in by a person deemed expensive. Individuals who can be expected to be profitable for the insurer can be encouraged to sign a contract by offering them supplementary services at a discount or, in the extreme case, outright payments.

Indirect risk selection, however, consists in designing benefit packages or by contracting with service providers who are attractive for low risks but unattractive for high risks. In particular, insurers may design their benefit package to attract low but not high risk. An example is a contract with a deductible. This is more appealing for low than for high risks as they face a lower probability of becoming ill and, therefore, of having to pay the deductible. The same reasoning applies to the design of the benefit package in general. For instance, an insurer who covers only few services for patients suffering from diabetes can expect these high risks to prefer another insurer. A straightforward counterstrategy is to impose a maximum deductible and a minimum benefit package. This may not be sufficient, however, because insurers can still try to attract low risks by writing policies with ample coverage of athletic medicine and well-baby care. If these benefits are included in the mandatory package, they will also have to be financed by high risks who have no interest in them (Kifmann, 2002). It may therefore be necessary to specify a maximum benefit package as well.

There are a number of options for complementary regulation designed to limit risk selection. The first is a central fund running an RAS, which pays to the insurer the difference between the expected healthcare cost of the insured and of the average of the respective population. *Ex ante* equalization of expected healthcare expenditures has the crucial advantage of preserving incentives for cost control but is restricted by the availability of data needed to determine payments from the fund. An alternative are cost-reimbursement schemes. These can be based on total cost, costs by service type, and individual healthcare expenditure. In the latter case, the individuals whose healthcare expenditure is reimbursed can be determined prospectively or retrospectively. Various functional forms of reimbursement can be employed. For example, regulation may prescribe a high-risk pool, in which expenditures of the x% most expensive insured (which are identified on the basis of past experience) are covered by the central fund (Van Barneveld *et al.*, 1996). By contrast, Kifmann and Lorenz (2011) found with data of a Swiss health insurer risk selection can most effectively be prevented if costs are reimbursed only up to a limit.

Income-Related versus Flat Contributions

A second feature of SHI which is under scrutiny is the base on which contributions are levied, where the choice is:

1. between income-related and flat contributions, and, in the first case;
2. on which types of income should be included:
   a. only earnings, or
   b. income from all sources,

and whether or not an income ceiling shall apply.

Presently, all countries with an SHI system except Switzerland levy contributions only on the basis of labor income. (In the Netherlands, the employer's share of 50% is levied on labor earnings, whereas the employees pay a flat fee.) Historically, this was an application of the principle of equivalence between contributions and benefits as long as the majority of benefits consisted of income replacement in times of sickness. In the meantime, with the rise in scientific medicine throughout the twentieth century the percentage of income replacement in total health insurance expenditures has dropped to the single digits so that this justification is no longer valid.

Nowadays, health insurance benefits are virtually the same for all members (except for differences in risk discussed in Section Social Health Insurance and Equity), and thus income-related contributions constitute pure income redistribution between high and low earners. A possible justification would be the principle of ability to pay. But this is very imperfectly measured if only labor income is taken into account, in particular with the declining share of labor earnings in national income in recent years. Moreover, wage-related contributions imply a significant additional tax on labor, and therefore distort the labor-leisure choice as well as the decision whether to work in the official sector or in the shadow economy. These important disadvantages have to be weighed against a single argument in favor of this particular contribution base, viz. the low costs of collecting the contributions at the source, i.e., as a payroll tax from the employer.

Besides the efficiency reasons just mentioned, there are additional arguments for uncoupling health insurance contributions from income. In particular, the decision making on copayments and other features of the benefit package is distorted if contributions are differentiated according to income: low-income persons have an incentive to opt for 'too much' insurance coverage whereas the opposite is true for high-income voters.

The only European country with flat contributions is Switzerland, whereas the Netherlands have a mixed system in which the employer's contribution is wage related and the employee's contribution is flat. However, even in Switzerland, effective contributions are not completely independent of income from the point of view of the insured, because low-income households receive a so-called 'premium subsidy.' This subsidy varies from canton to canton and covers that part of the total premium of household members which exceeds x% of household income (where x is usually between 7 and 10). Thus effectively, the total contribution amounts to x% of income up to an income ceiling that depends on the number and age of household members and can be calculated as total contribution divided by x. Thus, this system has the same effects like an income-dependent contribution in which total income from all sources is taken into account.

It is sometimes argued that the volume of premium subsidies, which have to be financed from general taxation, will place an enormous stress on the government budget and will therefore be a constant matter of political debate.

Therefore, a better and politically more stable method for compensating the losers from a transition to flat contributions could be a general reform of the tax transfer system, in which social assistance transfers and child allowances are increased and the income tax schedule is appropriately changed (higher initial tax exemption and higher marginal tax rates) so that the pure income redistribution, which is now implicit in the system of health insurance contributions, is performed within the general budget but now with a broader tax base.

Positive Theory: The Political Economy of Social Health Insurance

Having discussed normative theories of the design of SHI, it is important to assess what can be expected from political decisions in democracies. In particular, the theory should explain:

1. the apparent lack of generosity of the benefit package of SHI, characterized by the massive use of nonprice rationing methods and the simultaneous presence of private financing of supplementary healthcare services; and
2. the widespread phenomenon of financing SHI with income-dependent contributions (or taxes) and no risk-rated premiums, in which there is a twofold redistribution from the low to the high risks and from high to low earners.

As to the first point, several models have been put forward (e.g., by Breyer, 1995; Gouveia, 1997), which show that under plausible assumptions on preferences, the majority of voters will support a two-tier system in which citizens can top up their SHI coverage with PHI (for an opposite result, cf. Hindriks and De Donder, 2003). This is a typical 'ends-against-the-middle' result because the groups of voters who are in favor of a small SHI system are members of the lowest and highest income brackets, whereas middle- and high-income earners end up buying supplementary coverage.

To the second point, Kifmann (2005) showed that there can be majority support for a system with income-dependent contributions if the choice of regime is taken at the 'constitutional stage,' i.e., before the individuals know their health risk, whereas the details of SHI are decided at a later stage after risk types have been revealed. Then even a high earner can vote in favor of a redistributive SHI if the alternative is private insurance with risk-rated premiums and no insurance against a deteriorating risk type over time. This is particularly likely if individuals are sufficiently risk averse and the premiums in a system with risk rating sufficiently dispersed, whereas income inequality in the society should not be too extreme to make the implicit income redistribution too expensive for the high earners. This may explain why in many countries with SHI, contributions are levied only on labor incomes and even on those only up to a ceiling to limit the volume of income redistribution. However, separating 'pure' income redistribution from SHI through flat premiums may be more efficient, but not politically feasible because in such a system the political support for SHI with a generous benefit package, which comes only from the high-risk group, may be too small.

Disclaimer

Part of the material used in this survey is adapted from Zweifel and Breyer (2006). Valuable comments from the Editors and Mathias Kifmann are gratefully acknowledged.

See also: Access and Health Insurance. Adoption of New Technologies, Using Economic Evaluation. Collective Purchasing of Health Care. Demand for and Welfare Implications of Health Insurance, Theory of. Health Care Demand, Empirical Determinants of. Health Insurance and Health. Long-Term Care Insurance. Mandatory Systems, Issues of. Modeling Cost and Expenditure for Healthcare. Moral Hazard. Performance of Private Health Insurers in the Commercial Market. Pharmaceutical Pricing and Reimbursement Regulation in Europe. Prescription Drug Cost Sharing, Effects of.

Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Rationing of Demand. Risk Classification and Health Insurance. Risk Equalization and Risk Adjustment, the European Perspective. Risk Selection and Risk Adjustment. Supplementary Private Health Insurance in National Health Insurance Systems. Supplementary Private Insurance in National Systems and the USA. Switching Costs in Competitive Health Insurance Markets. Value-Based Insurance Design. Welfarism and Extra-Welfarism. Willingness to Pay for Health

References

Arnott, R. and Stiglitz, J. E. (1986). Moral hazard and optimal commodity taxation. *Journal of Public Economics* **29**, 1–24.

Breyer, F. (1995). The political economy of rationing in social health insurance. *Journal of Population Economics* **8**, 137–148.

Cochrane, J. (1995). Time-consistent health insurance. *Journal of Political Economy* **103**, 445–473.

Ellis, R. P. and Manning, W. G. (2007). Optimal health insurance for prevention and treatment. *Journal of Health Economics* **26**, 1128–1150.

Gouveia, M. (1997). Majority rule and the public provision of a private good. *Public Choice* **93**, 221–244.

Hindriks, J. and De Donder, P. (2003). The politics of redistributive social insurance. *Journal of Public Economics* **87**, 2639–2660.

Kifmann, M. (2002). Community rating in health insurance and different benefit packages. *Journal of Health Economics* **21**, 719–737.

Kifmann, M. (2005). Health insurance in a democracy: Why is it public and why are premiums income related? *Public Choice* **124**, 283–308.

Kifmann, M. and Lorenz, N. (2011). Optimal cost reimbursement of health insurers to reduce risk selection. *Health Economics* **20**, 532–552.

Mas-Colell, A., Whinston, M. D. and Green, J. R. (1995). *Microeconomic theory.* New York, Oxford: Oxford University Press.

Newhouse, J. P. (1996). Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature* **34**, 1236–1263.

Pauly, M. V., Kunreuther, H. and Hirth, R. (1995). Guaranteed renewability in insurance. *Journal of Risk and Uncertainty* **10**, 143–156.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**, 629–650.

Van Barneveld, E. M., van Vliet, R. C. J. A. and van de Ven, W. P. M. M. (1996). Mandatory high-risk pooling: An approach to reducing incentives for cream skimming. *Inquiry* **33**, 133–143.

Zweifel, P. and Breyer, F. (2006). The economics of social health insurance. In Jones, A. M. (ed.) *The elgar companion to health economics*, pp. 126–136. Cheltenham: Edward Elgar.

Further Reading

Zweifel, P., Breyer, F. and Kifmann, M. (2009). *Health economics,* 2nd ed. New York: Springer.

Spatial Econometrics: Theory and Applications in Health Economics

F Moscone and E Tosetti, Brunel University, Uxbridge, UK

Glossary

Cross-section dependence This is the dependence among population units, such as individuals, households, cities, industries or countries, in a given cross-section.

Panel data A panel data set is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual in the sample.

Introduction

Spatial econometrics is concerned with measuring and modeling the correlation of observations generated by the inherent spatial structure of the data (Anselin, 1988). Such correlation, known as spatial dependence, may arise from local interaction of individuals, or from unobserved characteristics that are concentrated across space and that affect the variable of interest.

In health economics, spatial dependence may occur, for example, because individuals seek advice by speaking with neighbors regarding a variety of decisions concerning their health: the treatment to be purchased, the hospital wherein to be admitted, the diet to be undertaken. Local interaction can thus lead to an emergent collective behavior that empirically translates into a structure correlation among statistical units in the data. Spatial dependence may also arise if health providers engage in some forms of local strategic interactions, perhaps due to oligopolistic positions or agglomeration economies, when deciding the price to charge or the quality of health services to supply. At aggregate level of analysis, spatial correlation is likely to be present in the data if the investigator cannot observe important risk factors affecting the variable of interest, such as air pollution, migration, and criminality, which could be linked to regional rather than simply local trends, influencing prevalence and need across a wide geographical area. Spatial correlation can also be caused by a variety of measurement problems often found in applied work, or by the particular sampling scheme used to select units. An example is the lack of concordance between the delineation of observed spatial units, such as the region or the country, and the spatial scope of the phenomenon under study (Anselin, 1988). When the sampling scheme is clustered, potential correlation may also arise between respondents belonging to the same cluster. Indeed, units sharing observable characteristics, such as location or industry, may also have similar unobservable characteristics that would cause the regression disturbances to be correlated (Moulton, 1990).

This article provides a survey of econometric methods that are proposed to deal with spatial dependence in the context of linear panel data regression models. It then illustrates the application of spatial econometric methods to tackle problems in health economics. The discussion on these techniques is confined to linear panels with continuous dependent variable, whereas spatial discrete choice models will not be reviewed. Further, owing to space limitations, nonparametric methods for estimation of spatial models will not be discussed.

The plan of the remainder of the article is as follows. Section Spatial Weights and the Spatial Lag Operator introduces the notions of spatial weights and spatial lag. Sections Spatial Dependence in Panel Data Models, Estimation and Heterogenous Panels provide a review of spatial models and discuss their estimation under strictly exogenous regressors, whereas dynamic spatial models are treated in Section Dynamic Panels with Spatial Dependence. Section Testing for Spatial Independence introduces testing for spatial independence. Applications in health economics problems are reviewed in Section Applications of Spatial Econometrics in Health Economics, and Section Concluding Remarks concludes the article.

Spatial Weights and the Spatial Lag Operator

In spatial econometrics, the neighbor relation is typically expressed by the means of a nonnegative matrix, known as spatial weights matrix. In a spatial weights matrix, often indicated by W, the rows and columns correspond to the cross-section observations (e.g., individuals, regions, or countries), and the generic element, $w_{ij,}$ can be interpreted as the strength of potential interaction between units i and j. The specification of W is generally arbitrary, typically based on some measures of distance between units, using, for example, contiguity or geographic proximity, or more general metrics, such as economic, political, or social distance. To avoid nonlinearity and endogeneity problems, spatial weights should be exogenous to the model, a condition that is not guaranteed when using more general distance metrics. By convention, the diagonal elements of the weighting matrix are set to 0, implying that an observation is not a neighbor to itself. Further, to facilitate the interpretation of estimates in spatial models, W is typically row-standardized so that the sum of the weights for each row is 1, ensuring that all the weights are between 0 and 1. Finally, although most empirical works assume that weights are time-invariant, these can vary over time.

An important role in spatial econometrics is played by the notion of spatial lag operator. Let z_{it} be the observation on a variable for the ith cross-section unit at time t for $i=1, 2,\ldots, N$; $t=1, 2,\ldots, T$. Let $z_t=(z_{1t}, z_{2t},\ldots, z_{Nt})'$, and $W=\{w_{ij}\}$ be a time-invariant $N\times N$ spatial weights matrix. The spatial lag of z_t is given by Wz_t, with generic ith element

$$\sum_{j=1}^{N} w_{ij}z_{jt}$$

Hence, a spatial lag operator constructs a new variable, which is a weighted average of neighboring observations, with

weights reflecting distance among units. The incorporation of these spatial lags into a regression specification is considered in the next section.

Spatial Dependence in Panel Data Models

Spatial Lag Models

Several problems in the social sciences require the inclusion in the regression model of spatial lags of the dependent variable among the regressors. Under this specification,

$$y_{it} = \alpha_i + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \beta' x_{it} + u_{it}, \quad i = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, T \qquad [1]$$

where x_{it} is a $k \times 1$ vector of observed regressors on the ith cross-section unit at time t, u_{it} is the error term, and ρ and β are unknown parameters to be estimated. The group effects, α_i, could be either considered fixed, unknown parameters to be estimated, or draws from a probability distribution. Section Estimation will discuss estimation under these two alternative frameworks. For the time being, it is assumed that regressors are strictly exogenous, and nonstochastic.

It is often convenient to rewrite eqn [1] in stacked form:

$$y_t = \alpha + \rho \mathrm{W} y_t + X_t \beta + u_t \qquad [2]$$

where $y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)'$, $X_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})'$, and $u_t = (u_{1t}, u_{2t}, \ldots, u_{Nt})'$. Under certain asymptotic conditions on the spatial weights matrix, the correlation between the spatial lag of the dependent variable, $\mathrm{W}y_t$ and the error term, u_t, is nonzero, if $\rho \neq 0$. For this reason, conventional estimators of parameters ρ and β are inconsistent, and alternative estimation approaches, such as maximum likelihood (ML) and generalized method of moments (GMM), are needed.

Spatial Error Models

Another way to incorporate spatial dependence in the regression equation is by allowing disturbances to be spatially correlated. Consider the simple linear regression in stacked form:

$$y_t = \alpha + X_t \beta + u_t \qquad [3]$$

where the notation is as above. There exist few main approaches to assign a spatial structure to the error term u_t (as seen in Section Fixed effects specification, if α is assumed to be random, then a spatial structure could also be assigned to it); the intent is to represent the covariance as a simpler and lower dimensional matrix than the unconstrained.

One way is to define the covariance between two observations directly as a function of the distance between them. Accordingly, the covariance matrix for the cross-section at time t is $E(u_t u_t') = f(\theta, \mathrm{W})$, where θ is a parameter vector and f is a suitable distance decay function, such as the negative exponential. The decaying function suggests that the disturbances should become uncorrelated when the distance separating the observations is sufficiently large. One shortcoming of this method is that it requires the specification of a functional form for the distance decay, which is subject to a degree of arbitrariness.

An alternative strategy consists of specifying a spatial process for the error term, which relates each unit to its neighbors through W. The most widely used is the Spatial Autoregressive (SAR) specification. Proposed by Cliff and Ord (1969), the SAR process is

$$u_t = \delta \mathrm{W} u_t + \varepsilon_t \qquad [4]$$

where δ is a scalar parameter, and $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{Nt})'$, with $\varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2 \mathrm{I}_N)$. Other spatial processes suggested to model spatial error dependence, although less used in the empirical literature, are the Spatial Moving Average (SMA) and the Spatial Error Component (SEC) specifications. The first, proposed by Haining (1978), assumes that

$$u_t = \delta \mathrm{W} \varepsilon_t + \varepsilon_t \qquad [5]$$

where ε_t is as above. According to the SEC specification, introduced by Kelejian and Robinson (1995),

$$u_t = \delta \mathrm{W} \psi_t + \varepsilon_t \qquad [6]$$

where $\psi_t = (\psi_{1t}, \psi_{2t}, \ldots, \psi_{Nt})'$ and $\psi_{it} \sim IID(0, \sigma_\psi^2)$. A major distinction between the SAR and the other two specifications is that in the first there is an inverse involved in the covariance matrix. This has important consequences on the range of dependence implied by its covariance matrix. Indeed, even if W contains few nonzero values, the covariance structure induced by the SAR is not sparse, linking all the units in the system to each other, so that a perturbation in the error term of one unit will be ultimately transmitted to all other units. Conversely, for the SMA and SEC, the only off-diagonal nonzero elements of the covariance matrix are those corresponding to nonzero elements in W. Under certain invertibility conditions, spatial processes eqns [4]–[6] can all be written as special cases of the following general form

$$u_t = R \varepsilon_t \qquad [7]$$

where R is a $N \times N$ matrix. For example, for an invertible SAR process $R = (I_N - \delta \mathrm{W})^{-1}$, whereas in the case of an SMA, $R = I_N + \delta \mathrm{W}$.

Conventional panel estimators such as the fixed effects (FE) or random effects (RE) estimators of slope coefficients in eqn [3] with spatially dependent errors are $\sqrt{NT}$-consistent under broad regularity conditions and strictly exogenous regressors. However, these estimators are in general not efficient because the covariance matrix of errors is nondiagonal and the elements along its main diagonal are not constant.

Estimation of spatial models is considered next.

Estimation

Maximum Likelihood Estimator

The theoretical properties of quasi-ML estimator in a single cross-section framework have been studied by Anselin (1988) and Lee (2004), among others. More recently, considerable work has been undertaken to investigate the properties of ML

estimators in panel data, in the presence of spatial dependence and unobserved time-invariant heterogeneity.

Fixed effects specification

For ML estimation of spatial regression models, it is convenient to consider the general case of a spatial lag model having SAR errors:

$$y_t = \alpha + \rho W_1 y_t + X_t\beta + u_t \quad [8]$$

$$u_t = \delta W_2 u_t + \varepsilon_t \quad [9]$$

where the spatial lags in the dependent variable and in the error term are constructed using two (possibly different) spatial weights matrices, W_1 and W_2. Suppose that the group effects are treated as fixed and unknown parameters, and that $\varepsilon_{it} \sim IID(0,\sigma_\varepsilon^2)$ Lee and Yu (2010a) propose a transformation of the above model to get rid of the FE, and then use ML to estimate the remaining parameters, ρ, β, δ, and σ_ε^2. Specifically, the authors suggest to multiply all variables by a $T\times(T-1)$ matrix, P, having as columns, the $(T-1)$ eigenvectors associated to the nonzero eigenvalues of the deviation from the mean transformation, $M = I_T - 1_T\left(1_T'1_T\right)^{-1}1_T'$, where 1_T is a T-dimensional vector of 1. It is easily seen that $1_T'P = 0$ so that such transformation removes the individual-specific intercepts. After the transformation, the effective sample size reduces to $N(T-1)$, and because $P'P = I_{T-1}$, the new error term has uncorrelated elements. Under some identification conditions, the estimator of the unknown parameters, obtained by maximizing the transformed model's log-likelihood function, is consistent and asymptotically normal when either N or T, or both, are large.

Random effects specification

This formulation assumes that the group effects, α_i, are random and independent of the exogenous regressors. In this case, following Baltagi *et al.* (2009), a general specification can be suggested by assuming that spatial processes apply both to the random group effects and the remainder disturbances:

$$y_t = \rho W_1 y_t + X_t\beta + v_t \quad [10]$$

$$v_t = \alpha + u_t \quad [11]$$

$$\alpha = \gamma W_2 \alpha + \mu \quad [12]$$

$$u_t = \delta W_3 u_t + \varepsilon_t \quad [13]$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_N)'$, and it is assumed that $\mu_i \sim IID(0,\sigma_\mu^2)$ and $\varepsilon_{it} \sim IID(0,\sigma_\varepsilon^2)$. The above model, by distinguishing between time-invariant spatial error spillovers and spatial spillovers of transitory shocks, encompasses a variety of econometric specifications that are proposed in the literature as special cases. If the same spatial process applies to α and u_t (i.e., $\delta = \gamma$ and $W_2 = W_3$), this model reduces to that proposed by Kapoor *et al.* (2007); if $\gamma = 0$, it simplifies to that considered by Baltagi *et al.* (2003).

Consistency of estimator for the unknown parameters based on maximization of the model's log-likelihood is established in Baltagi *et al.* (2009). A set of joint and conditional specification Lagrange Multiplier (LM) tests for spatial effects within the RE framework are proposed by Baltagi *et al.* (2009). These statistics allow testing model (10)–(13) against their restricted counterparts: the Anselin model, the Kapoor *et al.* models, and the RE model without spatial correlation.

Instrumental Variables and GMM

In the presence of heteroskedasticity, the ML estimator for spatial models under the incorrect assumption of spherical disturbances is generally inconsistent. As an alternative, instrumental variables (IV) and GMM techniques have been suggested.

In a single cross-section setting, Kelejian and Prucha (1998) propose a simple IV strategy to deal with the endogeneity of the spatially lagged dependent variable Wy_t that consists of using as instruments, the spatially lagged (exogenous) explanatory variable WX_t. The IV approach can be easily adapted in the context of spatial panel data models with either FE or RE. Hence, the Hausman's specification test can be used to choose between FE and RE specification (Mutl and Pfaffermayr, 2011).

GMM estimation of spatial regression models for a single cross-section has been originally advanced by Kelejian and Prucha (1999). The authors focus on a regression equation with SAR disturbances and suggest the use of three moment conditions that exploit the properties of disturbances implied by a standard set of assumptions. Estimation consists of solving a nonlinear optimization problem, which yields a consistent estimator under a number of regularity conditions. Considerable work has been carried to extend this procedure in various directions. Liu *et al.* (2010) suggest a set of moments that encompass Kelejian and Prucha conditions as special cases. Kelejian and Prucha (2009) generalize their original work to include spatial lags in the dependent variable and allowing for heteroskedastic disturbances. This setting is extended by Kapoor *et al.* (2007) to estimate a spatial panel regression model with group error components and by Moscone and Tosetti (2011) for a panel with fixed effects. One advantage of the GMM procedure over ML is that it is computationally simpler, especially when dealing with unbalanced panels.

Heterogenous Panels

For panel data studies with large N and small T, observations are usually pooled and homogeneity of the slope coefficients is assumed. The latter is a testable assumption, which is often rejected in practice. A recent literature argues in favor of heterogenous estimates and suggests the following specification with heterogenous slopes

$$y_{it} = \alpha_i' d_t + \beta_i' x_{it} + u_{it} \quad [14]$$

where $d_t = (d_{1t}, d_{2t}, \ldots, d_{nt})'$ is a $n\times 1$ vector of observed common effects (e.g., a time trend), x_{it} is a k-dimensional vector of strictly exogenous regressors, and β_i follow the random coefficient model $\beta_i = \beta + v_i$, with $v_i \sim (0,\Omega_v)$. It is further

assumed that errors are generated by a spatial process having form (7), such as SAR, SMA, or SEC, where ε_t follows a covariance-stationary process. Pesaran and Tosetti (2011) focus on estimation of the cross-section means of parameters, $\beta = E(\beta_i)$, by conventional FE estimator, and by the following mean group estimator

$$\hat{\beta}_{MG} = N^{-1} \sum_{i=1}^{N} \hat{\beta}_i \qquad [15]$$

where

$$\hat{\beta}_i = \left(X_i' M_D X_i \right)^{-1} X_i' M_D y_i$$

$y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $X_t' = (x_{i1}, x_{i2}, \ldots, x_{iT})$, $M_D = I_T - \mathrm{D}(D'D)^{-1} D'$, and $D' = (d_1, d_2, \ldots, d_T)$. The authors show that under some regularity conditions, as N and T tend to infinity, $\hat{\beta}_{MG}$, the asymptotic distribution of $\hat{\beta}_{MG}$, (as well as that of the conventional FE estimator) does not depend on the particular spatial structure of the error, u_{it}, but only on Ω_v. Robust estimators for the variances of $\hat{\beta}_{MG}$ can be obtained following the nonparametric approach employed in Pesaran (2006). One advantage of this method is that it does not require *a priori* knowledge of the spatial arrangement of cross-sectional units. Indeed, misspecification of the spatial weights matrix may lead to substantial size distortions in tests based on the quasi-ML estimators of β_i (or β).

Temporal Heterogeneity

Temporal heterogeneity may be incorporated in a spatial version of the Seemingly Unrelated Regression Equations (SURE) approach, as suggested by Anselin (1988). This approach, suitable when N greatly exceeds T, permits slope parameters to vary over time, and errors are allowed to be both spatially and serially correlated. In its more general form, the spatial SURE is

$$y_t = \rho_t \mathrm{W}_1 y_t + X_t \beta_t + u_t \qquad [16]$$

$$u_t = \delta_t W_2 u_t + \varepsilon_t \qquad [17]$$

where β_t, ρ_t, and δ_t are time-varying parameters, and ε_t satisfies $E(\varepsilon_t \varepsilon_s') = \sigma_{ts} I_N$. Let Ω be a $T \times T$ positive definite matrix with elements σ_{ts}. ML or GMM techniques can be used to estimate the above model.

Dynamic Panels with Spatial Dependence

In the recent years, considerable work has been undertaken on estimation of panel data models that include both spatial and temporal dynamics. A variety of spatiotemporal models have been proposed in the literature. Consider the following general dynamic spatial panel:

$$y_t = \alpha + \gamma y_{t-1} + \rho \mathrm{W} y_t + \lambda \mathrm{W} y_{t-1} + X_t \beta + u_t \qquad [18]$$

The above model can be classified into different cases depending on the eigenvalue matrix of its reduced form. In particular, eqn [18] is stable if $\gamma + \rho + \lambda < 1$; spatial cointegration takes place when $\gamma + \rho + \lambda = 1$; whereas under the explosive case, $\gamma + \rho + \lambda > 1$. Under $\gamma + \rho + \lambda < 1$, Yu *et al.* (2008) derive the ML of the FE specification, showing that when T is large relative to N, the estimators are consistent and asymptotically normal, whereas when $N/T > 0$, the limit distribution is not centered around 0, in which case, the authors propose a bias correction. Under $\gamma + \rho + \lambda = 1$, the ML estimator is consistent and asymptotically normal as in the stationary case, although spatial cointegration yields a singular asymptotic covariance matrix for the ML estimator. When $\gamma + \rho + \lambda > 1$, the ML is not tractable, although it turns tractable after applying a transformation that renders stable, the explosive variables (Lee and Yu, 2010b).

Testing for Spatial Independence

Spatial econometrics literature proposes a number of statistics for testing the null hypothesis of spatial independence, i.e., $H_0 : E(u_{it} u_{jt}) = 0, \; i \neq j$ in Model (3). The majority of these tests have been studied only in the case of a single cross-section. One of the most commonly used is the Moran's statistic (Kelejian and Prucha, 2001), which, when extended to a panel set up, takes the form

$$CD_{\mathrm{Moran}} = \frac{\sum_{t=1}^{T} \hat{u}_t' \mathrm{W} \hat{u}_t}{\left(T \hat{\sigma}_\varepsilon^4 \sum_{i=1}^{N} \sum_{j=1}^{i-1} \left(w_{ij} + w_{ji} \right)^2 \right)^{1/2}} \qquad [19]$$

where $\hat{\sigma}_\varepsilon^2$ is a consistent estimator for σ_ε^2, and $\hat{\mathrm{u}}_t$ is a consistent estimator of regression errors. The CD_{Moran} is asymptotically normally distributed.

The information on the distance among units can also be used to build 'local' versions of some statistics as proposed in the panel literature to test against generic forms of cross-section dependence. For example, the local CD_P test proposed by Pesaran (2004) is

$$CD_{P,\mathrm{Local}} = \sqrt{\frac{T}{S_0}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \hat{\rho}_{ij} \right) \qquad [20]$$

where $S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, and $\hat{\rho}_{ij}$ is the sample pairwise correlation coefficient computed between FE residuals of units i and j. The $CD_{P,\mathrm{Local}}$ test is asymptotically normally distributed. The reader is referred to Moscone and Tosetti (2009) for a review of this literature.

Applications of Spatial Econometrics in Health Economics

The methods described in this article have been employed to study a variety of problems in regional and urban sciences, geography, economics, crime analysis, environmetrics, epidemiology, and public health. The recognition of a marked geographical concentration for many health indicators has encouraged a wide use of spatial methods to analyze health economics issues. This section first illustrates empirical evidence on spatial concentration for a number of health indicators, and then considers works on local interaction among healthcare providers.

Health Outcomes, Risk Factors, and Health Needs

There exists a growing literature adopting spatial econometric methods to model geographical clustering of various health conditions such as health status and mortality, obesity, and diseases, both communicable, like poliomyelitis, influenza, and HIV, and noncommunicable, like diabetes, cardiovascular problems, and mental health disorders. Lorant *et al.* (2001) study the impact on mortality of deprivation, measured by the Townsend index, and of a set of socioeconomic indicators, in Belgium at municipality level from 1985 to 1993. The authors estimate a spatial lag model and find evidence of high significant spatial effects ($\rho = 0.6$) in mortality and of positive influence of deprivation on mortality. Chen *et al.* (2010) investigate the impact of access to chain grocers on body mass index (BMI) in Indiana (USA) in 2005. The authors estimate a spatial lag model by ML to control for possible 'obesity epidemic' effects, allowing the influence of access to chain grocers on BMI to differ depending on whether or not, a person lives in a low-income community. Empirical results suggest that improvement in access to chain grocer access significantly reduces the average BMI, in low-income communities. Congdon (2002) studies geographical variations in mental health outcomes proxied by hospital and community referrals for a set of diagnostic categories, using data on people living in a London health authority over the period 1994–99. Using Bayesian methods, the author derives an index of needs for mental health problems that includes spatial dependence and a set of sociodemographic variables, such as deprivation, community integration, and ethnicity. Hence, the author compares the forecasting performance of the developed index with that of traditional needs indices and shows that the forecast performance in predicting referrals improves consistently when accounting for spatial effects. One policy implication of the above studies is that formulae used to allocate healthcare resources across geographical areas could be ameliorated by incorporating spatial correlation.

Health Expenditure

Recent works in health economics and in the medical literature indicate that one important element explaining variations in health expenditure is represented by the spillover effect, that is, expenditure on health services in one locality can have beneficial or harmful effects across a wider geographical area. A number of factors can justify such wider effects. For example, a municipality may choose a particular course of action so as to persuade individual service users, families, or indeed service providing bodies, to migrate into or out of their area. Such flow can be encouraged by ensuring that health expenditure, clinical activity, or health policy is more (or less) attractive than that offered in neighboring authorities. Politicians may adopt this strategy as voters perhaps judge them relative to those in nearby localities. A municipality good (or bad) performance may encourage neighboring municipalities to mimic (or avoid) the activities and expenditure patterns associated with such performance.

A large number of papers have empirically tested the above hypotheses. One influential example is the work by Baicker (2005) that explores the extent to which health spending in one state is influenced by the spending in neighboring states. The author adopts IV and GMM approaches to estimate a spatial lag model for 48 contiguous US states, in the years from 1983 to 1992. One conclusion of this work is that states, in response to $1 increase in neighbors' expenditure, raise their own expenditure by almost a full dollar. The reader is referred to Revelli (2006) and Moscone *et al.* (2007) for studies on the UK. Recent studies investigate the long-run dynamics of health spending. For example, Moscone and Tosetti (2010), using a panel of 49 US States over the period 1980–2004, estimate a regression equation for health spending assuming that errors are spatially correlated and also depend on a set of unobserved common factors. The authors find evidence of sizeable spatial correlation in health spending even after controlling for unobserved effects.

A number of works look at healthcare resources consumption and utilization rather than expenditure. For example, Filippini *et al.* (2010) study the demand of antibiotics in 240 Swiss regions in 2002. The authors estimate a spatial error model to account for infection spreading and find that dispensing practices induce higher rate of antibiotic consumption, even after controlling for patient characteristics, epidemiological variables, access to drug treatment, and spatial dependence. Joines *et al.* (2003) investigate the determinants of hospital admission rates in California and find significant spatial effects in hospitalization rates.

Hospital Competition and Agglomeration

Mobley (2003) adopts spatial econometric methods to study hospital competition under managed care in the State of California in the years 1993 and 1998. The author considers a SURE model with spatial lags of the dependent variable and estimates it by ML. Empirical results show that the price charged by a hospital is affected by the price set by neighboring hospitals, suggesting that such information may be used to design antitrust policies. Moscone *et al.* (2011) study hospital competition exploring the determinants of patients' hospital choices, using 144 Italian hospitals in the years from 2004 to 2007. The authors conclude that the likelihood of choosing an hospital by an individual is significantly influenced by the experience in utilization of health services by patients living in the same postal code. However, the use of neighborhood information on average does not seem to lead patients to high quality hospitals. Cohen and Paul (2008) investigate why hospitals concentrate across territory, using data on 93 Washington state hospitals during 1997–2002. The authors estimate a system of cost function and input demand equations, which include an agglomeration variable as a cost shift factor, measured by the spatial lag of labor forces in neighboring hospitals. Results show significant agglomeration economies, perhaps due to cost saving generated by knowledge sharing with adjacent hospitals, labor market pooling, or lower employment search costs.

Concluding Remarks

This article has surveyed the most recent econometric methods for panel data dealing with spatial effects. Recent developments in spatial econometrics offer new methods for

representing the spatiotemporal dynamics of many health economics phenomena. However, the range of spatial techniques adopted until now in health economics is rather limited, when compared to the methods developed in the literature. For instance, only few works have incorporated in their specification time-invariant unobserved heterogeneity and/or temporal dynamics. The use of recently developed techniques in spatial econometrics may offer insights and raise new questions in several areas of health economics.

Acknowledgment

The authors acknowledge the financial support from ESRC (Ref. no. RES-061-25-0317). They would like to thank the editor John Mullahy and Anirban Basu for their helpful comments on an earlier version of this article.

See also: Competition on the Hospital Sector. Dynamic Models: Econometric Considerations of Time. Modeling Cost and Expenditure for Healthcare. Panel Data and Difference-in-Differences Estimation

References

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baicker, K. (2005). The spillover effects of state spending. *Journal of Public Economics* **89**, 529–544.

Baltagi, B., Song, S. and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* **111**, 123–150.

Baltagi, B. H., Egger, P., and Pfaffermayr, M. (2009). A generalized spatial panel data model with random effects. *CPR Working Papers No. 113*. Syracuse, NY: Center for Policy Research, Maxwell School, Syracuse University.

Chen, S., Florax, R. J., Snyder, S. and Miller, C. C. (2010). Obesity and access to chain grocers. *Economic Geography* **86**, 431–452.

Cliff, A. D. and Ord, J. K. (1969). The problem of spatial autocorrelation. In Scott, A. J. (ed.) *London papers in regional science*. London: Pion.

Cohen, J. P. and Paul, C. M. (2008). Agglomeration and cost economies for Washington. *Regional Science and Urban Economics* **38**, 553–564.

Congdon, P. (2002). A model for mental health needs and resourcing in small geographic areas: A multivariate spatial perspective. *Geographical Analysis* **34**, 168–186.

Filippini, M., Masiero, G. and Moschetti, K. (2010). Dispensing practices and antibiotic use. *Working Papers no. 1006*. Bergamo, Italy: Department of Economics and Technology Management, University of Bergamo.

Haining, R. P. (1978). The moving average model for spatial interaction. *Transactions of the Institute of British Geographers* **3**, 202–225.

Joines, J. D., Hertz-Picciotto, I., Carey, T. S., Gesier, W. and Suchindran, C. (2003). A spatial analysis of county-level variation in hospitalization rates for low back problems in North Carolina. *Social Science & Medicine* **56**, 2541–2553.

Kapoor, M., Kelejian, H. H. and Prucha, I. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics* **140**, 97–130.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* **17**, 99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* **40**, 509–533.

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the moran I-test with applications. *Journal of Econometrics* **104**, 219–257.

Kelejian, H. H. and Prucha, I. R. (2009). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* **157**, 53–67.

Kelejian, H. H. and Robinson, D. P. (1995). Spatial correlation: A suggested alternative to the autoregressive model. In Anselin, L. and Florax, R. J. (eds.) *New directions in spatial econometrics*, pp. 75–95. Berlin: Springer-Verlag.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**, 1899–1925.

Lee, L. F. and Yu, J. (2010a). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* **154**, 165–185.

Lee, L. F. and Yu, J. (2010b). Some recent developments in spatial panel data models. *Regional Science and Urban Economics* **40**, 255–271.

Liu, X., Lee, L.-F. and Bollinger, C. (2010). Improved efficient quasi maximum likelihood estimator of spatial autoregressive models. *Journal of Econometrics* **159**, 303–319.

Lorant, V., Thomas, I., Deliege, D. and Tonglet, R. (2001). Deprivation and mortality: The implications of spatial autocorrelation for health resources allocation. *Social Science and Medicine* **53**, 1711–1719.

Mobley, L. R. (2003). Estimating hospital market pricing: An equlibrium approach using spatial econometrics. *Regional Science and Urban Economics* **33**, 489–516.

Moscone, F., Knapp, M. and Tosetti, E. (2007). Mental health expenditure in England: A spatial panel approach. *Journal of Health Economics* **26**, 842–864.

Moscone, F. and Tosetti, E. (2009). A review and comparison of tests of cross section independence in panels. *Journal of Economic Surveys* **23**, 528–561.

Moscone, F. and Tosetti, E. (2010). Health expenditure and income in the US. *Health Economics* **19**, 1385–1403.

Moscone, F. and Tosetti, E. (2011). GMM estimation of spatial panels with fixed effects and unknown heteroskedasticity. *Regional Science and Urban Economics* **41**, 487–497.

Moscone, F., Tosetti, E. and Vittadini, G. (2011). Social interaction in patients' hospital choice: Evidence from Italy. *Journal of the Royal Statistical Society: Series A* **175**, 453–472.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* **72**, 334–338.

Mutl, J. and Pfaffermayr, M. (2011). The Hausman test in a cliff and ord panel model. *Econometrics Journal* **14**, 48–76.

Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. *CESifo Working Paper Series 1229*. Munich: CESifo Group Munich.

Pesaran, M. H. (2006). Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica* **74**, 967–1012.

Pesaran, M. H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics* **161**, 182–202.

Revelli, F. (2006). Performance rating and yardistick competition in social service provision. *Journal of Public Economics* **90**, 459–475.

Yu, J., de Jong, R. and Lee, L. F. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both N and T are large. *Journal of Econometrics* **146**, 118–137.

Specialists

DJ Wright, University of Sydney, Sydney, NSW, Australia

Introduction

Specialists have a unique position in the health system as they provide health-care services to patients and so are agents of patients, but in addition, they provide patients to, and order services from, other health-care providers (e.g., hospitals) and so are agents of other health-care providers as well. In the economics literature this situation is known as common agency. These two agency relationships are characterized by asymmetric information, in that patients have limited medical knowledge about the type of specialist to visit, the quality of any particular specialist, or the appropriate treatment, whereas hospitals have limited knowledge about the preferences of specialists over income and patient welfare.

In addition to having a unique position in the health-care system, specialists operate in a unique market. Patients' health-care expenditures are often covered by public or private insurance and so the demand for specialists' services is often not rationed by price. The payment that specialists receive for their services is either regulated by the public purchaser, the private insurer, or a private purchaser such as a Preferred Provider Organization or a Health Maintenance Organization. As a result, the quality and the quantity of services specialists provide depend very much on the incentives contained in their payment schemes.

In this environment, some interesting questions are: (1) what factors lead to particular specialty choices by doctors, (2) how are patients allocated between general practitioners and specialists, (3) is the quality and quantity of services provided by specialists efficient, and (4) how are specialists allocated between hospitals. These questions are addressed below.

The Allocation of Doctors

An important result in economics is the First Fundamental Theorem of Welfare Economics that states that competitive markets are efficient. A fundamental characteristic of competitive markets is free entry. However, entry into the medical profession is not free as places at medical schools are restricted by governments and entry into specialties is restricted by various professional bodies. As a result, the allocation of physicians between specialties is often inefficient. For example, in many countries, the market for specialties is characterized by excess supply, whereas the market for general practitioners is characterized by excess demand.

These inefficiencies have led researchers to investigate what factors are important in the specialty choice decision of doctors as these factors can then be manipulated to achieve a more efficient outcome. A number of US and Canadian studies have found that graduates of medical schools are more likely to choose specialties with less demanding workloads and higher expected income. In the US, the income elasticity of a particular specialty choice has been estimated to be 1.4, where the income elasticity of a particular specialty choice was defined as the percentage increase in the number of medical students who rank a particular specialty first associated with a 1% increase in expected income. In Canada, the income elasticity of any specialty choice (relative to general practice) has been estimated to be 0.2, where the income elasticity of any specialty choice was defined as the percentage increase in the number of physicians who chose any specialty (relative to general practice) associated with a 1% increase in the fee-per-consultation. Specialty choice is quite responsive to expected income. Therefore, policy makers can influence specialty and general practitioner choice through income differentials and by changing the work environment.

The Allocation of Patients

Patients with a medical condition have limited medical knowledge and so do not know whether a general practitioner (GP) or a specialist is needed to treat the condition. Even if patients knew they need to see a specialist, they do not know which type, for example, does a patient with chest pain need to see a cardiologist or a thoracic surgeon? To analyze the first situation, consider a model in which treatment by a specialist is more expensive than by a GP, but for some conditions treatment is only successful if completed by a specialist. In this latter case, if the GP treats the patient, the patient suffers a waiting loss as specialist treatment is delayed. To determine which type of condition the patient has required the GP to expend effort. In this environment, if specialist costs and patient waiting costs are high, then a gate-keeping system, in which patients must get a referral from a GP before seeing a specialist, is more efficient than a system in which referrals are not needed. Under gate-keeping, a GP payment scheme, which involves a bonus for nonreferral and another bonus for not providing treatment provides GPs with an incentive to undertake effort and only refer those patients with the condition that only specialists can treat. This saves on patient waiting costs and ensures that patients with the condition which is treatable by GPs are not referred to specialists. GP fund-holding, whereby GPs are given a budget from which they pay for a range of elective procedures for their patients, also provides incentives for GPs not to refer patients to specialists and hospitals as it acts like a bonus for nonreferral. However, this is at the cost of not referring patients who should be referred.

The second situation is similar to the first. In a gate-keeping system the GP makes a diagnosis and makes an informed decision about which type of specialist the patient should see. However, in a non gate-keeping system, the uninformed patient chooses which type of specialist to see and if they make an incorrect decision, they waste scarce specialist time and delay appropriate treatment. Where specialist costs are

high and delay in treatment is costly, a gate-keeping system dominates a non gate-keeping system.

Specialist Quality and Quantity of Care

The benefit a patient, who seeks diagnosis or treatment from a specialist, receives depends on the quality of care provided by the specialist, Q, the quantity of medical services delivered by the specialist, q, and the quantity of medical services the specialist orders from other health-care providers such as hospitals, q_0. Let this benefit be given by $B=B(Q,q,q_0)$, where benefit is increasing in all three variables. Patient benefit also depends on the quality of the services provided by other health-care providers, but this is outside the domain of the specialist. The cost of the specialist providing services is given by $C=C(Q,q)$ and the cost of the services provided by other care providers is $C_0=C_0(q_0)$. Both these costs are increasing in all variables. The efficient provision of quality occurs where the extra patient benefits a unit of quality generates equals the extra cost of providing that unit. A similar condition applies for the efficient provision of the quantity of specialist services and the quantity of services ordered from other health-care providers.

Illness is an uncertain event and potential patients are often covered by either private or public health insurance for any health expenditures they incur. The terms of these insurance contracts affect the decision to seek care from a specialist and then the specialist determines the quality and quantity of care once care has been sought. Therefore, the payment scheme that the public or private insurer offers to specialists and the incentives it contains will be a major determinant of the quality and quantity of specialist care.

If quality and quantity were observable and verifiable by courts, a contract between the public or private purchaser and the specialist could be written in which payment was conditional on a particular quality and quantity of service. Competition between private purchasers or maximization of welfare by the public purchaser would then ensure the efficient provision of quality and quantity of health care. However, although the quantity of services is observable and verifiable the same is not true of quality. The quality of a medical service has many dimensions, for example, the appropriate diagnosis of a patient's condition, the appropriate treatment given diagnosis, and appropriate pain management. These dimensions are likely to vary from condition to condition and from patient to patient and so are not easy to specify in a contract in a verifiable manner. Therefore, efficiency cannot be achieved by the writing of conditional payment contracts.

Although quality is not verifiable in courts, it nevertheless maybe be observable to patients. If patients can observe the quality of provider care and the demand for a particular provider is increasing in quality, it is well known that a prospective payment (a payment per patient) coupled with fee-for-service (a payment per service) can achieve efficient provision of provider quality and quantity of service. Essentially the fee-for-service is set to achieve the efficient provision of quantity and the prospective payment is set to achieve the efficient quality as providers compete for valuable patients.

For specialist services, the assumption that patients can observe quality is problematic. Patients in general have little or no information concerning what is the appropriate treatment. In addition, their relationship with specialists is often once-off and so they cannot use past experiences as an indicator of quality as they might do with other health-care providers with whom they have long-term relationships. As competition for patients cannot be relied upon to ensure efficient quality what institutions or avenues exist which might?

Institutions

Specialists need to be licensed to practice, but given evidence from the US that current specialist licensing arrangement guarantee nothing more than minimum specialist expertise and little specialist learning postmedical school, licensing does not provide an indication of specialist quality. In addition, specialist quality is not ensured through selfregulation as review boards fail to respond to patient complaints, and when they do, they rarely impose serious disciplinary sanctions.

Tort law is a vehicle through which specialists who are negligent are liable for any damage that results from not exercising due care. This gives specialists an incentive to maintain quality, but not a very strong one as evidence suggests patients are not very good at detecting negligence when it occurs.

Unlike patients, gate-keeping GPs do form long-term relationships with specialists. As a result, over time, they are able to monitor many patient outcomes from treatment by particular specialists and base referral decisions on these outcomes. This is an imperfect mechanism for ensuring specialist quality as GP referral decisions are not only based on past patient outcomes, but on other factors, for example, friendship. Furthermore, where GPs perfectly observe specialist quality and base their referrals only on quality, specialists have an incentive to overprovide quality to receive more referrals. This incentive to overprovide quality is mitigated if specialists are paid by a combination of a prospective payment and fee-for-service.

In summary, licensing, tort law, and GP gate-keeping provide only limited incentives for specialists to provide quality efficiently.

Specialist Preferences

An avenue for the efficient provision of quality and quantity exists if specialists not only value income, but also the welfare of their patients. Such altruism on the part of providers is thought to be an important characteristic of medical services. Therefore, it is assumed that specialist utility is given by $U=\alpha B(Q,q,q_0)+I-C(Q,q)$, where α is the weight the specialist attaches to patient welfare and I is income. Under the assumption that specialist payments are regulated in most countries by public or private purchasers, specialist income is assumed to consists of a prospective payment, P, and cost reimbursement of $r\cdot C(Q,q)$, where r is the proportion of cost that is reimbursed, that is, $I=P+r\cdot C(Q,q)$. It is assumed that cost is observable.

If the payment scheme involves only full-cost reimbursement, so that $P=0$ and $r=1$, specialist utility is $U=\alpha B(Q,q,q_0)$. The specialist maximizes utility by choosing quality so that the extra benefit the patient gets from an extra unit of quality is zero. The specialist is not concerned with the extra cost involved in providing an extra unit of quality as costs are fully reimbursed. Therefore, relative to the efficient quality, the specialist provides 'too much' quality. Similarly, the specialist provides 'too many' services and orders 'too many' services from other providers relative to the efficient quantities.

However, if the payment scheme involves only a prospective payment, so that $P>0$ and $r=0$, specialist utility is $U=\alpha B(Q,q,q_0)+P-C(Q,q)$. The specialist maximizes utility by choosing quality so that the extra benefit the patient gets from an extra unit of quality weighted by α equals the extra cost of providing that unit of quality. If the specialist values patient benefit and income equally, that is, if $\alpha=1$, then the specialist provides the efficient quality. However, if the specialist values patient benefit less than income, $\alpha<1$, then the specialist provides 'too little' quality relative to the efficient quality. Similarly, the specialist provides 'too few' services relative to efficient provision, but orders 'too many' services from other providers relative to efficient provision as the specialist bears none of the costs of the other providers' services.

Finally, if the payment scheme involves both a prospective payment and some cost reimbursement, so that $P>0$ and $r<1$, specialist utility is $U=\alpha B(Q,q,q_0)+P-(1-r)C(Q,q)$. The specialist maximizes utility by choosing quality so that the extra benefit the patient gets from an extra unit of quality weighted by α equals the extra cost of providing that unit of quality weighted by $1-r$. If r is chosen so that $r=1-\alpha$, then the specialist provides the efficient quality and also the efficient quantity of services. It is still the case that 'too many' services are ordered from other health-care providers.

So even though quality is not observed, a payment scheme, which is a mix of a prospective payment and partial cost reimbursement (supply-side cost sharing) is able to induce the specialist to provide the efficient quality and quantity of own services. Having patient welfare in the specialist utility function provides an incentive for the specialist to provide quality and reimbursing some of the cost of quality provision reinforces this incentive.

The costs of the specialist not only depend on quality and quantity, but also on the amount of unobservable cost reducing effort the specialist expends. In this case, specialist costs are given by $C(Q,q,e)$, where costs are decreasing in cost reducing effort e. The cost of this effort to the specialist is $v=v(e)$ and is increasing in effort. If the specialist payment scheme depends only on q and contains no cost reimbursement, specialist utility is $U=\alpha B(Q,q,q_0)+P(q)-C(Q,q,e)-v(e)$ and the specialist has strong incentives for cost reduction, but weak incentives for quality provision. However, if the specialist payment scheme involves some cost reimbursement, then specialist utility is $U=\alpha B(Q,q,q_0)+P(q)-(1-r)C(Q,q,e)-v(e)$, and the specialist has weak incentives for cost reduction, but strong incentives for quality provision. The optimal payment scheme involves some cost reimbursement as long as $\alpha<1$.

Different specialists have different degrees of altruism, different α's. In this case, it is optimal to offer a menu of payment schemes to specialists and allow each specialist to choose the one that is best for them. These payment schemes involve a fixed payment component and a cost reimbursement component with a greater fixed payment component being associated with a smaller cost reimbursement component. Specialists who are more altruistic choose a scheme with a greater fixed component and a smaller cost reimbursement component. This has intuitive appeal as more altruistic specialists have a greater incentive to provide quality and so are given strong incentives for cost reduction.

The consensus from this theoretical literature is that quality provision above some minimum amount ensured by licensing requirements, malpractice liability, or GP referral decisions requires specialists to value the welfare of patients. If specialists value patient welfare equivalently to their own income, then a prospective payment with no cost reimbursement attains efficient provision of quality and quantity. However, if specialists value patient welfare less than their own income, then regardless of whether costs depend on cost-reducing effort or not, the optimal payment scheme involves some degree of cost reimbursement. Full-cost reimbursement is not optimal as 'too much' quality is provided and 'too little' cost-reducing effort is undertaken.

Given the specialist values patient benefit, the specialist orders 'too many' services from other health providers. This is because the specialist bears none of the cost of these services. However, if the specialist values the welfare of patients and the other health provider's profit, then efficiency can be achieved. To help clarify the exposition, let the other health provider be a hospital. Assume the hospital is paid with a prospective payment and some cost reimbursement, then hospital profit is $\Pi=P-(1-r)C_0(q_0)$. Specialist utility is $U=\alpha B(Q,q,q_0)+\Pi$. If $\alpha<1$, and if the proportion of cost reimbursement is chosen so that $r=1-\alpha$, then the specialist orders the efficient quantity of services from the hospital. Here it is not how the specialist is paid that is fundamental, but rather how the hospital is paid. Once again, some cost reimbursement is optimal.

It should be noted that little has been written about an environment in which insured patients do not know specialist quality and specialist payments are unregulated.

Empirical Evidence Concerning Specialist Payment Schemes

Although the theory above suggested that efficiency of quality provision required at least some cost reimbursement, this is not how specialists are usually paid. In practise, specialists are paid in a number of ways including salary, fee-for-service, and combinations of both salary and fee-for-service. Salary is similar to a prospective payment in that it does not depend on the quantity of services delivered to a patient. Theory suggests it will induce 'too little' quality and 'too few' services relative to efficient provision. Fee-for-service involves a payment for each service delivered so depending on the level of the fee 'too little' or 'too many' services will be provided relative to efficient provision. If the fee is set greater than the extra benefit of an additional service at efficient provision weighted by $(1-\alpha)$, then too many services relative to efficient

provision will be provided. Fee-for-service will also induce 'too little quality.'

It is useful to think of Q and q as the amount of time the specialist devotes to quality and quantity provision, respectively, where quality is increasing in the time spent gaining and retaining expertise and quantity is increasing in the time spent delivering services. That is, the specialist devotes time to two tasks. The time devoted to both tasks in total is $Q+q$. Salary provides 'weak incentives' for quality and quantity provision, whereas fee-for-service provides strong incentives for quantity. Relative to payment by salary, with payment by fee-for-service the specialist provides a higher quantity as it is strongly rewarded and a lower quality as its marginal cost of provision has increased. Because quantity is higher and quality is lower, it is not clear whether the patient is better-off.

A number of empirical studies have examined quantity choices under various payment schemes. The usual finding is that the quantity of services provided by specialists is greater under payment by fee-for-service than under payment by capitation (a prospective payment, where the specialist receives a fixed payment per patient) or salary.

The predictions of the multitasking framework have also been examined empirically in an environment in which specialist have a choice between being paid by fee-for-service or by a mixed payment scheme (salary plus fee-for-service). As the mixed scheme has a smaller fee-for-service component, multitasking theory suggests specialists who choose it will provide less services (lower quantity), but the services will be of a higher quality. Empirical evidence indicates that specialists who switched from being paid by fee-for-service to the mixed scheme reduced their volume of services by 6.2%, but increased the average time spent with patients by 3.8%. This indicates a substitution from quantity to quality as theory suggests. The welfare implications of the policy change are unclear.

Theory and the empirical evidence suggest that pay for performance initiatives should be viewed with caution if not all aspects of performance can be measured and rewarded. This is because specialists will substitute into tasks that are measurable and rewarded and out of task that are not measurable and not rewarded. In such an environment 'weak incentives' such as those provided by salary or a mixed payment schemes can be optimal.

Specialist and Patient Selection

If specialists are remunerated through a prospective payment, they have an incentive to refuse treatment to patients who require many services as these services are costly to provide and these costs are not reimbursed. That is, specialists have an incentive to 'dump' high-cost patients. In addition, specialists have an incentive to attract low-cost patients and they do this by overproviding services. That is, specialists have an incentive to 'cream skim' low-costs patients.

In addition to specialists providing services to patients, specialists provide patients to hospitals and order hospital services for these patients. If hospitals are paid full-cost reimbursement plus a profit margin per service, then they have an incentive to employ specialists that value patient welfare highly (a high α) as they will order many services for their patients. However, if hospitals are paid a prospective payment, then they have an incentive to employ specialists who do not value patient welfare highly. Therefore, the manner in which hospitals are paid has implications for the type of specialists hospitals employ.

It turns out that the way specialists are paid also has implications for the type of specialists hospitals employ and the type of patients they service. Assume there are two types of patients who differ in their length of stay in hospital. There are two types of hospitals, private and public. Private hospital profit increases with length of stay though at a decreasing rate, that is, the second day a patient stays in hospital generates less profit than the first day. Therefore, the private hospital makes more profit with many short stays than with fewer long stays and so prefers to treat short length of stay patients. The public hospital is benevolent and is indifferent between the type of patients it treats. All specialists value income, but differ in their preference for fairness, where fairness is defined as treating all patients equally regardless of type.

In this setting, if private hospitals pay specialists by fee-for-service and public hospitals pay specialists by salary, then (1) specialists who place a relatively high value on income work in private hospitals and treat short stay patients as this maximizes the number of patients they treat, maximizes their utility, and also maximizes the profit of the private hospital and (2) specialists who place a relatively high value on fairness work in public hospitals as this maximizes their utility and the utility of the benevolent public hospital. As healthy short stay patients are treated in private hospitals, public hospitals treat unhealthy long stay patients. Hospitals have selected the type of specialists that they employ through the payment scheme offered and this payment scheme induces these specialists to select the type of patient the hospital prefers. The result is consistent with the empirical observation that specialist physicians are paid by either salary or fee-for-service, with salary being more common in the public sector than the private sector.

See also: Competition on the Hospital Sector. Efficiency and Equity in Health: Philosophical Considerations. Income Gap across Physician Specialties in the USA. Moral Hazard. Physician-Induced Demand. Primary Care, Gatekeeping, and Incentives. Quality Reporting and Demand. Risk Selection and Risk Adjustment

Further Reading

Brekke, K. R., Nuscheler, R. and Straume, O. R. (2007). Gatekeeping in health care. *Journal of Health Economics* **26**, 149–170.

Chalkley, M. and Malcomson, J. M. (1998a). Contracting for health services with unmonitored quality. *Economic Journal* **108**, 1093–1110.

Chalkley, M. and Malcomson, J. M. (1998b). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* **17**, 1–19.

Chalkley, J. M. and Malcomson M. (2000). Government purchasing of health services. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, vol. 1, ch. 15, pp. 847–890. Amsterdam: Elsevier.

Cooper, R. A. and Aiken, L. H. (2001). Human inputs: The healthcare workforce and medical markets. *Journal of Health Politics, Policy and Law* **26**, 925–938.

Dumont, E., Fortin, B., Jacquemet, N. and Shearer, B. (2008). Physicians' multitasking and incentives: Empirical evidence from a natural experiment. *Journal of Health Economics* **27**, 1436–1450.
Ellis, R. P. (1998). Creaming, skimping, and dumping: Provider competition on the intensive and extensive margins. *Journal of Health Economics* **17**, 537–555.
Ellis, R. P. and McGuire, G. (1986). Provider behaviour under prospective reimbursement. *Journal of Health Economics* **5**, 129–151.
Gagne, R. and Leger, P. T. (2005). Determinants of physicians' decisions to specialize. *Health Economics* **14**, 721–735.
Gawande, A. (2002). *Complications: A surgeon's notes on an imperfect science*. New York: Henry Holt and Company.
Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* **7**, 24–52, (Special Issue, January).
Jack, W. (2005). Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* **24**, 73–93.
Ma, C.-T. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy* **3**, 93–112.
Marinosos, B. G. and Jelovac, I. (2003). GPs' payment contracts and their referral practice. *Journal of Health Economics* **22**, 617–635.
Newhouse, J. P. (2002). *Pricing the priceless: A health care conundrum*. Cambridge, MA: MIT Press.
Nicholson, S. (2002). Physician specialty choice under uncertainty. *Journal of Law and Economics* **20**, 816–847.
Sage, W. S. (2002). Putting the patient in patient safety. *Journal of the American Medical Association* **287**, 3003–3005.
Simoens, S. and Giuffrida, A. (2004). The impact of physician payment methods on raising the efficiency of the healthcare system. *Applied Health Economics and Health Policy* **3**, 39–46.
Shafrin, J. (2010). Operating on commission: Analyzing how physicians financial incentives affect surgery rates. *Health Economics* **19**, 562–580.
Sloan, F. A., Mergenhagen, P. M., Burfield, W. B., Bovbjerg, R. R. and Hassan, M. (1989). Medical malpractice experience of physicians: Predictable or haphazard? *Journal of the American Medical Association* **262**(23), 3291–3297.
Thornton, J. (2000). Physician choice of medical specialty: Do economic incentives matter? *Applied Economics* **32**, 1419–1428.
Weiler, C., Hiatt, H., Newhouse, P., *et al.* (1993). *A measure of malpractice: Medical injury, malpractice litigation, and patient compensation*. Cambridge: Harvard University Press.
Wright, J. (2007). Specialist payment schemes and patient selection in private and public hospitals. *Journal of Health Economics* **26**, 1014–1026.

Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies

H Haji Ali Afzali and J Karnon, The University of Adelaide, Adelaide, SA, Australia

Introduction

In countries such as Australia, the UK, and Canada, public funding decisions for health technologies (e.g., pharmaceuticals) are based predominantly on cost-effectiveness data. National reimbursement bodies in these countries, such as the Pharmaceutical Benefits Advisory Committee in Australia and National Institute for Health and Clinical Excellence in the UK, provide guidance to the government, which form the basis of decisions around public funding of new health technologies. Most often, this process involves using decision analytic models to inform reimbursement approval at the national level.

Decision analytic models are now an expected part of economic evaluations and are used to synthesize data from a variety of sources, link intermediate outcomes to final outcomes (e.g., quality-adjusted life-years), and extrapolate beyond the data observed in a clinical trial.

The process of developing a model requires three key choices to be made regarding (1) model structure (e.g., health states included in the model, transitions between them, and the choice of a modeling technique), (2) analytical methods such as the perspective taken (e.g., government and society) and the discount rate assumed, and (3) model input values (e.g., incidence of disease and costs of treatment). Recommendations to address uncertainty around the choice of analytical methods and model parameterization (values assigned to the model inputs) (i.e., (2) and (3) above) are generally well established and continually refined in the guidelines developed by national reimbursement bodies. Although the impact of uncertainty around the choice of model structure and making incorrect structural assumptions on model outputs, and hence on funding decisions, is acknowledged, relatively little attention has been paid to address these issues in guidelines.

The article focuses on the processes required to specify a model structure and on the uncertainty specifically arising from the choice of a model structure (i.e., structural uncertainty). The specification of a model structure involves the choice of health states or events to include in the model, and the relationships to be represented between those states.

There are a range of issues to be addressed; firstly defining an appropriate model structure, and subsequent considerations around the feasibility or applicability of the appropriate structure. Key issues include the transparency of the model to end users and the potential effects of model complexity on the process of establishing the internal and external validity of the model. In some cases, model structure may be amended on the basis of the time available to implement, validate, and analyze a model, and/or the data available to populate a model.

Another key part in the development of a model is the choice of modeling techniques, which provide an implementation framework that is used to implement a defined model structure. Inappropriate choice of modeling techniques may influence the outputs of a decision analytic model, and hence needs to be taken into account.

Decision trees are not generally applicable to the process of extrapolation, and so the major area of choice for models that estimate long-term costs and benefits is around the use of cohort-based models (predominantly cohort state-transition models), and individual-based models (generally, either state-transition models or discrete event simulation (DES)). Individual sampling models allow attributes to be assigned to patients that reflect baseline characteristics of patients and/or events experienced within the model. Such characteristics can only be represented in a cohort model by increasing the number of health states included in the model structure. Individual-based models require more analysis time as individuals rather than cohorts are run through the model.

The preceding elements of model-based studies represent the structural features of a decision analytic model. Although concerns have been raised regarding assumptions incorporated into model structures, there is a lack of clarity around the choice of model structure.

The article unfolds as follows: The first section outlines issues around the choice of an appropriate conceptual framework. This framework should reflect the natural history of the condition under study, and defines the states/events to be represented, the relationships between them and the effect of patient characteristics on the probability and timing of events. The second section discusses the development of an appropriate modeling technique (i.e., the choice of an appropriate implementation framework). The third section briefly provides a guide to the terminology used in defining different types of uncertainty around decision analytic models with a focus on structural uncertainty. Then the methods that can be used to deal with structural uncertainty are explored. The concept of reference models and their potential benefits are discussed in section five. Thereafter, the authors illustrate the application of the proposed framework for defining an appropriate model structure, taking major depressive disorder as a case study. Conclusions are formed in the final section.

Choice of an Appropriate Model Structure: Conceptual Framework

All valid models are based on an appropriate conceptual framework. The specification of an appropriate conceptual framework involves two key features, that is, the structural assumptions that inform the choice of health states/events to include in the model, and the relationships to be represented between those states (i.e., transitions between clinical events). The International Society for Pharmacoeconomics and

 doi:10.1016/B978-0-12-375678-7.01401-2

Outcomes Research – The Society for Medical Decision Making Joint Modeling Good Research Practices Task Force reports on the conceptualization of a model provides guidance for the development of an appropriate model structure.

A realistic model structure should comprise key clinically relevant events relating to the natural history of the condition being evaluated. To develop an appropriate conceptual framework, a thorough review of the clinical literature related to the condition under study should be undertaken. This review will document the progression of the condition and summarize a set of main clinical events and their interrelationships, as well as relevant patient's attributes that influence disease pathways and/or response to treatment. A review of the existing models should also be undertaken to inform the importance of identified states/events from the economic perspective. The clinical events identified will be disaggregated where there are likely to be important differences between events with respect to expected disease progression, associated costs, or associated outcomes (e.g., quality of life). If the current evidence is conflicting, more than one plausible model framework can be proposed.

There are additional issues to consider during the specification of a model structure. The choice of model structure should balance the potential value of additional model complexity, that is, increasing the likelihood of identifying important differences in the costs and benefits of the alternative health care interventions being compared, against the potential for reduced transparency, and ease of implementation, which may affect internal validity (i.e., the likelihood of undetected errors in the model). Another issue to consider is the availability of data to populate and externally validate the model (e.g., the extent to which predicted model outputs replicate the observed data, or correspond with outputs from other models in the same area). These issues are related, and all concern the credibility of model outputs.

Transparency and Validity of the Framework

Transparency refers to the ability of the end user (or a delegate of the end user, i.e., an independent reviewer) to understand and assess the implemented model. Where a specific end user is identified, the importance of this issue should be established in consultation with the end user.

The importance of model transparency may be ameliorated if the analyst can clearly demonstrate the face validity of the model. This may take the form of an expert review of the implemented model by an analyst who was not directly involved in the implementation of the model, or who was independent of the entire evaluation. Inputs from experts will be sought as to whether the proposed model structure(s) sufficiently reflects the relevant disease process and disease management pathways. Alternatively, a set of analytic checks may be defined that demonstrate the internal accuracy of the model, for example, using extreme parameter values. In defining the final model structure, the analyst should consider the extent to which they will be able to meet transparency criteria and/or demonstrate the validity of the model.

More complex model structures generally require the estimation of a greater number of input parameters, which may preclude the estimation of particular parameter values or require the use of less reliable data sources, such as data elicited from experts. Alternatively, more complex models may provide greater flexibility with respect to identifying targets for calibration or external validation.

Data Sources for Model Population

A systematic review of the literature and a thorough investigation of relevant data sources (e.g., published literature, national registries, or patient-level database) should be undertaken to identify all potentially useful sources of data, to which appropriate analytic methods can be applied to populate and validate the model. However, if this fails to identify relevant and sufficiently valid data to inform the estimation of all the required input parameters to construct the model, an expert elicitation process can be undertaken to explore the possibility of using expert opinion to fill gaps in data.

It should not be assumed that all missing parameters can be elicited from experts, as there is evidence that experts are sometimes unable or unwilling to estimate parameters about which they are too uncertain.

Expert elicitation is subject to a range of biases; both intentional and unintentional (e.g., recall bias). Before a decision is made to use elicited parameter values, and thus maintain the preferred model structure, values elicited from experts should be validated by asking additional questions in which elicited parameter values can be compared with empirical data. Elicited parameter values can also be cross-checked, that is, comparing expert data from independent sources. It is also important to represent the certainty with which different parameters can be estimated by experts, for which established methods can be applied that also provide transparency.

Missing parameter values may also be estimated using calibration techniques, by identifying sets of input parameter values that produce model outputs that are similar to target values (i.e., observed estimates of the output parameters). Calibration is a useful process even when empirical estimates for all input parameters are available, especially for complex models with many uncertain parameters.

If specific input parameters cannot be directly estimated, and relevant calibration targets cannot be identified (i.e., targets that are part-determined by parameters for which direct data is absent), potential modifications to the model structure should be considered. Decisions around data-related structural modifications should be informed by the importance of the parameters that would be omitted from a modified model structure.

Ideally, the importance of parameters with missing values would be established by implementing and populating the originally specified model structure and testing the sensitivity of the model's outputs to extreme values for the missing parameters. The paradox of this process is that the more complex model is likely to be used regardless of the outcome of this testing process – if the model is not sensitive to these parameters then a less complex model would suffice, but the analyst has already built the more complex model.

Time Constraints

In some cases, it may also be necessary to consider the timelines for the economic evaluation. If a model is being developed to inform a national reimbursement body, which has requested results from the evaluation in a time period that precludes adequate processes for the population, validation, and analysis of a preferred model structure, it may be advisable to reduce the complexity of an ideal model structure.

In all cases where a preferred model structure is modified on the basis of any of the above considerations, the process and rationale for making such decisions should be documented and reported. Here, the 'art of modeling' may be used, whereby the modeling team use their combined clinical and analytical experience to consider and report the potential effects of any structural modifications on the estimated costs and benefits, and in particular on the differences in cost and benefits between competing interventions.

Choice of an Appropriate Modeling Technique: Implementation Framework

There are various modeling techniques that can be used for the economic evaluation of health care interventions. Full details of different types of modeling techniques are beyond the scope of this article and need not be rehearsed here. In this section, the authors briefly present the key features of the three techniques commonly being used in the literature, that is, decision trees, cohort state-transition models, and individual-based models.

Decision Trees

Decision trees (DTs) are the simplest modeling techniques and are most appropriate for modeling interventions in which the relevant events occur over a short time period. The main limitation of decision trees is their inflexibility to model decision problems, which involve recurring events and are ongoing over time. In general, DTs are constructed with three types of nodes, namely decision nodes, chance nodes, and terminal nodes. A simple illustrative decision tree is presented in Figure 1. The tree flows from left to right starting with a decision node (square) representing the initial policy decision (alternative interventions in Figure 1). Each management strategy is then followed by chance nodes (circles) representing uncertain events (i.e., 'disease free' or 'dead' in Figure 1), which will have probabilities attached to them. Finally, endpoints of DTs are represented by a terminal node (triangle) at the right of the tree. The outcome measures (e.g., utility value) are generally attached to these endpoints. Costs, however, are attached to events within the tree, as well as to endpoints. The expected values (costs and effectiveness) associated with each branch are estimated by 'averaging out' and 'folding back' the tree from right to left.

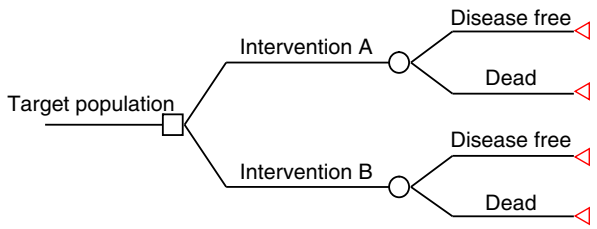


Figure 1 An illustrative example of a decision tree.

Cohort State-Transition Models

Cohort state-transition models (CSTM) are the standard technique used to model the economic impact of health care interventions over time. Using these models to capture long-term costs and benefits (e.g., lifetime), disease progression is conceptualized in terms of a discrete set of health 'states' and the 'transitions' between them. In health technology assessment, most state-transition models are discrete-time models in which the time horizon of the analysis is split into cycles of equal length (Markov cycles).

A CSTM assumes a homogenous population cohort moving between states (or staying in the same state) in any given cycle. Transitions between states during cycles are based on a set of conditional probabilities (i.e., transition probabilities). These probabilities are conditional upon the current health state. The movement between discrete health states (such as depressive episodes or remission) continues until patients enter an absorbing state (e.g., 'death') or up to the end of the specified time horizon. Costs and outcome (e.g., utility weights representing quality of life) are attached to each health state (i.e., state rewards) and transitions (i.e., transition rewards if appropriate) in the model. Expected costs and quality-adjusted life-years (QALYs) are estimated over the time horizon of the model as the sum of the time spent in each health state multiplied by the respective cost and utility weights for each health state.

CSTMs, however, suffer from the lack of memory (i.e., Markovian assumption) in which transition probabilities are not influenced by pathways taken to a particular health state. By creating separate states in a Markov model, it is technically possible to address the above issues. However, this can result in an unwieldy number of events/states and may make model implementation, checking, and analysis difficult.

Individual-Based Models

Individual-based state-transition models

These models are able to carry histories whilst remaining a manageable size. These models include all the key features of the cohort Markov family. However, assigning relevant attributes to individuals rather than to health states within a model means that the effect of the Markovian assumption is removed. They transit individual patients through the model rather than proportions of a cohort. If and where the patient moves during cycles is determined by random numbers drawn from a uniform distribution. A large number of patients are run through the model and the mean costs and QALYs gained across all patients are estimated. The principal advantage of these models is that patients' treatment and/or disease history can be captured, and used to inform subsequent transition probabilities applied to each patient within the model.

A limitation of individual-based state-transition models, however, is that time is managed through a fixed cycle length (e.g., monthly) by which the model moves forward. This may not reflect accurately the length of time spent in certain states as patients can only experience one event within each cycle. It is possible to address this issue by selecting shorter cycles (i.e., weekly) or linking separate models each with different cycle lengths. However, it may be easier to use a more flexible individual-based technique, i.e., discrete event simulation (DES).

Discrete event simulation

DES is a very flexible model that describes the flow of individuals through the treatment system. In DES, patients can be assigned attributes such as age, gender, or disease history which are assumed to influence patient's pathways through the model.

DES can accommodate differing cycle lengths more easily and with greater accuracy than state-transition models. DES uses a stochastic process to simulate events for an individual by sampling probabilities from survival distributions. Times (to next possible events) are sampled and the earliest time represents the next event for a particular individual. These events are added to a calendar and probability distributions are updated conditionally on patient history. This means that the time spent in a particular state can be estimated exactly.

Uncertainty in Decision Analytic Models: Structural Uncertainty

Given the lack of complete information about the key aspects of a decision analytic model (e.g., choice of the health states/events or true values of costs and effects of a particular intervention in a given population), uncertainty is inherent within any model-based evaluation. Uncertainty (or sensitivity) analysis is necessary so that policy-makers can incorporate information on the accuracy of model outputs into decisions around funding of the competing interventions being evaluated, as well as decisions regarding the need for additional information.

To illustrate different types of uncertainty associated with decision analytic models, the authors present a simple state-transition depression model (Figure 2). The model presented in Figure 2 consists of three states: 'well,' 'depression,' and 'dead.' The aim is to estimate differences in the time spent in each state by patients receiving alternative technologies, over a defined time horizon (e.g., patients' lifetime). Costs and outcomes are then attached to the time spent in each state to estimate the costs and benefits of alternative technologies. In state-transition models, time progresses in equal increments (e.g., monthly) known as cycles. The arrows represent possible transitions between states, for which transition probabilities are estimated. Relevant attributes (which influence disease pathways or response to treatment) can be assigned to patients, and updated during the course of running the model. For example, the probability of recurrence (i.e.,'well' to 'depression') increases with the number of previous depressive episodes that patients experienced.

Three broad forms of uncertainty have been distinguished: parameter, methodological, and structural. Parameter uncertainty concerns the uncertainty around the true value of a given parameter within the model (e.g., probabilities of moving between 'well' to 'depression'). Methodological uncertainty relates to the choice of analytic methods such as the perspective taken (e.g., society and government) with an impact on, for example, the process of identifying resource items. Issues around parameter and methodological uncertainties are generally dealt with, for example, by using probabilistic sensitivity analysis and by prescribing a 'reference case,' respectively. Structural uncertainty arises from the assumptions imposed by the modeling framework and refers to the structural features of the chosen model, that is, the choice of clinical events represented in a model (e.g., adding a 'partial response' state to the model presented in Figure 2), and the possible transitions between them.

It is recognized that the choice of model structure can lead to different results, hence different reimbursement decisions. Although concerns have been raised regarding assumptions incorporated into model structures, and that structural uncertainty may have a greater impact on the model's results than other sources of uncertainty, relatively little attention has been paid to the representation of structural uncertainty. Different ways to address structural uncertainty are discussed in Section 'Uncertainty in Decision Analytic Model.'

Addressing Structural Uncertainty

Like uncertainty around methodological issues and the value of input parameters, structural uncertainty cannot be eliminated, and needs to be handled appropriately. Different approaches are used to characterize structural uncertainty.

The simplest approach to representing structural uncertainty is to implement and analyze a range of alternative model structures, and to present the results from each model as scenario analyzes. However, this approach puts the burden of assessing the relative credibility of the alternative structures on the end user, which may result in only a superficial and subjective assessment that will not reflect the full value to the

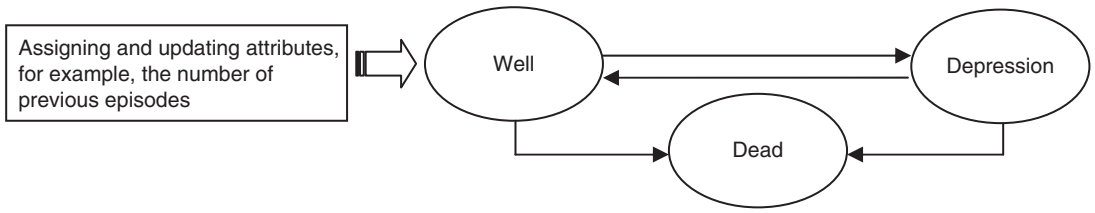


Figure 2 A simple state-transition depression model.

decision-making process of implementing multiple model structures.

For the characterization of structural uncertainty, two techniques have been identified: model selection and model averaging. Both techniques involve developing a set of plausible models using different structural assumptions.

The 'model selection' approach ranks alternative models according to some measure of prediction performance or goodness of fit, and selecting the single model that maximizes that particular criterion. This approach does not represent structural uncertainty, rather it identifies the model that is believed to be the most likely to be the correct model. The effect of selecting a single model will depend on the relevance of the criterion used to select the model and the magnitude of the difference in the predictive power of the alternative model structures. Moreover, given the effort required to build and analyze a range of alternative model structures, it would appear wasteful to report results based only on a single model when there is a likelihood that the other models could be a more correct model.

'Model averaging' methods assess structural uncertainty by assigning weights to a range of alternative model structures according to some measure of model adequacy. Weighted model outputs are then estimated across all model structures included in the model averaging process. The weightings should reflect the probability that each model is the correct model, based on some measure of predictive power (e.g., Akaike's information criterion). The main limitation of the quantitative approach to model averaging is the necessary restriction of the process to represent the elements of structural uncertainty that are linked to outputs for which observed data are available.

Other model averaging applications have elicited weights to be applied to each of the defined model structures. Here, the analyst must consider the potential for biased responses and superficial assessment of the relative merits of the alternative models.

Alternatively, a discrepancy approach has been applied to represent the uncertainty around an implemented model structure. Unlike model averaging, the discrepancy approach does not assess the adequacy of the model structure in relation to alternative structures. Rather, the joint effects of structural and parameter uncertainty are estimated, i.e., the estimated distribution of the costs and benefits of the competing interventions reflects structural and parameter uncertainty.

The application of the discrepancy approach requires the identification of points of discrepancy (i.e., structural uncertainty) within a model structure, and the subjective estimation of the magnitude and variance of the discrepancy between predicted outputs from the model using true values for each input parameter and the true values of the predicted parameters.

Improving the Accuracy and Consistency of Model-Based Evaluations: The Case for Reference Models

One of the main consequences of structural uncertainty is that alternative economic evaluations for a specific disease use alternative model structures (e.g., alternative model structures to evaluate different pharmaceuticals for treating depression). In addition to the structural uncertainties around each model structure, indirect comparisons of the cost-effectiveness of these interventions are further hindered by the diversity in model structure. Using different model structures in evaluations of alternative technologies for the same condition in submissions to national reimbursement bodies can lead to inconsistent public funding decisions because changes in model structure and analysis can produce substantially different results. This increases the likelihood of incorrectly identifying a new technology as being cost-effective (and vice versa).

There is potential for strategic behavior when defining a model structure, by focusing on aspects of the disease process that are targeted by particular technologies, for example, if drug A reduces risk of event X but not event Y and event Y has significant costs and/or quality of life effects, if the model structure represents only X then important aspects of the disease are not included and the results are biased.

To address these concerns, there is a need for a detailed and transparent framework for developing an appropriate common model structure (reference models) for specific diseases (e.g., depression, colorectal cancer, etc.) for economic evaluations to inform public funding decisions. The structure of a reference model defines the clinical events to be represented, the relationships between the events, and the effect of patient characteristics on the probability and timing of events. The model should accurately represent both the knowledge and uncertainty about states/events relating to the disease progression on the basis of the best available evidence. Reference models can be applied to a wide set of interventions for a specific disease (e.g., drugs and procedures that may target alternative mechanisms or stages of disease). For example, a reference model for depression could be used to estimate the costs and benefits of competing antidepressants, as well as evaluating the costs and benefits of alternative models of care (e.g., usual care vs. enhanced care).

By reflecting the natural history of the condition under study more accurately, reference models can improve the accuracy, comparability, and transparency of public funding decisions for new health technologies. However, we emphasize that reference models are not intended to replace structural sensitivity analyzes, and approaches to address structural uncertainty might be still required if there is insufficient evidence (or conflicting evidence) to support an appropriate model structure.

A national reimbursement body could commission the development of reference models for key disease areas, which are then developed according to best practice (as outlined in the preceding sections). The resulting models could be subjected to a thorough process of structural uncertainty analysis, and would ultimately provide a comprehensive and unbiased representation of the disease, including all important clinical and economic aspects (e.g., costs and quality of life effects), which may affect the long-term estimation of costs and benefits of the alternative health care interventions being compared. In terms of the feasibility of using reference models, one of the possibilities will be, for example, to provide the reference model to any industry applicant intending to submit a technology in the relevant disease area by national

reimbursement bodies. Applicants could update the inputs used to populate the model, but would be expected to revalidate the model. Likewise, applicants would be free to use an alternative model structure, but it would have to be fully justified. The outputs of alternative models can be compared against the reference model, which provides a basis for confirming any claimed advantages of new model structures.

Case Study: Major Depressive Disorder

After reviewing the course of depression, this section discusses the issues surrounding structural assumptions used to inform a preferred model structure and the choice of modeling techniques, using examples from published model-based economic evolutions of interventions from the treatment of depression.

Background to the Course of Depression

Major depressive disorder (referred to as depression henceforth) is the most common mental health disorder. Depression is associated with a considerable functional impairment, morbidity, and premature mortality. It is increasingly recognized as a chronic disease characterized by multiple acute episodes/relapses.

A common set of terminology describing different stages of disease progression has been proposed. 'Response' represents a significant reduction (50% or more reduction from baseline scale scores) in depressive symptoms. 'Remission' represents a period during which the patient is either symptom free or has no more than minimal symptoms. 'Recovery' is defined as an extended asymptomatic phase, which lasts for more than 6 months. 'Relapse' is a flare up of the depressive episode, which occurs after remission, whereas recurrence is a new depressive episode that occurs after recovery.

Chronic depression is one of the clinically meaningful structural aspects of depression and is defined as a persistent depressive episode, which is continuously present for at least 2 years, or an incomplete remission between episodes with a total duration of illness of at least 2 years. High-risk patients and those who have not responded adequately to outpatient treatment can be admitted to hospitals. Finally, all patients can die, with an increased risk of death due to suicide while in a depressive episode.

Patients with depression may be treated with different classes of antidepressant drugs, including selective serotonin reuptake inhibitors and serotonin norepinephrine reuptake inhibitors. After a successful short-term treatment, antidepressants need to be continued for a period to prevent relapse (continuation therapy) and to prevent recurrence (maintenance therapy). Psychotherapy or more commonly a combination of pharmacotherapy and psychotherapy are other options in treating patients with depression.

Choice of an Appropriate Conceptual Framework

The choice of health states/events included in the model is considered as a key part in developing decision analytic models. An appropriate model structure should reflect a number of key clinically relevant health states relating to the course of the disease under study. Over recent years, various model structures were used in depression studies.

Some model-based studies of depression included only response as a primary outcome measure in their model. This model structure is not appropriate as it is likely to bias results in favor of treatments with higher rates of response, which may be the more or less effective treatment overall.

In some studies, the possibility of relapse (i.e., remission to depressive episodes) and/or recurrence (i.e., recovery to depressive episodes) was not considered. By excluding these clinical events, studies implicitly assume that the incidence of such is unlikely to differ between competing management strategies. As a recurring illness, this simplification in the model structure in depression studies will favor treatments with higher remission rates. A more appropriate model structure is to represent both levels of treatment success (i.e., response and remission) and treatment failure (i.e., relapse and recurrence).

Hospital admission for high-risk patients or those who have not responded adequately to outpatient treatment is another relevant clinical event. A realistic model of depression should represent this event as it represents clinical practice more accurately.

'Chronic depression' is another relevant event that is important from both clinical and economic perspectives. Approximately 20% of patients with acute episodic illness will develop a chronic course of depression. It has been noted that chronic depression, compared with nonchronic depression, is associated with increased health care utilization, more suicide attempts, and greater functional impairment requiring a longer duration of treatment.

Finally, given a baseline increased risk of suicide at rates 10–20 times above general population rates, the omission of such a state from a depression model will likely bias the model in favor of less effective interventions.

Accepting the above course of disease, the model structure and relationships between events are presented in Figure 3. In this model, depressive episodes represent relapse or recurrence.

Choice of an Appropriate Computational Framework

Some studies of depression used DTs to evaluate costs and effects of alternative management strategies. Given that depression is a chronic illness with recurrent episodes, the main limitation of DTs is their inflexibility to model long-term events. The likely impact of the choice of a DT with a short time horizon on the cost-effectiveness results, and hence policy decisions, will vary according to the short-term cost-effectiveness results. If a more effective treatment (i.e., fewer deaths and/or more remission) can demonstrate cost-effectiveness within a short time horizon, it may be reasonable to expect that this result would not change over an extended time horizon. The likely magnitude of potentially underestimated cost differences should be considered, but a DT may be appropriate in this scenario. Where the more effective treatment is not shown to be cost-effective using a DT, it may be possible to imply the likely cost-effectiveness over a longer

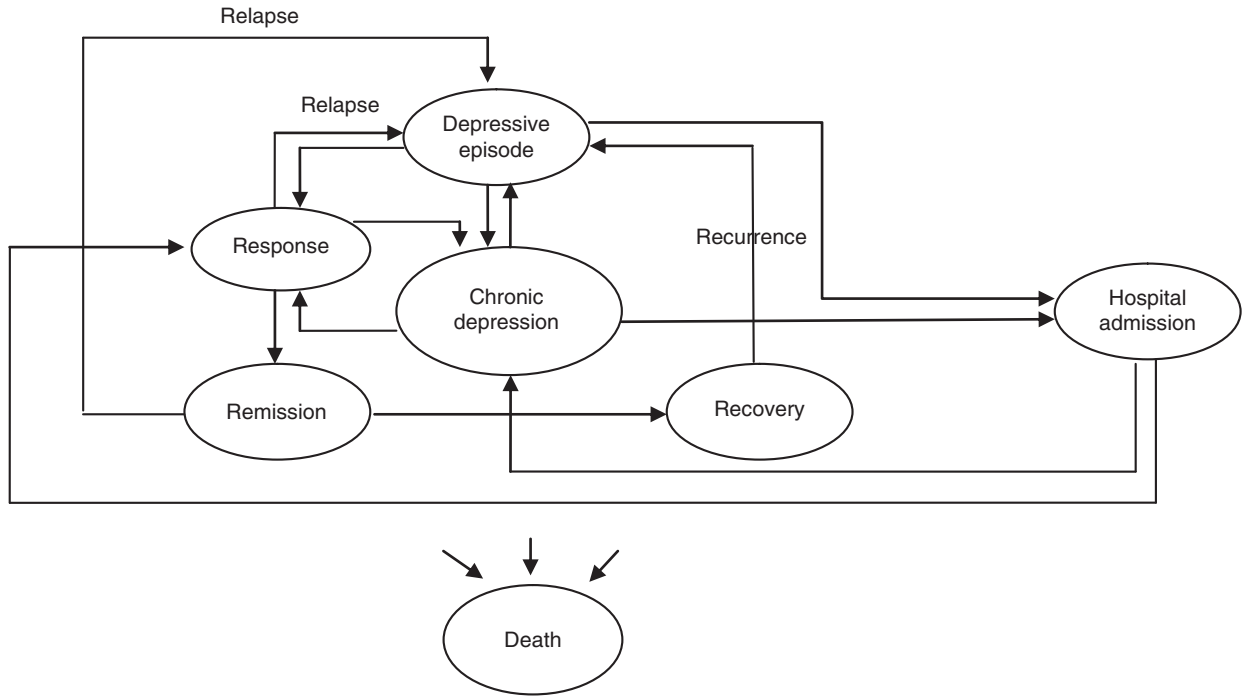


Figure 3 The conceptual framework. (a) Bubbles show the states/events included in the model, and arrows indicate the relationships between the events. (b) There is always a possibility of dying (absorbing state).

time horizon, though a modeled extrapolation provides a more explicit and transparent process of estimation.

By including recurring events and breaking up a decision tree to sequential time periods, it is technically possible to capture recurring events over longer time periods. This, however, progressively increases the size of the model and the use of an alternative modeling technique might be preferred.

To address the above issue, some model-based studies of depression employed CSTMs. These models allow for an evaluation of recurring events over a longer period of time, for example, multiple depressive and recovery episodes. However, the lack of memory (i.e., Markovian assumption) is considered as a major limitation of these models. In the case of depression, the above assumption means that an individual in the 'full remission' state with only one previous depressive episode has the same probability of experiencing 'recurrence' as a patient with a history of multiple depressive episodes. This, however, does not appropriately represent the natural history of depression as the risk of recurrence has been found to increase with the number of previous depressive episodes. One way to overcome this limitation is to create separate states to represent differing numbers of previous depressive episodes, for example, 'remission, one previous episode,' 'remission, two previous episodes,' etc. As implied, this process can soon result in an unwieldy number of health states, and may make model implementation, checking, and analysis difficult.

By modeling individual patient pathways, individual-based state-transition models are able to carry patient histories whilst remaining a manageable size.

Assigning attributes such as the severity of episodes to individuals rather than to health states within a model means that the effect of the Markovian assumption is removed. Individual-based models then facilitate the estimation of probability and timing of experiencing relevant events as a function of patient-recorded attributes (e.g., age, gender, severity, comorbidity, and psychotherapy as attributes). By updating attributes, probabilities can change over time while the model is running.

One practical limitation associated with the use of individual-based state-transition models is the need to define a fixed cycle length by which the model moves forward in time. The cycle length should represent a clinically relevant time span. Modeling studies of depression that use a fixed monthly cycle length may not reflect accurately the length of time spent in certain states as patients can only experience one event within each cycle. It is possible to address this limitation by selecting shorter cycles (i.e., weekly) or linking separate models each with different cycle lengths. However, it may be easier to use a more flexible modeling technique, that is, DES. No model-based studies of depression using DES were located.

DES can accommodate differing cycle lengths more easily and with greater accuracy than state-transition models. In a DES, time does not move forward in cycles, but rather with respect to the sampled timing of events, for example, if a patient relapses after 2 weeks, the model moves forward from time zero (treatment initiation) to time 14 days. If the next event (e.g., remission) occurs 2 months later, then the DES would move forward directly to that time point. Using survival distributions, times (to next possible events) are sampled, and the earliest time represents the next event for a particular

individual. This means that the time spent in a particular state can be estimated exactly.

Considering disease characteristics, the authors argue that DES is the appropriate modeling technique for depression studies to project life-time benefits and resource costs of management options in patients with depression.

Data Sources

A variety of data sources are used to derive clinical parameters.

Clinical trials and observational studies

Using data derived from clinical trials or observational studies, it is widely accepted that 'response' is defined as at least a 50% improvement in baseline scores recorded by the most cited assessment tools, i.e., the Hamilton Depression Rating Scale (HAM-D). HAM-D is considered as the gold standard to measure the severity of depression and is used in the majority of clinical trials of depression. 'Remission' is mostly defined as a score of 7 or less on the 17-item HAM-D. The occurrence of 'depressive episodes' is commonly defined when patients meet the criteria identified by DSM IV.

Retrospective databases

Clinical outcome measures using standard assessment tools are not typically recorded in data sources such as claims databases and medical notes. Thus, proxy measures such as treatment changes should be defined. For example, relapse can be defined as a subsequent antidepressant prescription less than 6 months after the antidepressant stop date. Another form of observational data is claims databases (e.g., US managed care database).

Expert opinion

Expert opinion is another source of data, which was used in a few studies to estimate probabilities of clinical inputs such as dropout rates, reasons for discontinuation, and nondrug-specific events such as relapse rate after discontinuation. An elicitation method is intended to link an expert's underlying opinions to an expression of these in a statistical form, and is an appropriate technique to fill gaps in data where no published information is available. However, it is important to have transparent criteria for the selection of participants, and to recognize the potential for biased estimates that may be in favor of the participant personal experiences (e.g., more frequent patients with the most severe symptoms) and/or preferences.

Conclusions

To date, relatively little attention has been paid to the processes and issues around the specification of appropriate decision analytic model structures, but better specified decision analytic models will contribute to more consistent and better informed decisions for the allocation of limited resources.

This entry has addressed issues around the appropriate choice of model structure for the purposes of economic evaluations of health care technologies, including discussions around the choice of appropriate modeling technique, representation of structural uncertainty, and the potential benefits of the development of disease-specific reference models. The disease area of depression was used to illustrate the issues involved in the process of specifying an appropriate model structure.

See also: Adoption of New Technologies, Using Economic Evaluation. Information Analysis, Value of. Searching and Reviewing Nonclinical Evidence for Economic Evaluation. Statistical Issues in Economic Evaluations. Synthesizing Clinical Evidence for Economic Evaluation

Further Reading

Barton, P., Bryan, S. and Robinson, S. (2004). Modelling in the economic evaluation of health care: Selecting the appropriate approach. *Journal of Health Services Research and Policy* **9**, 110–118.

Bojke, L., Claxton, K., Sculpher, M., et al. (2009). Characterizing structural uncertainty in decision analytic models: A review and application of methods. *Value in Health* **12**, 739–749.

Brennan, A., Chick, S. and Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies. *Health Economics* **15**, 1295–1310.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A., et al. (2012). Model parameter estimation and uncertainty analysis: A report of the ISPOR-SMDM Modelling Good Research Practices Task Force-6. *Medical Decision Making* **32**, 722–732.

Brisson, M. and Edmunds, W. J. (2006). Impact of model, methodological, and parameter uncertainty in the economic analysis of vaccination programs. *Medical Decision Making* **26**, 434–446.

Eddy, D. M., Hollingworth, W., Caro, J. J., et al. (2012). Model transparency and validation: A report of the ISPOR-SMDM Modelling Good Research Practices Task Force–7. *Medical Decision Making* **32**, 733–743.

Haji Ali Afzali, H. and Karnon, J. (2011). Addressing the challenge for well informed and consistent reimbursement decisions: The case for reference models. *Pharmacoeconomics* **29**, 823–825.

Haji Ali Afzali, H., Karnon, J. and Gray, J. (2012). A critical review of model-based economic studies of depression: Modelling techniques, model structure and data sources. *Pharmacoeconomics* **30**, 461–482.

Jackson, C., Sharples, L. D. and Thompson, S. G. (2010). Structural and parameter uncertainty in Bayesian cost-effectiveness models. *Applied Statistics* **59**, 233–253.

Karnon, J. and Brown, J. (1998). Selecting a decision model for economic evaluation: A case study and review. *Health Care Management Science* **1**, 133–140.

Le Lay, A., Despiegel, N., Francois, C., et al. (2006). Can discrete event simulation be of use in modelling major depression? *Cost effectiveness and resource allocation* **4**, 19.

Roberts, M., Russell, L. B., Paltiel, D., et al. (2012). Conceptualizing a model: A report of the ISPOR-SMDM Modelling Good Research Practices Task Force-2. *Value in Health* **15**, 804–811.

Russell, L. B. (2005). Comparing model structures in cost-effectiveness analysis. *Medical Decision Making* **25**, 485–486.

State Insurance Mandates in the USA

MA Morrisey, University of Alabama at Birmingham, Birmingham, AL, USA

Introduction

Insurance markets in the US traditionally have been regulated at the state level. This tradition was reinforced by the 1945 McCarran–Ferguson Act, which exempted the business of insurance from federal antitrust oversight as long as the individual states regulated the insurance business. Much of the early health insurance regulation related to reserve requirements and sales practices although the states did impose premium taxes. The number and types of mandates proliferated in the past 40 years.

Insurance mandates are defined as state laws that require a health insurer to include specific categories of individuals, providers, or services within the scope of coverage provided. Less restrictive laws only require that the insurer offer specific coverages to purchasers. Some analysts also have included laws affecting specific types of insurers, particularly managed care plans, as insurance mandates as well. The early mandates of the 1950s tended to expand the cohort of covered persons to include newborns and handicapped children. Thus, if a policy was to be sold in the state it must include coverage for newborns.

Prevalence of State Insurance Mandates

Most mandates apply to employer-sponsored health insurance coverage, although nongroup mandates also exist. There were few statutes until the 1970s when the number of laws began to proliferate. Although they only consider provider and service mandates, Laugesen *et al.* (2006) reported that of nearly 1500 state mandates enacted between 1949 and 2002, 12% were enacted in the 1970s, 25% in the 1980s, 39% in the 1990s, and another 16% in the first 3 years of the 2000s.

The Council for Affordable Health Insurance annually compiles a listing of all state health insurance mandates (Bunce and Wieske, 2010). Their 2010 edition reports the presence of 2156 mandates. The states display considerable heterogeneity in the number of mandates they enact. Idaho and Alabama have the fewest mandates with 16 and 19 provisions, respectively. Four states (Rhode Island, Maryland, Minnesota, and Texas) each have more than 60 provisions. Table 1 reports the most common mandates by category.

Federal Insurance Mandates

It is only relatively recent that the Congress has enacted health insurance mandates. The first was the 1979 Pregnancy Discrimination Act, which required that pregnancy be covered as a medical condition in most employer-sponsored plans. The 1986 Consolidated Omnibus Budget Reconciliation Act provided continuation coverage for 18–36 months for persons separated from employer-sponsored coverage. In 1996, the Health Insurance Portability and Accountability Act, (first) Mental Health Parity Act, and Newborns and Mothers' Health Protection Act were enacted. Other federal laws followed. The enactment of state and federal mandates is not independent of each other. As Laugesen *et al.* (2006) noted, 50 jurisdictions had enacted breast reconstruction legislation by the time the federal legislation was passed in 1998 and 32 states had enacted maternal and newborn minimum stay provisions before or coincident with the 1996 federal legislation. These patterns suggest a common underlying demand for the legislation and may suggest that the federal legislation is merely a reflection of extant state practice.

Rationales for Mandates

There are three rationales for the enactment of a mandate. The first is lack of knowledge. The argument is that individuals and their employer agents underestimate the value of the coverage. By requiring coverage people get the coverage they would have purchased had they been better informed. The second rationale is adverse selection. The argument holds that both low- and high-probability individuals would like to buy the

Table 1 Most prevalent health insurance mandates, by category

Health insurance mandates	*Number of states with the law*
Service mandates (1251 total mandate laws)	
Mammography screening	50
Maternity minimum stay	50
Breast reconstruction	50
Mental health parity	48
Diabetic supplies	47
Alcohol/substance abuse	46
Emergency room service	45
Provider mandates (558 total mandate laws)	
Chiropractor	44
Psychologist	44
Optometrist	41
Dentist	33
Podiatrist	33
Nurse practitioner	29
Nurse midwife	27
Individual mandates (347 total mandate laws)	
Newborn	51
Continuation employee	46
Continuation dependent	45
Adopted children	44
Disabled dependent adult	42
Conversion to nongroup coverage	41
Dependent student/adult	34

Source: Reproduced from Bunce, V. C. and Wieske, J. P. (2010). *Health insurance mandates in the states 2010.* Alexandra, VA: Council for Affordable Health Insurance. Available at: http://www.cahi.org/cahi_contents/newsroom/article.asp?id=1036 (accessed 06.06.13).

 doi:10.1016/B978-0-12-375678-7.00925-1

coverage, for, say, *in vitro* fertilization. However, when a single insurer offers the coverage those with a high probability of use are disproportionately attracted to the plan, raising plan premiums substantially. The higher premiums lead low probability users to forego coverage. Had the coverage been in all plans, the cost of insurance would have risen only slightly. Thus, mandating coverage allows low risk individuals who value the coverage to actually obtain it at premiums they were willing to pay. The third rationale is one of public choice. The argument is that advocates of a particular mandate will tend to be the providers of the particular service who petition the state legislature for statutes that expand coverage for themselves, the services they provide, and the people who are most likely to use their services. Opponents will be individuals and their employer and/or insurer agents who directly or indirectly face the costs of the law. There has been no empirical work testing the first two rationales. The public choice approach enjoyed some research interest in the 1990s with attention focused on mental health services and laws affecting the composition of managed care provider panels, all with results consistent with the public choice argument (see Jensen and Morrisey, 1999 for review).

One implication of the public choice model is that opposition should be reduced when potential opponents are exempt from the law. The Employee Retirement and Income Security Act (ERISA) of 1974 effectively made larger employers exempt from state insurance regulations. From this perspective it is not coincidental that the growth of state mandate legislation began in the 1970s when key elements of the opposition no longer benefited by expending political capital to oppose the laws.

To try to control the proliferation of mandates, half of the states have enacted mandated benefit review laws, which provide for a review of factors such as the cost and social impacts, medical efficacy, and quality of care before the legislature votes on a prospective mandate. There has been no evaluation of the effectiveness of these laws.

Economics of Employer-Sponsored Health Insurance and State Insurance Mandates

Maximizing behavior on the part of employees and employers implies that workers are paid what they are worth, i.e., their marginal revenue product. Compensation may take many forms: wages, pensions, vacations, health insurance, etc. However, adding an element to the compensation package, other things equal, necessitates taking something else out. When a mandated benefit is added to the bundle the premium will increase. If premiums increase, wages or some other form of compensation must be reduced.

If workers do not sufficiently value the new compensation package they will seek employment in firms that do not offer the new costly benefit or may choose compensation that excludes health insurance entirely. Firms may seek a legal status that exempts them from the mandate. If wages or other forms of compensation can not adjust, perhaps because of minimum wage laws, one would expect reductions in employment. In the following four sections, the empirical evidence associated with these hypotheses will be explored.

Cost of Mandates

The key issue in the chain of economic logic is that a mandate results in higher health insurance premiums. There are three issues bound up in this straightforward proposition. The first question is whether a particular benefit raises the health insurance premium of a firm that now must provide the new coverage. Second is the question of the cost of the mandate, per se. Suppose a chiropractor coverage mandate would raise premiums by US$10 per worker per month, but employees are willing to pay US$6. The benefit cost is US$10, but the cost of the mandate is only US$4 per worker. Finally, if the firm already offers the coverage that is the subject of the mandate, then there is no cost to the mandate for the firm or its workers. The cost of the legislation is the extra burden it imposes, not necessarily the dollar cost of the benefit.

Many of the cost analyses of health insurance mandates are actuarial studies. These works draw on a distribution of the likely claims experience for a health service, such as chiropractic care. The expected claims experience per worker is the additional cost of insurance to a firm that did not offer the benefit previously. Bunce and Wieske (2010) provided ranges of actuarial costs for each of the mandates they report. Mental health parity (e.g., covering mental health illnesses equivalently to physician health illnesses) would increase premiums by 5–10%, *in vitro* fertilization by 3–5%, alcoholism/substance abuse by 1–3%, and most other coverages by less than 1%. Kominski *et al.* (2006) used an actuarial model to estimate the costs of seven proposed mandates in California. The costs of these relatively small-scale mandates were estimated to increase premiums by 0.006–0.2%. The actuarial approach overstates the cost of a mandate for two reasons. First, additional services used under the new benefit may offset, somewhat, utilization of other services that were already covered. Second, workers are likely to have been willing to pay something for the mandated coverage, and indeed, some may already have the coverage.

In principle the first problem is easy to overcome. One could estimate a firm specific hedonic premium regression that includes various benefits along with a set of control variables. The coefficients on the specific benefits reflect the premium cost of adding each benefit, given the coverage already provided by the firm. Thus, any service substitution is accounted for. Unfortunately, this is not an estimate of the cost of a mandate, per se; rather it is the net cost of the benefit to the firm (see Jensen and Morrisey, 1999 for review).

Acs *et al.* (1992) were the first to directly estimate the presence of state insurance mandates on the premiums paid by firms. Using a cross section of more than 2500 firms in 1989, they concluded that each additional mandate increased health insurance premiums for large firms by US$1.50. Although this is a direct estimate of the cost of mandates, it is not without problems. First, the count of the number of mandates in a state forces each mandate to have an equal effect. Almost certainly some mandated benefits are substantially more costly than others. Second, the estimate is likely endogenous. That is, it may be that the legislature enacted particular mandates in their state because they were already commonly offered by large employers in the state or because residents of the state were perceived to benefit from the coverage.

Wage and Benefit Adjustments to Mandates

If mandates are costly, their inclusion in employer-sponsored health insurance plans should lead to reductions in wages or other forms of compensation. This adjustment in the compensation bundle is referred to as 'compensating differentials.' There is one particularly strong analysis relating to state insurance mandates. In the mid-1990s, 23 states enacted legislation requiring that maternity services be included as a covered benefit in employer-sponsored health insurance. Gruber (1994a) examined the effects of the enactment of this law in Illinois, New York, and New Jersey compared to five states that had not enacted the legislation. His approach was to compare: the hourly wages of affected people to those unaffected, in states that enacted the law to those that did not, in the periods before and after the date of enactment. Affected people were employed married women aged 20–40 years, that is, those potentially likely to use the new benefit. Unaffected people were defined as employed single men aged 20–40 years and all employed people aged 40–60 years.

This is the so-called triple-differences model. Gruber's work is increasingly the standard approach to addressing the effects of insurance mandates. Its strength is that it controls for trends occurring over the period before and after the enactment, for differences in the states enacting and not enacting the laws, and for differences that might exist in the average productivity of people affected and not affected by the law. Gruber found that the net effect of the maternity mandate was to reduce the average hourly wages of the affected group, in the enacting states after passage of the law by 5.4% relative to unaffected groups in nonenacting states over the same period. This amount was consistent with the actuarial estimates of the cost of the coverage. This is very strong evidence that the effect of a binding mandated benefit law is to reduce other elements of the compensation bundle. In short, it implies that workers pay for much of the cost of a mandate.

Self-Insured Plans: Avoiding Mandates

The 1974 federal ERISA legislation exempts self-insured firms from state insurance regulation. One way to avoid the costs of unwanted health insurance mandates would be to obtain coverage through a self-insured plan. Virtually all large employers offer plans that are self-insured. Even a small employer could provide self-insured coverage; they would do so by buying stop-loss coverage (sometimes called reinsurance) that limits their liability once claims in total or for individual cases reach a specified threshold. Nearly 60% of insured workers were in self-insured plans in 2009.

The empirical work on the relationship between state mandates and the propensity to self-insure comes from the late 1980s and early 1990s. Jensen *et al.* (1995) examined the effects of costly mandates and other factors on the switch from conventional to self-insured coverage in the 1981–84 and 1984–87 periods. Most mandates had a positive but statistically insignificant effect on the self-insurance decision. They did find that state premium taxes and high risk pool taxes were strongly associated with switching. Subsequent work found no consistent effects of mandates on the self-insurance decision and some suggestion that self-insured status was associated with firms that had multiple locations and who may have been trying to avoid conflicting state regulation.

Mandates and Coverage Decisions

Jensen and Gabel (1992) were among the first to estimate the effect of state insurance mandates on the probability that an employer would offer insurance coverage. Using 1985 data from small employers they concluded that every additional mandate in a state reduced the probability that a firm would offer coverage by 1.5%. Surprisingly, they also found that some arguably high cost mandates such as alcohol and drug abuse treatment and continuation of coverage requirements had no statistically significant effects. Gruber (1994b) examined the effects of five costly mandates: alcoholism treatment, drug abuse treatment, mental illness, chiropractic services, and continuation of coverage using Current Population Survey (CPS) data from 1979, 1983, and 1988. He found no effects of the mandates on the probability of having coverage.

Sloan and Conover (1998) examined data from 1989 to 1994 CPS. They measured regulation as a count of the number of state mandates in effect and concluded that eliminating state mandates would reduce the number of uninsured in the state by 20% and 25%. In a recent working paper Ma (2007) updated these analyses using the 1996–2002 CPS. His focus was on two alternative sets of high cost mandates and unlike the earlier studies in this section he used state and year differences-in-differences to control for contemporaneous trends in coverage and systematic differences across states. He found no statistically significant effects of mandates on coverage decisions.

This set of work highlights two crucial problems that have tended to undermine much of the work that has looked broadly at the effects of state mandates. First, state laws differ markedly with respect to the likely costs they impose on potential purchasers and the nature of the coverage that they actually mandate. The research has been rather cavalier with measuring regulation. Efforts to aggregate state laws via counts of laws, for example, at best can only obtain an average effect of an average mandate. It seems increasingly clear that there is as yet no good empirical way to look at the overall effects of state mandates. One must look at the effects of mandates individually and provide some assurance that the laws under study are homogeneous. Second, state insurance mandates are not enacted randomly. They result from legislative actions that in turn reflect issues such as resident preferences, existing levels of coverage, and the influence of providers. Failure to account for the endogeneity of the laws is almost certainly responsible for much of the inconsistency and uncertainty surrounding a lot of the existing empirical work.

The New Generation of Mandates Research

Several mandate-specific studies have been undertaken in the past decade. They focus on the careful measurement of a particular mandate across the states and they rigorously deal with the endogeneity issue by using differences-in-differences (or sometimes triple differences) or instrumental variables.

These studies have not uniformly concluded that the laws have affected use or outcomes. However, they have given much greater attention to measuring the law and they employ much more sophisticated methods to account for the endogeneity of the legislation. Several examples are noteworthy.

Bitler and Carpenter (2011) examined the effects of mammography screening mandates. They used person-level data from the Behavioral Risk Factor Surveillance System (BRFSS) over the 1987–2000 period and concluded that the mandates account for approximately 7% of the doubling of screening observed over the study period. This conclusion resulted from a triple-difference analysis that compared screening in states that did and did not enact the mandate, before and after enactment, among women at ages that were and were not recommended for screening.

Klick and Stratmann (2007) used the 1996–2000 BRFSS data to examine the effects of mandates that cover diabetes supplies, treatment, and services on obesity among diabetics. They too used the triple-difference model (states that did and did not enact the mandate, before and after enactment, and for those with and without a self-reported diagnosis of diabetes). Their hypothesis was that the presence of the mandate lowered the cost of treatment and, therefore, provided incentives for individuals to let their health deteriorate. They conclude that diabetics in states that enacted the coverage mandates, had greater increases in body mass index (BMI) compared to changes in BMI for nondiabetics in states without mandates over the same period. They also found that the results depended on the use of the triple-difference approach. Less rigorous models suggest that the mandates lowered BMI. In related work also focusing on the moral hazard effects of insurance mandates, Klick and Stratmann (2006) examined the effects of mental health mandates (that included substance abuse treatment) on the per capita consumption of beer in the state. They argue that the mandate lowers the cost of alcohol abuse treatment, should it be needed. They explicitly model the enactment of the law using the enactment of medical malpractice damage caps, diabetes mandates, and the number of nonpsychiatrist physicians per capita as instruments. Their estimates suggest that the mandate enactment was associated with an additional consumption of two cases of beer per capita.

In less controversial work, Liu *et al.* (2004) examined the effect of the so-called drive-by delivery laws. These provisions were enacted to prevent insurers from insisting that women and newborns be discharged from a hospital too quickly after a birth. The researchers used hospital discharge abstract data from 18 states over the period 1995–98. Their differences-in-differences analysis concluded that the laws resulted in 11% longer lengths of stay for vaginally delivered newborns. The effects were larger when the law was a 'mother's decision' rather than specifying a 'physician decision.' The effects were smaller among patients who were likely to be exempt from the law due to ERISA.

Summary

State insurance mandates have proliferated since the mid-1970s. The economics of the laws suggest that they should increase insurance premiums, reduce coverage, and lead to greater self-insured status among employers. More generally, they may affect the use and outcomes of mandated services and may influence the take-up of health insurance. The empirical work has been inconsistent in its findings. This stems from measurement issues in classifying and aggregating the laws and a failure to account for the nonrandom enactment of laws in the states. The new generation of research, largely undertaken in the 2000s and building on Gruber's (1994b) seminal work on maternity mandates, has used much more sophisticated techniques and provided much greater confidence in the evaluations of how the laws have affected behavior.

See also: Health Insurance and Health. Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare. Health Insurance in the United States, History of. Private Insurance System Concerns

References

Acs, G., Winterbottom, C. and Zedlewski, S. (1992). Employers' payroll and insurances costs: Implications for play or pay employer mandates. In *Health benefits and the workforce*, pp. 195–230. Washington, DC: U.S. Department of Labor, Pension and Welfare Benefits Administration.

Bitler, M. P. and Carpenter, C. (2011). Insurance mandates and mammography. *National Bureau of Economic Research Working Paper 16669*. Cambridge, MA: National Bureau of Economic Research.

Bunce, V. C. and Wieske, J. P. (2010). *Health insurance mandates in the states 2010*. Alexandra, VA: Council for Affordable Health Insurance. Available at: http://www.cahi.org/cahi_contents/newsroom/article.aspid=1036 (accessed 06.06.13).

Gruber, J. (1994a). The incidence of mandated maternity benefits. *American Economic Review* **84**, 622–641.

Gruber, J. (1994b). State mandated benefits and employer provided health insurance. *Journal of Public Economics* **55**, 433–464.

Jensen, G. A., Cotter, K. D. and Morrisey, M. A. (1995). State insurance regulation and an employer's decision to self insure. *Journal of Risk and Insurance* **62**, 185–213.

Jenson, G. A. and Gabel, J. (1992). State mandated benefits and the small firm's decision to offer health insurance. *Journal of Regulatory Economics* **4**, 379–404.

Jensen, G. A. and Morrisey, M. A. (1999). Employer-sponsored health insurance and mandated benefit laws. *Milbank Quarterly* **77**(4), 425–459.

Klick, J. and Stratmann, T. (2006). Subsidizing addiction: Do state health insurance mandates increase alcohol consumption? *Journal of Legal Studies* **35**, 175–198.

Klick, J. and Stratmann, T. (2007). Diabetes treatments and moral hazard. *Journal of Law and Economics* **50**, 519–538.

Kominski, G. F., Ripps, J. C., Laugesen, M. J., Cosway, R. G. and Pourat, N. (2006). The California cost and coverage model: Analyses of the financial impacts of benefit mandates for the California legislature. *Health Services Research* **41**(3 Part II), 1027–1044.

Laugesen, M. J., Paul, R. R., Luft, H. S., Aubry, W. and Ganiats, T. G. (2006). A comparative analysis of mandated benefit laws, 1949–2002. *Health Services Research* **41**(3 Part II), 1081–1103.

Liu, Z. W. H., Dow and Norton, E. C. (2004). Effect of drive-through delivery laws on postpartum length of stay and hospital charges. *Journal of Health Economics* **23**(1), 129–156.

Ma, A. (2007). Another look at the effect of state mandates for health insurance benefits. *Working Paper*. Philadelphia, PA: Wharton Research Scholars Journal, University of Pennsylvania.

Sloan, F. A. and Conover, C. J. (1998). Effects of state reforms on health insurance coverage of adults. *Inquiry* **35**(3), 280–293.

Statistical Issues in Economic Evaluations

AH Briggs, University of Glasgow, Glasgow, Scotland, UK

Introduction

As health economic evaluation has become increasingly popular, so it has become much more common that individual patient data are collected alongside clinical trials. This has opened up the possibility of using statistical methods to analyze health economic data with the purposes of informing health economic evaluation.

In this entry of the encyclopedia, statistical methods for analyzing health economic data are reviewed. The second section focuses specifically on issues relating to the analysis and testing of cost data, including issues related to nonnormality of the data, censoring of cost-data and other forms of missingness. The third section deals with characterizing uncertainty in cost-effectiveness (CE) analysis, focusing on the commonly employed incremental CE ratio statistic. The fourth section focuses on the net-benefit statistic as an alternative approach to characterizing uncertainty in CE analyses. A final section offers some concluding comments and links to other articles that address similar material, in particular the increasing use of individual data analyses within CE modeling to address issues of heterogeneity in CE.

Analyzing Individual-Level Cost Data

Where health care resource information has been collected alongside randomized controlled trials, so the formal statistical testing of cost differences has become possible. Nevertheless, a number of problems plague cost data in particular. First, that cost data are often highly nonnormal, exhibiting excess zeros, skewness, and kurtosis. Second, that censoring of cost occurs, just as it does for other trial endpoints. Third, that cost data are often missing, due to the nature of the data collection process.

Non Normality

It is well recognized that statistical analysis of healthcare resource use and cost data poses a number of difficulties related to the distribution of these data: they often exhibit substantial positive skewness, can have heavy tails and often have excess zeros representing a proportion of subjects that are not users of health care resources. In medical statistics, the standard approach for handling such nonnormal data has been to use nonparametric methods, such as rank order statistics. However, in health economics it is widely accepted that it is the estimated population mean cost that is the statistic of interest to policy makers – as only the mean cost (when multiplied by the number of subjects) recovers the total cost of care for the patient group.

In a recent review of the literature on methods for the statistical modeling of health care resource and cost data, Mihaylova and colleagues distinguish two broad areas that have developed in relation to the statistical analysis of these data. In the 'randomized evaluation' field, healthcare resource use and cost data are collected alongside randomized controlled trials, in order to study the impact of interventions on average costs and test mean cost differences. These studies are used to evaluate the CE of healthcare interventions and guide treatment decisions. The 'health econometrics' field is characterized by the use of large quantities of observational data to model individual healthcare expenditures, with a view to understanding how individual characteristics influence overall costs. Observational data are vulnerable to biases in estimating effects due to nonrandom selection and confounding that are avoided in randomized experimental data.

The literature on evaluating costs for the purpose of CE analysis and that on health econometrics have developed largely independently. Mihalova and colleagues provide a review of the analytical approaches to estimating mean resource use and costs with a particular focus on mean cost differences for evaluative purposes. Although the fundamental interest relates to the raw cost scale, analysis can be performed on a different scale for the purposes of estimation providing a mechanism exists for returning to the original cost scale.

The objective of their review was to examine the state-of-the-art of statistical analysis of healthcare resource use and cost data, by identifying the methods employed, their ability to address the challenges of the data and their ease for general use. They proposed a framework to guide researchers when analysing resource use and costs in clinical trials and a summary of this framework is reproduced in Table 1.

Their review identified 12 broad categories of methods: (I) methods based on the normal distribution, (II) models based on normality following a transformation of the data, (III) single-distribution generalized linear models (GLMs), (IV) parametric models based on skewed distributions outside the GLM family, (V) models based on mixtures of parametric distributions, (VI) two-part, hurdle and Tobit models, (VII) survival methods, (VIII) nonparametric methods, (IX) methods based on truncation or trimming of data, (X) data components models, (XI) methods based on averaging across a number of models, and (XII) Markov chain methods. Their recommendations were that, firstly, simple methods are preferred in large sample sizes (in the thousands) where the near-normality of sample means is assured. Secondly, in somewhat smaller sample sizes (in the hundreds), relatively simple methods, able to deal with one or two of the data characteristics studied, may be preferable but checking sensitivity of results to assumptions is necessary. More complex approaches for the analysis of mean costs in clinical trials that take into consideration the specific features of the data might lead to gains in precision and to more informative estimates if correctly specified, but run a risk of misspecification leading to biased results. Although some more complex methods hold promise for the future, these are relatively untried in practice and as such are not currently recommended for wider applied work.

 doi:10.1016/B978-0-12-375678-7.01418-8

Table 1 Summary characteristics of the reviewed method for modelling cost and resource use data in moderate size data

Analytical approach	*Features of data*				*Features of method*				
	Skewness	*Heavy tails*	*Excess zeros*	*Multimodality*	*Testing for cost difference*	*Covariate adjustment*	*Analysis on original scale/ No need to back transform*	*Works with small samples*[a]	*Ease of implementation*
I. Methods based on the normal distribution	L	L	L	L	H	H	H	L	H
II. Methods following transformation of data	H	H	L	L	P	H	L	P	P
III. Single-distribution	H	P	L	L	H	H	H	P L[b]	H
IV. Parametric models based on skewed distribution outside the GLM family	H	P	L	L	H	H	H	P L[c]	P
V. Model based on mixture of parametric distributions	H	H	P	H	P	H	H	P	L
VI. Two-part and hurdle models	H	P H[d]	H	L	P	H	L H[d]	P L[d]	P H[d]
VII. Survival (or duration) methods:									
(i) Semiparametric Cox and parametric Weibull proportional hazards model	H	H	H	P	P	H	H	P	H
(ii) Aalen additive hazard model	H	H	L	P	P	P	H	L	L
VIII. Nonparametric methods:									
(i) Central limit theorem and Bootstrap method	P	P	L	L	H	H	H	P	H

(*Continued*)

Table 1 Continued

Analytical approach	*Features of data*				*Features of method*				
	Skewness	*Heavy tails*	*Excess zeros*	*Multimodality*	*Testing for cost difference*	*Covariate adjustment*	*Analysis on original scale/ No need to back transform*	*Works with small samples*[a]	*Ease of implementation*
(ii) Nonparametric-modified estimators based on pivotal statistic or Edge worth expansion	H	P	L	L	P	L	H	P	H
(iii) Non-parametric density approximation	H	H	H	H	P	H	H	L	P L[e]
(iv) Quantile-based smoothing	H	H	H	H	H	H	P	L	L
IX. Methods based on data trimming	L	L	L	L	L	L	H	P L[f]	H
X. Data components models	H	P	P	H	P	H	H	P L[g]	L
XI. Model averaging	H	H	P	H	P	H	H	L	L
XII. Markov chain methods	H	H	P	H	P	P	H	L	L

[a]Small sample refers to tens to a few hundreds of participants.
[b]More complex GLMs require large sample sizes, as does checking parametric modeling assumptions.
[c]Checking parametric modeling assumptions needs large sample size.
[d]The ability to model heavy tails depends on the model used in the second part. If the model used in the second part is from Category II 'Methods based on normality following a transformation of the data' back transformation will be needed. Checking modeling assumptions needs reasonable sample size. The case of implementation depends on the models used in the two parts.
[e]These approaches are not available in standard statistical software.
[f]Depends on data; in small samples opportunities to check parametric model assumption are restricted.
[g]Models beyond those relying on multivariate normality will need large data sets. More details is provided in the review templates in the web appendix at http://www.herc.ox.ac.uk/downloads/support_pub.
Abbreviations: GLM, generalized linear model; H, High applicability; L, Low applicability; P, Possible applicability.
Source: Reproduced from Mihaylova *et al.* (2011).

One of the areas that Mihaylova *et al.* (2011) ruled outside of the scope of their review was the problem of censoring. However, particularly in the field of health economic evaluation conducted alongside clinical trials, the problem of censored cost data is highly prevalent. In the past, many analyses simply ignored this issue and presented analyses that assumed the data were uncensored. Fenn *et al.* (1995) argues that such an approach would lead to biased estimates of the true cost which, when adjusted for censoring, could be substantially higher. They went on to propose that standard survival analyses could be employed using the cost-scale in order to adjust for censoring. However, although this removed some of the problem, other authors demonstrated that the approach remained biased as the cost scale and the censoring event were no longer independent, an assumption required by the standard survival analysis methods.

Two general approaches have been shown to be capable of generating unbiased estimates of censored cost data. The first, known as the Kaplan–Meier Sample Average estimator is based on estimating a mean cost function over fixed time intervals based on the cost of the at risk population, then weighting those mean costs by the proportion surviving (estimated from the Kaplan–Meier survivor function) and summing across intervals to estimate total mean cost for the follow-up period. The other approach, based on Inverse Proportion Weighting (IPW) takes the inverse of the estimated survivor function for the censoring process as a weight to apply to the remaining observed data. Thus, at a given time point, if the censoring is 50% then each remaining data point receives a weight of 2 to reflect that the cost observed must count for both the observed subjects and the additional 50% that are censored at that point. The weighted costs in each time period are then summed and averaged to obtain the mean total cost.

These approaches can be extended into a regression framework in order to handle prognostic patient characteristics. Indeed, parameterizing the cost and survivor functions would also offer a crude approach to extrapolating beyond the data to estimate total lifetime cost. Although rarely acknowledged, the approach taken to estimating unbiased cost-estimates shares a common methodology with the estimation of quality adjusted survival analyses.

Other Forms of Missingness

The problem of missing data is not new and has received much attention in the statistical literature as to the appropriate methods for handling missing data. Although in principle, missing economic data alongside clinical trials is no different to other forms of missing data; the distributional form of cost data (as presented above) may provide challenges for the analyst. Furthermore, because economic evaluation is commonly 'piggy-backed' onto clinical trials, there is a danger that economic variables will be considered less important by researchers responsible for data collection which could result in higher rates of missingness.

Missing data can arise in a number of ways. Univariate missingness occurs when a single variable in a data set is causing a problem through missing values, although the rest of the variables contain complete information. Unit nonresponse describes the situation where for some people (observations) no data are recorded for any of the variables. More common, however, is a situation of general or multivariate missingness where some, but not all of the variables will be missing for some of the subjects. Another common type of missingness is known as monotone missing data, which arises in panel or longitudinal studies, and is characterized by information being available up to a certain time point/wave but not beyond that point.

Little and Rubin (2002) outline three missing data mechanisms:

1. Missing completely at random (MCAR): If data are missing under this mechanism then it is as if random cells from the rectangular data set are not available such that the missing values bear no relation to the value of any of the variables.
2. Missing at random (MAR): Under this mechanism, missing values in the data set may depend on the value of other observed variables in the data set, but that conditional on those values the data are missing at random. The key is that the missing values do not depend on the values of unobserved variables.
3. Not missing at random (NMAR): It describes the case where missing values do depend on unobserved values.

The difference between these mechanisms is quite subtle, particularly for the first two cases of MCAR and MAR. Briggs and colleagues give the following example related to resource/cost data. Consider a questionnaire distributed to patients, in order to ascertain their use of health care resources following a particular treatment intervention, where the response rate is less than 100%. The nonresponse is MCAR if the reason for failure to complete the questionnaire was unrelated to any prognostic variables in the data set. In practice, however, such a situation is unlikely. For example, retired patients may find more time to complete and return a questionnaire than those of working age. Also, being older on average, retired patients may make more use of health care resources. If having conditioned on the age and retirement status of the patients nonresponse is random then the missing data problem is considered MAR. However, it is possible that one of the reasons for nonresponse is that patients have been admitted to hospital. Now the missing data are NMAR because the value of the data that are not observed is driving the reason for nonresponse.

Note that the case of MCAR is quite rare – indeed the impact of administrative censoring is a special case of MCAR and survival analysis techniques employ the assumption that the censoring mechanism is independent of the event of interest when adjusting estimates for censoring. However, the case of NMAR is likely more common but difficult to demonstrate convincingly by the very nature of the problem of the values being related to the missing data problem. By far the majority of missing data methods relate to the attempt to correct statistically for the MAR case – using the observed data to predict the missing information in order to restore the full rectangular dataset for analysis. In general, multiple imputation, using a model-based imputation method is now readily implemented in most statistical packages and is a straightforward way to

appropriately correct for the MAR problem. Alternatively, IPW approaches have been demonstrated to perform well and can obviate the need for creating multiple data sets to inform analyses.

Characterizing Uncertainty for Cost-Effectiveness Ratios

This section focuses specifically on uncertainty in the incremental CE ratio (ICER) statistic, defined as $\text{ICER} = \Delta C/\Delta E$, where ΔE is the per patient mean difference (treatment minus control) in effectiveness and ΔC is the mean per patient difference in cost.

Confidence Intervals/Surfaces for Incremental Cost-Effectiveness Ratios on the Cost-Effectiveness Plane

This subsection considers the presentation of uncertainty on the CE plane and the specific issue of calculating confidence intervals for CE ratios.

The cost-effectiveness plane

The CE plane can be used to show the difference in effectiveness (ΔE) per patient against the difference in cost (ΔC) per patient. By plotting the effectiveness difference on the horizontal axis the slope of the line joining any point on the plane to the origin is equal to the ICER statistic.

One treatment is said to 'dominate' another, being less costly and more effective, if it is located in the northwest (NW) quadrant or the southeast (SE) quadrant of the CE plane. In these two circumstances it is clearly appropriate to implement the dominant treatment and no estimation of CE ratios is required. However, far more common is for one treatment to be more effective but also more costly. In such circumstances, a decision must be made as to whether the additional health benefits of the more effective treatment are worth the additional cost. If the ICER of the more effective therapy ($\Delta C/\Delta E$) – the slope of a straight line from the origin that passes through the (ΔE, ΔC) coordinate – is less than the acceptable 'ceiling ratio' of the decision maker (representing the willingness-to-pay for a unit of health gain) then the treatment should be adopted. This upper limit on CE can be given a value (R_c) and can be represented on the CE plane as a line passing through the origin with slope equal to R_c.

The use of the CE plane has previously been used to illustrate the CE of early endoscopy for dyspeptic patients versus no early endoscopy. In a clinical trial, the early endoscopy arm cost an additional £80 per patient and resulted in an additional 5% of patients free of dyspeptic symptoms at 12 months. This point is plotted on the CE plane in Figure 1 and the slope of the line joining that point to the origin represents the ICER of £1700 per patient free of dyspepsia at 12 months. Also shown on the CE plane are the standard confidence intervals for the difference in effect and difference in effect (the horizontal and vertical I bars respectively that cross at the point estimate). The box that is defined by these I bars (and which is also illustrated in Figure 1) represents an early attempt to approximate sampling uncertainty in the ICER. Nevertheless, subsequent methodological research has demonstrated the utility of two exact methods for confidence interval estimation – nonparametric bootstrapping or the parametric Fieller's approach. These are described and illustrated for the dyspepsia trial below.

Fieller's theorem

Fieller's approach is based on the assumption that the cost and effect differences follow a joint normal distribution, rather than the ratio itself.

The standard CE ratio calculation of $R = \Delta C/\Delta E$ can be expressed as $R\Delta E - \Delta C = 0$ with known variance $R^2\text{var}(\Delta E) + \text{var}(\Delta C) - 2R\,\text{cov}(\Delta E, \Delta C)$. Therefore, a standard normally

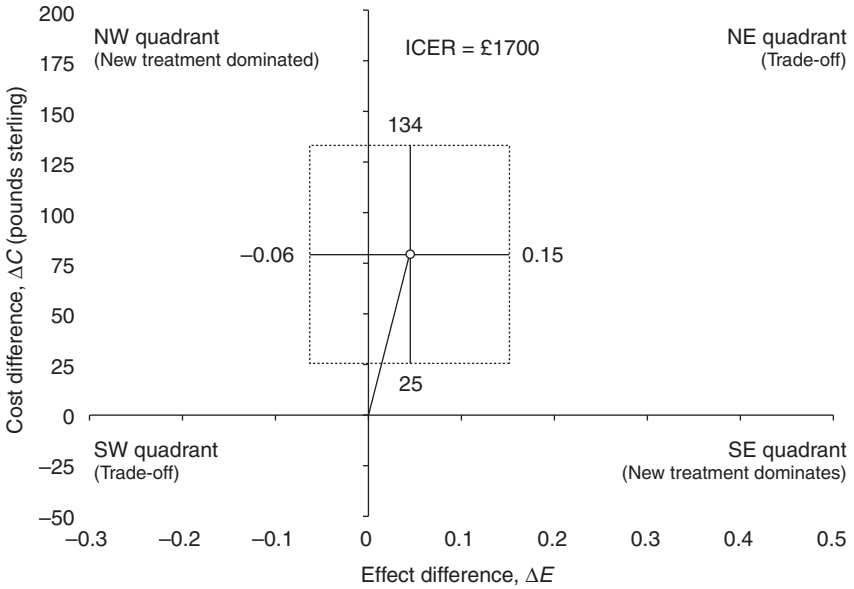


Figure 1 CE plane showing the location of results for the early endoscopy trial. ICER, incremental cost-effectiveness ratio; NE, northeast; NW, northwest; SE, southeast; SW, southwest. Reproduced from Briggs, A. H. (2004). Statistical approaches to handling uncertainty in health economic evaluation. *European Journal of Gastroenterology & Hepatology* **16**(6), 551–561, with permission from Wolters Kluwer Health.

distributed variable can be generated by dividing the expression through by its standard error:

$$\frac{R\Delta E - \Delta C}{\sqrt{R^2\text{var}(\Delta E) + \text{var}(\Delta C) - 2R\text{cov}(\Delta E, \Delta C)}} \sim N(0,1)$$

Setting this expression equal to the critical point from the standard normal distribution, $z_{\alpha/2}$ for a $(1-\alpha)100\%$ confidence interval, yields the following quadratic equation in R:

$$R^2\left[\Delta E^2 - z^2_{\alpha/2}\text{var}(\Delta E)\right] - 2R\left[\Delta E \cdot \Delta C - z^2_{\alpha/2}\text{cov}(\Delta E, \Delta C)\right] + \left[\Delta C^2 - z^2_{\alpha/2}\text{var}(\Delta C)\right] = 0$$

The roots of this equation give the Fieller confidence limits for the ICER.

Figure 2(a) shows the assumption of joint normality on the CE plane for the early endoscopy example: three ellipses of equal density are plotted covering 5%, 50%, and 95% of the integrated joint density. Also plotted are the estimated confidence limits using Fieller's theorem (£300 to −£1100), represented by the slopes of the lines on the plane passing through the origin. Note that the 'wedge' defined by the confidence limits falls inside the 95% ellipse – this is because Fieller's approach automatically adjusts to ensure that 95% of the integrated joint density falls within the wedge.

Bootstrapping

The approach of nonparametric bootstrapping is a re sampling procedure that employs raw computing power to estimate an empirical sampling distribution for the statistic of interest rather than relying on parametric assumptions. A number of authors have demonstrated its potential use for estimating confidence intervals for CE ratios. Bootstrap samples of the same size as the original data are drawn with replacement from the original sample and the statistic of interest is calculated. Repeating this process a large number of times generates a vector of bootstrap replicates of the statistic of interest, which is the empirical estimate of the statistic's sampling distribution.

One thousand bootstrapped effect and cost differences for the early endoscopy example are plotted on the CE plane in Figure 2(b). Confidence limits can be obtained by selecting the 2.5th and 97.5th percentiles of the bootstrapped replications ordered from most favorable to least favorable CE ratio – this effectively ensures that 95% of the estimated joint density falls within the wedge on the CE plane defined by the confidence limits. As is clearly apparent from Figure 3(b), the bootstrap estimate of the joint density and the bootstrap confidence limits (£300 to −£1200) are very similar to those generated by Fieller's theorem.

Estimation or Hypothesis Testing?

In practice, the example in Figure 1 is just one situation that can arise when analyzing the results of an economic analysis conducted alongside a clinical trial with respect to the significance or otherwise of the cost and effect differences. In fact, Briggs and O'Brien (2001) have argued that there are nine possible situations that could arise and these are illustrated on the CE plane in Figure 2 with multiple 'confidence boxes'.

In situations 1 and 2, one intervention has been shown to be significantly more effective and significantly cheaper than the other and is therefore clearly the treatment of choice. In situations 7 and 8 one treatment has been shown to be significantly more costly, but also significantly more effective. It is in these situations that it is clearly appropriate to estimate an ICER and where much research effort has been employed to ascertain the most appropriate method for estimating the ICER confidence interval.

A potential problem arises in the situations where either the cost difference (situations 3 and 5) or the effect difference (situations 4 and 6) is not statistically significant. (The dyspepsia

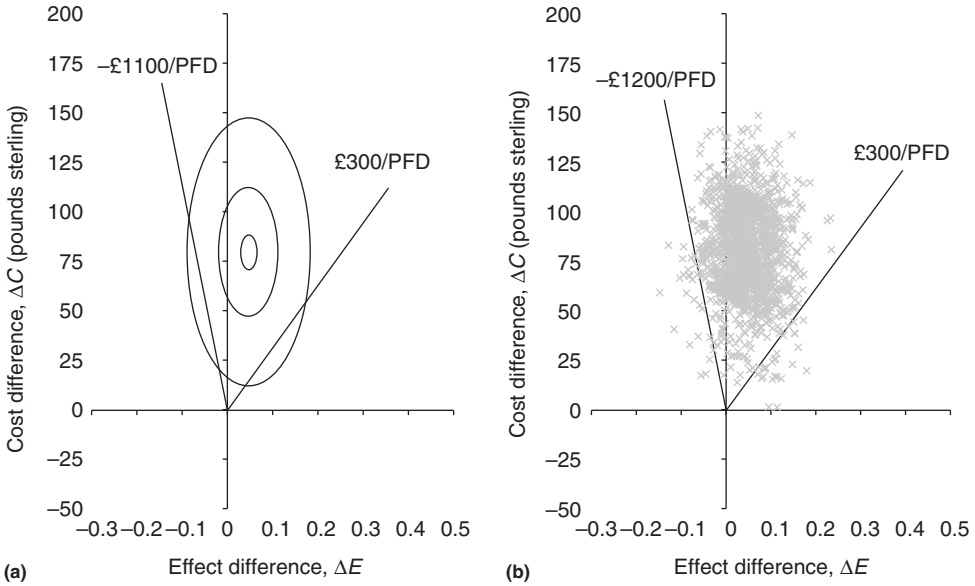


Figure 2 Confidence intervals for CE ratios for the early endoscopy study. (a) Parametric Fieller's theorem and (b) nonparametric bootstrapping. PFD, patients free of dyspepsia. Reproduced from Briggs, A. H. (2004). Statistical approaches to handling uncertainty in health economic evaluation. *European Journal of Gastroenterology & Hepatology* **16**(6), 551–561, with permission from Wolters Kluwer Health.

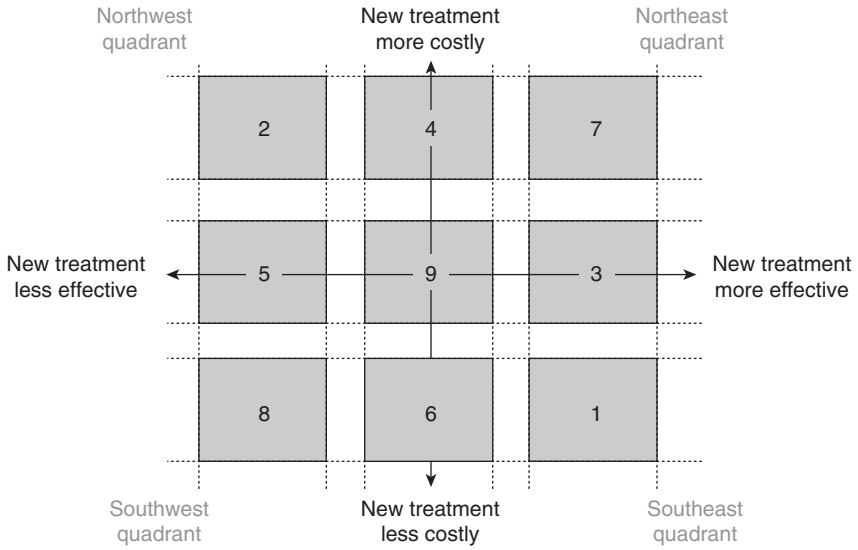


Figure 3 Nine possible situations that can arise concerning the significance (or otherwise) of cost and effect differences illustrated on the CE plane. Boxes indicate the area bounded by the individual confidence limites on cost and effect: statistically significant differences are indicated where the box does not straddle the relevant axis. Reproduced from Briggs and O'Brien (2001).

example falls into situation 4.) It is common to find analysts in these situations adapting the decision rule to focus only on the dimension where a difference has been shown. For example, it might be tempting in situation 4, as in the dyspepsia example, to assume early endoscopy has not better effectiveness than the no early endoscopy option and therefore focus the comparison only in terms of cost. This form of analysis – known as cost-minimization analysis – uses the logic that among outcome-equivalent options one should choose the less costly option.

The problem with this simple approach to decision making in situations where either cost or effect is not statistically significant is that it is based on simple and sequential tests of hypotheses. But the deficiencies of hypothesis testing (in contrast to estimation) are well known, and therefore the goal of economic evaluation should be the estimation of a parameter – incremental CE – with appropriate representation of uncertainty, rather than hypothesis testing.

Acceptability Curves

Although commentators are now largely agreed on the most appropriate methods for ICER confidence interval estimation, such intervals are not appropriate in all the nine situations outlined in Figure 3 above.

An important problem is that ratios of the same sign, but from different quadrants, are not strictly comparable. Negative ICERs in the NW quadrant of the plane (favoring the existing treatment) are qualitatively different from negative ICERs in the SE quadrant (favoring the new treatment) yet will be grouped together in any naïve rank-ordering exercise (note the treatment of negative ratios in the bootstrapping of the early endoscopy example above – because the negative ratios were in the NE quadrant they were ranked above the highest positive ratios to give a negative upper limit to the ratio). Similarly, positive ratios of the same magnitude in the SW and NE quadrants have precisely the opposite interpretation from the point of view of the intervention under evaluation. This is because the decision rule in the SW quadrant is the opposite of that in the NE. For example, an ICER of 500 may be considered as supporting a new treatment in the NE quadrant if society has set a ceiling ratio of 1000. However, in the SW quadrant this value of the ICER would be considered as support of the existing treatment rather than the new treatment. Again, any naïve ranking exercise could easily conflate ICERs with the same magnitude but with different implications for decision making.

Acceptability curves have been proposed as a solution to this problem. If the estimated ICER lies below some ceiling ratio, R_c, then it should be implemented. Therefore, in terms of the bootstrap replications on the CE plane in Figure 2(b), uncertainty could be summarized by considering how many of the bootstrap replications fall below and to the right of a line with slope equal to R_c lending support to the CE of the intervention. Alternatively, using an assumption of joint normality in the distribution of costs and effects, the proportion of the parametric joint density that falls on the cost-effective surface of the CE plane can be calculated. Because the appropriate value of R_c is itself unknown, it can be varied in order to show how the evidence in favor of CE of the intervention varies with the decision rule. The resulting acceptability curve for the early endoscopy example and based on the joint normal assumption shown in Figure 2(a) is presented in Figure 4.

This 'acceptability curve' presents much more information on uncertainty than do confidence intervals. The curve cuts the horizontal axis at the p-value (one-sided) for the cost difference (which is $p<.05$ in the early endoscopy example) because a value of zero for R_c implies that only the cost is important in the CE calculation. The curve is tending toward one minus the p-value for the effect difference (which in the early endoscopy example is $p=.20$), because an infinite value for R_c implies that effect only is important in the CE calculation. The median value ($p=.5$) corresponds to the point estimate of the ICER, £1700 for the early endoscopy example.

As well as summarizing, for every value of R_c, the evidence in favor of the intervention being cost-effective, acceptability

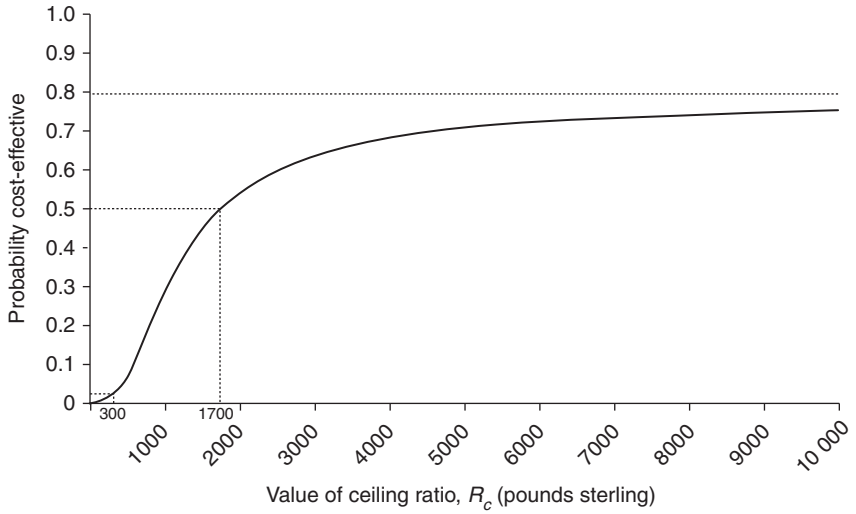


Figure 4 CE acceptability curve for the early endoscopy examples. Reproduced from Briggs, A. H. (2004). Statistical approaches to handling uncertainty in health economic evaluation. *European Journal of Gastroenterology & Hepatology* **16**(6), 551–561, with permission from Wolters Kluwer Health.

curves can also be employed to obtain a confidence interval on CE. The limits are obtained by looking across from the vertical axis to the curve at the appropriate points for the desired confidence level and reading off the associated CE value from the horizontal axis. For the early endoscopy example the 95% upper bound is not defined (because the curve is not defined at 0.975) and the 95% lower bound is equal to £300.

Development of Net-Benefit Solutions

This section describes the reformulation of the standard CE decision rule into one of two possible net-benefit statistics. The use of the net-benefit statistic to estimate acceptability is highlighted, and the section goes on to describe regression approaches that utilize individual-level net-benefits to estimate CE directly.

Net-Benefit Statistics

The algebraic formulation of the decision rule for CE analysis that a new treatment should be implemented only if its ICER lies below the ceiling ratio, $\Delta C/\Delta E < R_c$, can be rearranged in two equivalent ways to give two alternative inequalities on either the monetary scale (NMB) or on the health scale (net health benefit (NHB))

$$NMB: \quad R_c \cdot \Delta E - \Delta C > 0$$

$$NHB: \quad \Delta E - \frac{\Delta C}{R_c} > 0$$

These decision rules are entirely equivalent to the standard rule in terms of the ICER but have the advantage that the variance for the net-benefit statistics is tractable and the sampling distribution of the net-benefits is much better behaved. The variance expressions for net benefits on the cost or effect scales are given by

$$\text{var}(NMB) = R_c^2 \cdot \text{var}(\Delta E) + \text{var}(\Delta C) - 2R_c \cdot \text{cov}(\Delta E, \Delta C)$$

$$\text{var}(NHB) = \text{var}(\Delta E) + \text{var}(\Delta C)/R_c^2 - 2 \cdot \text{cov}(\Delta E, \Delta C)/Rc$$

Because both the net-benefit statistics rely on the decision rule R_c to avoid the problems of ratio statistics, so the net-benefit can be plotted as a function of R_c. Both formulations of net-benefit are illustrated in Figure 5 for the early endoscopy example: the upper pane shows NMB and the lower pane NHB as a function of the ceiling ratio R_c. The net-benefit curves cross the horizontal axis at the point estimate of CE (£1700 per patient free of dyspepsia). Where the confidence limits on net-benefits cross the horizontal axis gives the confidence interval for CE and this is shown between the two panes in Figure 5. The lower 95% confidence limit crosses the axis at £300, whereas the upper 95% limit does not cross the axis indicating that an upper 95% limit on CE is not defined for the early endoscopy example. Note the correspondence with the Fieller limits – this correspondence is explained by the fact that the two methods employ the exact same assumption of joint normality in costs and effects.

Acceptability Solutions

The net-benefit statistic provides a straightforward method to estimate the acceptability curve from Figure 4. The curve can be calculated from the *p*-value on the net-benefits being positive. Note that this gives the acceptability curve a frequentist interpretation, in line with the original paper that introduced the acceptability curve, although that same paper also labeled the vertical axis as 'probability cost-effective.' Strictly, such an interpretation requires a Bayesian view of probability, although it is straightforward to show that the frequentist curve based on the *p*-values for net-benefit and the Bayesian curve based on a normal likelihood and uninformative prior converge.

Regression Possibilities

The Section Net-Benefit Statistics illustrated how the ICER ratio statistic could be reformulated into a linear net-benefit statistic, by using the decision rule. Hoch and colleagues went on to demonstrate how the linearity of the net-benefit

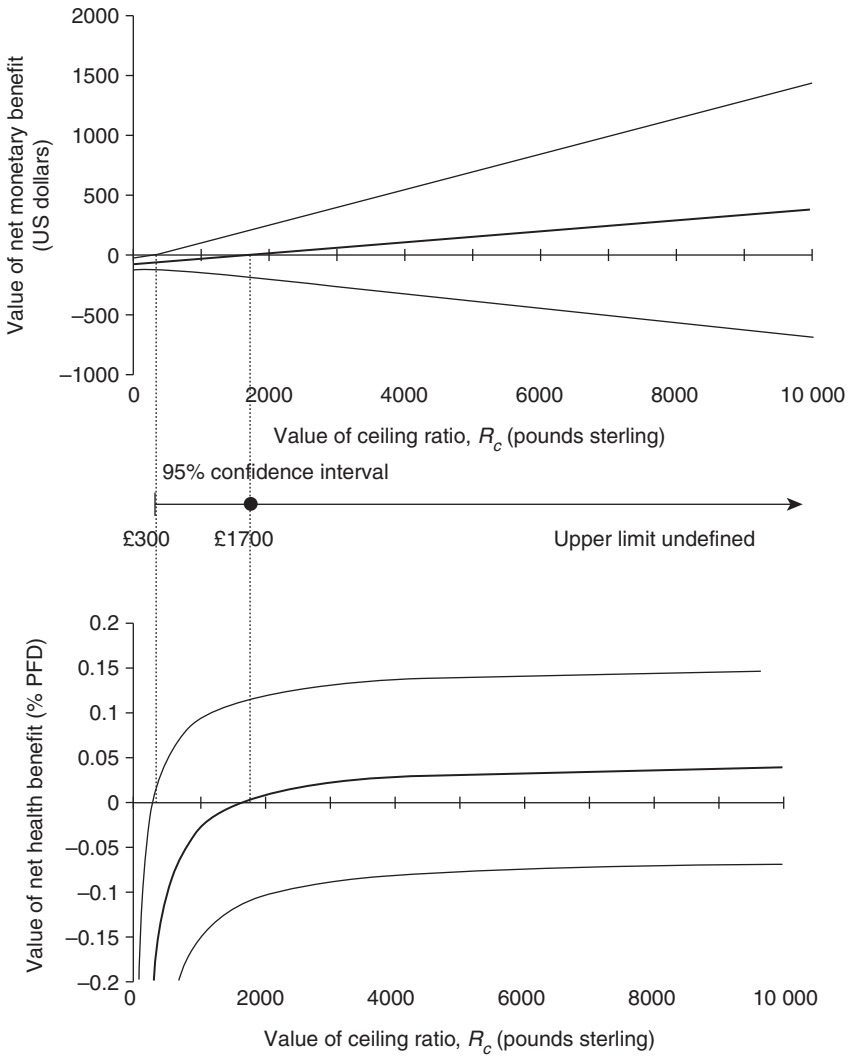


Figure 5 Value of net monetary benefit (upper pane) and net health benefit (lower pane) as a function of the ceiling ratio together with the corresponding confidence interval on cost-effectiveness for the early endoscopy example. Reproduced from Briggs, A. H. (2004). Statistical approaches to handling uncertainty in health economic evaluation. *European Journal of Gastroenterology & Hepatology* **16**(6), 551–561, with permission from Wolters Kluwer Health.

framework can be employed to directly estimate CE within a regression framework. By formulating a net-benefit value for each individual patient i as

$$NMB_i = \lambda \cdot E_i - C_i$$

where E_i and C_i are the observed effects and costs for each patient. At the simplest level, the following linear model

$$NMB_i = \alpha + \Delta t_i + \varepsilon_i \qquad \text{(Model 1)}$$

can be employed where α is an intercept term, t a treatment dummy taking the values zero for the standard treatment and one for the new treatment, and a random error term ε. The coefficient Δ on the treatment dummy gives the estimated incremental net-benefit of treatment and will coincide with the usual estimate of incremental net-benefit obtained by aggregating across the treatment arms in a standard CE analysis. Similarly, the standard error of the coefficient is the same as that calculated from the standard approach. However, the power of the regression approach comes from the ability to covariate adjust (Model 2) and/or look at interactions between covariates and treatments to explore potential subgroup effects (Model 3).

These models are given algebraically below

$$NMB_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \Delta t_i + \varepsilon_i \qquad \text{(Model 2)}$$

$$NMB_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \Delta t_i + t_i \sum_{j=1}^{p} \gamma_j x_{ij} + \varepsilon_i \qquad \text{(Model 3)}$$

where there are p prognostic covariates x.

In the context of an experimental design like a randomized controlled trial, the randomization process is expected to ensure a balance of both observed and unobserved confounding factors across the treatment arms. In this case, the use of prognostic covariates will not materially affect the magnitude

of the estimated CE, but may improve the precision of the estimate and lead to a corresponding narrowing of the estimated confidence intervals such that Model 2 should provide a more precise estimate of incremental net-benefit than Model 1.

The final term is the interaction between the treatment dummy and the prognostic covariates. The significance of the coefficients γ_j on the interaction between the covariates of the model and the treatment dummy represent the appropriate test for subgroup effects – although this does not protect against spurious subgroup effects being detected by chance. Where treatment effect modification is detected, the fact that CE varies for different types of patient may have important consequences for decision making.

Despite the potential of using regression for net-benefit, the use of bivariate regression, through techniques such as seemingly unrelated regression, is more powerful, in that the same explanatory variables do not need to be specified for both cost and effect part of the equations.

Conclusions

This article has explored the development of statistical techniques for analyzing cost and CE data where individual data on coat and effect are available, often as a result of collecting patient-level data alongside a clinical trial. The statistical techniques for examining mean cost differences require subtle changes to the often standard approaches recommended in general (medical) statistics texts. This is due to the focus of CE analysis on informing decision making and maximizing potential health gain from available resources. Hence the focus on testing mean differences rather than reliance on rank order statistics, and the development of adaptations to the Kaplan–Meier approach to estimate cost in the presence of censoring. The net-benefit statistic is much better behaved than the ICER statistic when it comes to statistical analysis and representing uncertainty in CE estimates. Although most authors will want to continue to present results to their audience in terms of traditional ICERs, net-benefit statistics remain an important tool for the analyst to generate statistical measures of uncertainty.

Although this article has focused on the analysis of data on costs and effects generated alongside clinical trials, it is of note that CE analysis conducted within a decision modeling framework often employ individual patient data analyses to inform parameter estimates within the models. By using covariate adjusted parameter estimates, these decision models can explore the potential for patient heterogeneity in CE estimates. Statistical techniques can also be used at the patient level or study level for synthesizing multiple sources of evidence through traditional meta-analyses of network meta-analysis techniques.

The interested reader may also find the following articles of interest in terms of statistical methods for health economic evaluation:

- Economic evaluation alongside clinical trials: issues of design.
- Using observational studies in economic evaluation.
- Reviewing and synthesis of clinical evidence for economic evaluation.
- Analysis of uncertainty.
- Heterogeneity (including subgroup analysis).
- Value of information.

References

Briggs, A. H. and O'Brien, B. J. (2001). The death of cost-minimization analysis? *Journal of Health Economics* **10**(2), 179–184.

Fenn, P., McGuire, A., Phillips, V., Backhouse, M. and Jones, D. (1995). The analysis of censored treatment cost data in economic evaluation. *Medical Care* **33**(8), 851–863.

Little, R. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. Hoboken, New Jersey, USA: Wiley.

Mihaylova, B., Briggs, A., O'Hagan, A. and Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Journal of Health Economics* **20**(8), 897–916.

Further Reading

Billingham, L. J., Abrams, K. R. and Jones, D. R. (1999). Methods for the analysis of quality-of-life and survival data in health technology assessment. *Health Technology Assessment* **3**(10), 1–152.

Briggs, A., Clark, T., Wolstenholme, J. and Clarke, P. (2003). Missing presumed at random: Cost-analysis of incomplete data. *Journal of Health Economics* **12**(5), 377–392.

Briggs, A. H., Wonderling, D. E. and Mooney, C. Z. (1997). Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Journal of Health Economics* **6**(4), 327–340.

Briggs, A. H. (2004). Statistical approaches to handling uncertainty in health economic evaluation. *European Journal of Gastroenterology & Hepatology* **16**(6), 551–561.

Briggs, A. H. A. (1999). Bayesian approach to stochastic cost-effectiveness analysis. *Journal of Health Economics* **8**(3), 257–261.

Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Statistics in Medicine* **15**(13), 1447–1458.

Delaney, B. C., Wilson, S., Roalfe, A., et al. (2000). Cost effectiveness of initial endoscopy for dyspepsia in patients over age 50 years: A randomised controlled trial in primary care. *Lancet* **356**(9246), 1965–1969.

Etzioni, R. D., Feuer, E. J., Sullivan, S. D., et al. (1999). On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics* **18**(3), 365–380.

Hoch, J. S., Briggs, A. H. and Willan, A. R. (2002). Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Journal of Health Economics* **11**(5), 415–430.

van Hout, B. A., Al, M. J., Gordon, G. S. and Rutten, F. F. (1994). Costs, effects and C/E-ratios alongside a clinical trial. *Journal of Health Economics* **3**(5), 309–319.

Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* **16**(3), 199–218.

O'Brien, B. J., Drummond, M. F., Labelle, R. J. and Willan, A. (1994). In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* **32**(2), 150–163.

O'Hagan, A. and Stevens, J. W. (2004). On estimators of medical costs with censored data. *Journal of Health Economics* **23**(3), 615–625.

Stinnett, A. A. and Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* **18**(supplement 2), S68–S80.

Tambour, M., Zethraeus, N. and Johannesson, M. (1998). A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care* **14**(3), 467–471.

Willan, A. R., Lin, D. Y., Cook, R. J. and Chen, E. B. (2002). Using inverse-weighting in cost-effectiveness analysis with censored data. *Statistical Methods in Medical Research* **11**(6), 539–551.

Supplementary Private Health Insurance in National Health Insurance Systems

M Stabile, University of Toronto, Toronto, ON, Canada, and National Bureau of Economic Research, Cambridge, MA, USA
M Townsend, University of Toronto, Toronto, ON, Canada

This article explores the economic theory and evidence regarding supplementary private health insurance in countries with national health insurance systems. It defines voluntary health insurance for the purpose of the article, and classifies the different roles played by voluntary private health insurance. It will then examine the economic literature on voluntary private health insurance, beginning with the theoretical literature before turning to the empirical evidence.

Defining Supplementary Private Health Insurance

Although private health insurance is available in all countries in the European Union, North America, Australia, and New Zealand, private insurance plays significantly different roles depending on the jurisdiction. Almost all the countries included above provide near universal statutory health insurance coverage for their citizens. The dynamics and scope of this statutory health insurance often influence the nature of private supplemental insurance in these jurisdictions. Thomson and Mossialos (2009) provide a useful classification for the role of private health insurance: supplementary, complementary, or substitutive. Their classification is adopted here.

Supplementary insurance generally provides access to services that are already available within the publicly financed health insurance scheme (presumably affording faster access, greater choice, or other amenities). Example jurisdictions include the UK, Australia, and Sweden. Supplemental private insurance markets tend to have small market shares. For example, the UK market covers approximately 10% of the population.

Complementary insurance generally offers services that are not covered under the statutory scheme such as prescription drugs. Example jurisdictions include Canada and Denmark. Some systems also allow for complementary private insurance to cover costs that are typically left outside the public system (e.g., insurance to cover the cost of user fees). Example jurisdictions include France. Market share for complementary insurance is generally higher given the nature of the insurance. In France, more than 90% of individuals have complementary insurance of some form.

Substitutive insurance generally covers people who are not covered under the statutory scheme. Example jurisdictions include Germany. Market share for substitutive insurance is generally smaller (in Germany the market share for private insurance is approximately 10%).

This article focuses on private insurance that is either supplementary or complementary as defined above. Further, the authors focus on private insurance that is voluntary and not statutory (as in the Netherlands) given that many of the potential market failures that occur do so in voluntary markets. It examines both the theoretical and empirical roles that supplementary private insurance can play and the interaction between public and private insurance within this context.

Theoretical Effects of Supplementary Private Health Insurance in National Health Insurance Systems

This section briefly reviews the theoretical effects of a parallel, privately financed system on the performance of the publicly financed system. It focuses on exploring the existing research on the relationship between supplementary private insurance (either as a substitute or complement as described above) and the national public insurance system. It is reasonable to assume that a privately financed system improves the welfare of those who use it as only those forms of private insurance that are voluntary are being considered. It is therefore possible that even if the effects on the public system are negative the overall welfare effects could be positive. However, because this is not a Pareto improvement, and the empirical evidence cited above suggests that the take up of privately financed care is generally small and concentrated among higher income individuals, the focus here will be on whether supplementary private insurance for health services can exist without harming the public system.

Economic theory regarding the effect of introducing a private alternative for publicly financed services is ambiguous. Models of the interaction between private and public insurance systems approach the problem along various population dimensions. For example, previous theoretical literature reviewed in Zweifel (2011) suggests that Pareto improvements are possible in models of differentiated risk types (high and low). Smith (2007) suggests in a two-income type model (rich and poor) that a first best solution is also feasible with supplementary private insurance but that political economy considerations and tax base erosion generally prevent the implementation of such a model.

Depending on assumptions regarding supply of physicians, demand for services, and the magnitude of the effects of conflicting incentives on providers, the theoretical effects of private insurance suggest several possibilities. It is possible that allowing patients to seek health care outside the public system would release public resources and lead to shorter waiting times both for users of the public and the private sector. Further, it is possible that a parallel private system could serve as a benchmark against which the public system could be compared, allowing health care administrators and political leaders to evaluate the efficiency of the public system meaningfully. However, a parallel private health care system may adversely affect the public system, resulting in, at best, no decline in public waiting lists and, at worst, substantial increases. Under certain assumptions, allowing private health care would induce a shift in health care resources from the public to the private sectors resulting in the crowding out of

 doi:10.1016/B978-0-12-375678-7.00924-X

public provision. Theoretically, physicians may have an incentive to increase waiting lists within the public system in order to encourage patients to switch to the private system, where they can bill more than in the public sector. The existence of a private system with deregulated prices can also reduce the monopsony power enjoyed by the public system and result in upward pressure on the prices in the public system. Cream skimming and dumping of risks by the private system can further increase the perpatient cost of the cases remaining in the public system.

Under other assumptions, an increase in supply afforded by a supplemental private system could be offset by an increase in demand for publicly funded health care, so that waiting lists could be relatively unchanged. In addition, introducing private health care may reduce political support for the public health care system by reducing the size of the coalition that uses the public system.

Models examining the effects of wait lists on the demand for private insurance suggest that longer waiting times for care in the public sector may in some cases increase demand for supplemental private insurance. The methods used to ration care within the public system can also have differing effects on the demand for supplemental private insurance. Private supplemental health insurance markets may be smaller when the public sector rations according to need versus random allocation of public resources. Income gradients caused by the introduction of private supplemental health insurance may also be larger under random rationing of care versus rationing according to need (Cuff *et al.*, 2012).

If private insurance complements the public system by covering costs or services not covered publicly such as user charges on the one hand, or additional health care services on the other, it may result in an inefficient level of utilization. In the case where the user charges are meant to help achieve the efficient level of utilization, it may undo these incentives, and may also result in cross subsidization from the tax payer to the user of private insurance. In the case where private coverage insures health care items not covered under the public system, an increase in utilization of privately financed care may increase the use of publicly financed care.

Empirical Evidence on the Effects of Supplementary Private Health Insurance in National Health Insurance Systems

Empirically evaluating the effects of a supplemental privately financed system is difficult due to the lack of counterfactuals and the multitude of differences in the interaction between publicly and privately financed systems across jurisdictions. Because identification is difficult, some of the evidence presented below is correlational evidence. This supports theoretical predictions. Other papers are able to use a variety of microeconomic strategies to tease out some causal relationships. Overall, what evidence there is suggests that there are some potentially negative consequences of privately financed systems on the public system. Evidence on those theoretical considerations with empirical support are presented below.

Public Sector Waiting Times and Demand for Care

Evidence using pooled cross sectional data from the UK suggests a positive relationship between public waiting lists and private insurance (Besley *et al.*, 1999). The positive correlation could be in response to long wait lists or because there is less attention paid to public lists in areas with higher levels of private insurance. Research from Australia examines the relationship between waiting times for care among patients waiting for elective care and the demand for private insurance using data on individual-specific (vs. average) expected wait times. The findings suggest that on average there is little relationship between expected wait times and the demand for private insurance but that for particular subpopulations, who have high probabilities of long waits, there is an increase in the probability of buying private insurance (Johar *et al.*, 2011).

A significant body of evidence reviewed in Thomson and Mossialos (2009) suggests that, in jurisdictions with both private and publicly financed treatment, patients in the private sector wait less than equivalent patients in the publicly financed sector. In those systems where doctors are able to operate in both public and private sectors (e.g., UK, Ireland, and Austria) evidence suggests that doctors give priority to private sector patients.

Evidence from the UK in the 1990s suggests that physicians who operate in both the public and private systems reduce their hours in the public system significantly and do not heed the requirement that publicly employed physicians only dedicate 10% of their earnings to private practice. Physicians who had dual practices earned on average 70% from the National Health Service (NHS) and 30% from private practice (Morris *et al.*, 2008). These authors also find a positive association between mean private income and waiting lists. They note, as above, that this relationship is not necessarily causal and that the causal relationship between wait times and physician effort in the private sector could run in either direction. Evidence from other jurisdictions is consistent with that of the UK, in that, dual practice physicians often do not work all of the contracted hours in the public sector in order to fulfill private sector demand. Evidence on the overall welfare implications of dual practice in developed health care systems is, however, still incomplete and is an area for further future research.

The evidence relating to how changes in public sector wait times impact on the demand for publicly financed service is inconclusive. For example, McAvinchey and Yannopoulos (1993) found that the long run elasticity of demand for NHS acute care with respect to the cost of waiting (a function of time and forgone income) were quite large. Their results suggest that a 1% decline in waiting times lead to a 4.79% increase in the demand for NHS acute care. This evidence implies that introducing a private system that reduces waiting times in the public sector may result in an increase in demand for care in the public sector. However, Francis and Frost (1979) examined the relationship between the number of hospital beds and the magnitude of waiting lists in the UK and concluded that the elasticity of the number of people on the wait list with respect to beds is 1, suggesting that if the number of hospital beds in the public system remains constant waiting lists do not decline regardless of private sector supply.

Evidence from Germany suggests that individuals with private insurance use more pharmaceuticals than individuals in the statutory social health insurance plans (Krobot *et al.*, 2004). However, Hullegie and Klein (2010) find a negative relationship between private insurance and visits to the doctor in Germany among those patients that have at least one doctor visit. They do not find significant differences in hospital stays. The authors suggest "private health insurance either has a positive effect on investment in prevention, because of the monetary incentives provided to the insured, or that privately insured patients receive more intense or better treatment each time they visit a doctor."

Costs

The theoretical evidence reviewed above suggests that supplemental private insurance may either increase or reduce overall public health care costs depending on the interaction between public and private financing. Evidence from Australia, which has promoted voluntary private health insurance along with the public system through the use of tax subsidies, finds that the combination of tax subsidies and the effects of private systems on the health care input costs (both in the short and long run) limit the potential cost savings for the public sector (Hurley *et al.*, 2002). The authors note that there is no conclusive evidence from Australia that shows a decline in public waiting times following the introduction of a parallel private system, nor that public costs were reduced when the overall cost of the policies are taken into account.

Hopkins and Zweifel (2005) note that an additional effect of subsidizing private insurance is that it encouraged policy holders to use more public hospital services and contributed to the government failing to meet its objective to relieve pressure on the public system. The evidence suggests that subsidized private supplemental insurance is a costly way to relieve public sector pressure.

Cream Skimming

Evidence suggests that private and public providers differ not only in the type of services they provide but also in the types of patients treated. Martin and Smith (1996) examined the determinants in length of stay in the NHS in Britain. They found that patients in the NHS who had more access to NHS hospitals were on average likely to experience shorter lengths of stay. They also found that the level of private health care facilities in the area has a positive impact on local NHS costs, suggesting that private health care tends to take the less severe cases so that those who remain in the NHS tend to have higher average costs. This result is significant in that not only does it suggest that the private sector would not serve as an effective benchmark for the public sector but it also implies that, with the introduction of private health care, costs in the public sector would rise.

Demand for Private Insurance

As noted in the introduction, the demand for supplemental private health insurance in most countries is relatively small, of the order of 10–15%. Evidence on the distribution of the demand suggests that, as would be expected, there is a strong correlation between income and take up of private insurance.

Evidence from Spain suggests a positive correlation between the demand for private supplemental health insurance and both wealth and education. Improved coverage, quality, and timeliness were the main reasons cited in Spain for purchasing additional coverage (Costa and Rovira, 2005). In Australia before 1998, only 20% of people with an annual income of less than US$20 000 had private coverage, compared to percentage of people with an annual income of US$100 000 or more (Tuohy *et al.*, 2004). In New Zealand, approximately 37% of the total population had private coverage. Approximately 60% of people with above-average incomes had private coverage, as compared to only 24% of people with below-average incomes. In the UK, individual private insurance is more prevalent among those with higher income and higher education. In the 1990s, 40% of people in the wealthiest 10% of the population were privately insured, whereas only 5% of people in the bottom 40% held private insurance (Tuohy *et al.*, 2004). Evidence on the relationship between health and risky behaviors and the demand for insurance suggests that both poor self-assessed health and risky behaviors are negatively associated with purchasing private insurance (Doiron *et al.*, 2008).

Evidence on the dynamics of using privately financed care suggests that there is considerable movement in and out of private care. Propper (2000) noted that although there is strong evidence of an association between past and current use of private care in the UK, there is also considerable cross-sectoral flow with past use of the NHS associated with current use of private care.

Complementary Private Health Insurance

Evidence on the relationship between privately and publicly financed services, when private finance is complementary, suggests that private financing may increase costs in the public system. In Canada, private insurance complements the public system by covering items not covered publicly – the largest of these being pharmaceuticals. Stabile (2001) found that individuals with private insurance for pharmaceuticals not only used more drugs but also used more publicly financed services such as doctors visits. Part of this was due to selection into private drug insurance but a large component was also due to the reduction in the cost to the patient of using both private and public services. Costs to the public system are also increased through tax expenditures used to subsidize the purchase of complementary private insurance. For example, Canada exempts employer payments for employee health insurance from the taxable income of the employee. Tax deductions are also available in Canada for the cost of privately purchased complementary care. Research examining the effects of such policies suggests that they increase the quantity of insurance demanded on the extensive margin and result in considerable tax expenditures (Smart and Stabile, 2005).

Evidence from France, which allows for private health insurance to reimburse copayments and charges in the public system, suggests that the private voluntary insurance

increases utilization and therefore publicly financed costs (Buchmueller *et al.*, 2004). Moreover, concern over the inequitable distribution of private insurance in France has led to public subsidies for private insurance for lower income families, further increasing public costs. These measures have not reduced the strong relationship between income and private insurance take up.

Conclusions

In summation, private supplemental insurance plays a large role in supplementing and complementing national health services. The literature on the effects of a supplemental, privately financed alternative for services that are also insured publicly on the overall health care system is ambiguous both theoretically and empirically. That said, the weight of the limited evidence available suggests that introducing a private system may result in a decline in the supply of medical services in the public system partially through physician time shifting, and partially through reduced attention to public lists, further resulting in longer waiting lists for patients who remain in the public system. There also appears to be a fair degree of uncertainty surrounding the impact of private insurance through its affect on waiting times on the total demand for health care and hence public insurance. This is in part a result of the difficulty of modeling both the demand and the supply side responses to changes in waiting times as well as the difficulty of determining the causal relationship between private insurance and wait times in the public system. There is some evidence that introducing a private health care system may result in a more complex case-mix in the public sector, resulting in either higher public costs or longer public waiting lists. The evidence across most jurisdictions suggests wealthier and more educated individuals are more likely to take up private insurance and that this is a stronger predictor than health status. Finally, there is little evidence that supplemental private insurance is able to achieve an often-stated goal of reducing pressure on the public system and reducing public sector costs.

Evidence from jurisdictions that use private supplementary insurance to complement the public system by covering charges or services not covered by the public system also suggests that private insurance increases overall demand – not only for those services that are privately covered but for those that are publicly covered as well. In addition, it serves to increase costs in the public sector through additional utilization and a reduction in the incentives brought about through cost sharing.

See also: Access and Health Insurance. Demand for and Welfare Implications of Health Insurance, Theory of. Health Insurance Systems in Developed Countries, Comparisons of. Private Insurance System Concerns. Social Health Insurance – Theory and Evidence. Supplementary Private Insurance in National Systems and the USA

References

Besley, T., Hall, J. and Preston, I. (1999). The demand for private health insurance: Do waiting lists matter? *Journal of Public Economics* **72**(2), 155–181.

Buchmueller, T., Couffinhal, A., Grignon, M. and Perronnin, M. (2004). Access to physician services: Does supplemental insurance matter? Evidence from France. *Health Economics* **13**(7), 669–687.

Costa, J. and Rovira, J. (2005). Why some people go private and others do not: Supplementary health insurance in Spain. *Public Finance and Management* **5**(4), 523–543.

Cuff, K., Hurley, J., Mestelman, S., Muller, A. and Nuscheler, R. (2012). Public and private health care financing with alternate public rationing rules. *Health Economics* **21**(2), 83–100.

Doiron, D., Jones, G. and Savage, E. (2008). Healthy, wealthy and insured? The role of self-assessed health in the demand for private health insurance. *Health Economics* **17**, 317–334.

Francis, B. and Frost, C. (1979). Clinical decision-making: A study of general surgery within trent RHA. *Social Science and Medicine* **13**, 193–198.

Hopkins, S. and Zweifel, P. (2005). The Australian health policy changes of 1999 and 2000: An evaluation. *Applied Health Economics and Health Policy* **4**(4), 229–238.

Hullegie, P. and Klein, T. (2010). The effect of private health insurance on medical care utilization and self-assessed health in germany. *Health Economics* **19**, 1048–1062.

Hurley J., Vaithianathan, R., Crossley, T. and Cobb-Clark, D. (2002). Parallel private health insurance in Australia: A cautionary tale and lessons for Canada. *Institute for the Study of Labor Research Paper Series #515*. Discussion Papers 515. Germany: Institute for the Study of Labor (IZA).

Johar, M., Jones, G., Keane, M., Savage, E. and Stavrunova, O. (2011). Waiting times for elective surgery and the decision to buy private health insurance. *Health Economics* **20**, 68–86.

Krobot, K., Miller, W., Kaufman, J., et al. (2004). The disparity in access to new medication by type of health insurance: Lessons from Germany. *Medical Care* **42**(5), 487–491.

Martin, S. and Smith, P. (1996). Explaining variations in inpatient length of stay in the National Health Service. *Journal of Health Economics* **15**, 279–304.

McAvinchey, I. and Yannopoulos, A. (1993). Elasticity estimates from a dynamic model of interrelated demands for private and public health care. *Journal of Health Economics* **12**, 171–186.

Morris, S., Elliott, B., Ma, A., et al. (2008). Analysis of consultants NHS and private incomes in England in 2003/4. *Journal of the Royal Society of Medicine* **101**(7), 372–380.

Propper, C. (2000). The demand for private health care in the UK. *Journal of Health Economics* **19**, 855–876.

Smart, M. and Stabile, M. (2005). Tax credits, insurance, and the use of medical care. *Canadian Journal of Economics* **38**(2), 345–365.

Smith, P. (2007). Provision of a public benefit package alongside private voluntary health insurance. In Preker, A., Scheffler, R. and Basset, M. (eds.) *Private voluntary health insurance in development*, pp. 147–167. Washington, DC: The World Bank.

Stabile, M. (2001). Private insurance subsidies and public health care markets: Evidence from Canada. *Canadian Journal of Economics* **34**(4), 921–942.

Thomson, S. and Mossialos, E. (2009). Private health insurance in the European Union. *Final Report Prepared for the European Commission, Directorate General for Employment, Social Affairs and Equal Opportunities*. London: LSE Health and Social Care London School of Economics and Political Science.

Tuohy, C., Flood, C. and Stabile, M. (2004). How does private finance affect public health care systems? Marshaling the evidence from OECD nations. *Journal of Health Politics, Policy and Law* **29**(3), 359–396.

Zweifel, P. (2011). Voluntary private health insurance. In Glied, S. and Smith, P. (eds.) *The Oxford handbook of health economics*, pp. 285–307. Oxford: Oxford University Press.

Supplementary Private Insurance in National Systems and the USA

AJ Atherly, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Introduction

Supplemental Insurance

In many countries with public health insurance coverage, individuals have the option to purchase additional coverage from private health insurers to supplement public coverage. This additional coverage can either duplicate public coverage or fill in gaps (supplement) in the public plan, such as covering services outside the public benefit package or filling in cost-sharing gaps in the public coverage. The Organization for Economic Co operation and Development (OECD) defines supplementary coverage as

> Private health insurance that provides cover for additional health services not covered by the public scheme. Depending on the country, it may include services that are uncovered by the public system, such as luxury care ,elective care, long-term care, dental care, pharmaceuticals, rehabilitation, alterative or complementary medicine, etc., or superior hotel and amenity hospital services.

Supplementary policies are commonly held in many OECD countries, including Australia, Canada, Germany, Ireland, Switzerland, the Netherlands, the UK, and the US Medicare population. Although the precise rules vary by country, the key motivation is that many types of public insurance leave policy holders with substantial potential liability for out-of-pocket expenses (Figure 1).

What Does Supplemental Insurance Cover?

Supplemental insurance coverage varies depending on the rules of the particular country as to what it is allowed to pay for and what the public program includes. Typically, coverage may include cost sharing for publicly provided services (e.g., France, USA), coverage for services outside the public benefit package (e.g., Canada, Germany), particularly dental services (e.g., Australia, Japan, UK), and superior amenities, such as private rooms in hospitals (e.g., Italy, UK). In some situations, supplemental coverage may also serve to provide swifter access to some services (e.g., Norway, UK).

What, precisely, supplemental insurance covers varies not only across countries but also for particular countries over time as the rules for what can be covered and the public benefit packages change. For example, supplements for the Medicare program in the USA often provided coverage for prescription drugs before 2006; when prescription drug coverage was added to the benefit package in 2006, supplemental policies often dropped that benefit. Currently, discussions are under way in the USA to limit the amount of cost-sharing supplemental insurers can cover.

Examples of what supplemental insurance policies cover in different countries are as follows:

- *Australia*: medications not covered by the public system; dental services, aid, and appliances; copayments for covered services;
- *Canada*: prescription drugs; dental care; nonhospital institutions (long-term care), vision care, and over-the-counter medications;
- *France*: dental and vision services, and copayments for covered services;
- *Germany*: uncovered services; access to better amenities and some copayments;
- *Italy*: over-the-counter drugs, dental care, access to better hospital amenities, and improved provider choice;
- *Japan*: dental services;
- *Netherlands*: adult dental care;
- *New Zealand*: copayments and cost sharing, elective surgery in private hospitals, private outpatient specialist consultations, and faster access to nonurgent treatment;
- *Norway*: shorter waiting times for publicly covered elective services.

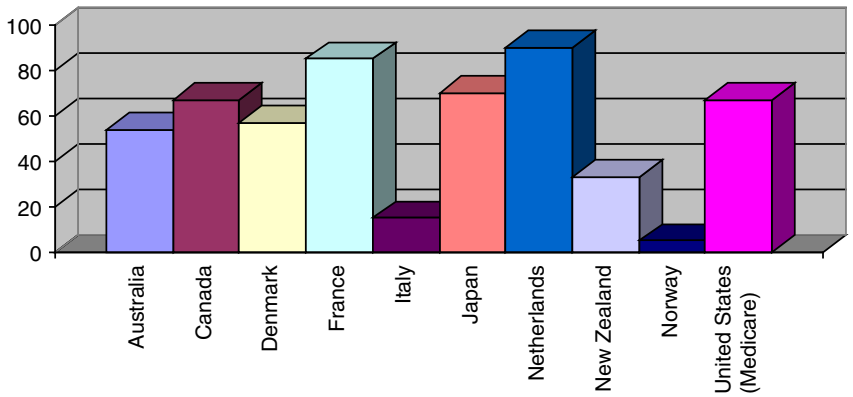


Figure 1 Percentage of population with supplemental insurance, selected countries.

 doi:10.1016/B978-0-12-375678-7.00917-2

Duplicative Coverage

In some situations, coverage can be purchased that effectively duplicates or replaces the publicly provided coverage. This then provides a private alternative to the public system. OECD defines duplicative coverage as

> Private insurance that offers coverage for health services already included under public health insurance. Duplicate health insurance can be marketed as an option to the public sector because, while it offers access to the same medical services as the public scheme, it also offers access to different providers or levels of service. It does not exempt individuals from contributing to public insurance.

Examples of this approach to supplemental insurance include Australia, Ireland, and the UK. Some countries, such as Canada, explicitly prohibit duplicative coverage.

Duplicative coverage is distinct from supplemental insurance in a number of different ways. Most importantly, duplicative coverage is intended to replace the public coverage – it is essentially an opt-out from the public system. In contrast, supplemental insurance is intended to enhance the public program and improve it by either providing more extensive coverage, reducing financial risk, or providing better access to publicly funded services.

Controversies Regarding Supplemental Insurance

Supplemental insurance has been the subject of some criticism. One of the principal criticisms is that if supplemental insurance provides valuable financial protection, it implicitly suggests that the public system is inadequate because additional coverage is necessary. If additional coverage is necessary, this then implies that there are important gaps in the public coverage. Supplemental insurance, because premiums are not based on income, thus allows higher income individuals better financial protection than lower income individuals, which undermines the goal of public coverage for health care.

Another criticism is that in some cases, the supplementary coverage can have problematic interactions with the public coverage. This can involve risk selection, increases in public costs due to the presence of the supplemental policy, and/or distortions in terms of access, such as wait times, to public care that serves to feed the demand for privately insured care. For example, in both France and the USA, studies have found that the removal of cost sharing by supplemental insurance leads to the increased use of services in the publicly funded program.

Supplemental Insurance in the USA

In the USA, supplemental insurance is most common among beneficiaries of the publicly funded Medicare program, and is sometimes referred to as 'Medigap' plans. Medicare supplemental insurance can either help pay for Medicare cost sharing or provide coverage for services not included in the Medicare benefit package, such as health insurance outside the USA.

The Value of Supplemental Insurance Plans

Medicare is the largest publicly provided health insurance program in the USA, with approximately 49 million enrollees in 2012. Although there are a number of ways to gain Medicare eligibility, the most common is through age eligibility, which occurs at age 65. Although Medicare covers many medical services, it often does not cover the services in full. For example, in 2012, Medicare Part A would pay for the full cost of a hospitalization for days 1–60, with a $1156 deductible (equal to the cost of the first day of the hospitalization). For the typical Medicare beneficiary – with a median income of $22 000 – Medicare cost sharing creates a substantial financial liability. If a Medicare beneficiary used all of the Part A (inpatient) covered services in 2012, the total cost sharing would be approximately $54 910, which includes a $8670 copay for hospital days 61–90, a $34 680 copay for hospital days 91–150, and $11 560 for skilled nursing facility (SNF) for days 21–100. For hospital stays beyond 150 days and SNF stays beyond 100 days, there is generally no coverage.

Medicare gaps are of two different types:

- Cost sharing for covered services
- Benefit limits

For example, for Medicare Part B, there is a $100 annual deductible plus a 20% copayment for covered services. There are also many services outside the benefit package, such as eyeglasses and hearing aids. Supplemental insurance can address either of these program limitations.

Sources of Supplemental Insurance Plans

There are two main sources of private supplemental insurance plans: employers and individual purchase. Individual purchase plans (which are often referred to as the 'Medigap' plans) are designed to be integrated with Medicare; in contrast, employer supplements are often extensions of medical insurance provided for active workers and thus not optimally designed for coordination with Medicare. Employer plans are provided to retired workers, with eligibility rules that often mirror early retirement rules. Typically, eligibility is dependent on the employee's age and length of service with the firm.

For employer plans, Medicare is considered the primary payer with the supplemental/employer plan serving as the secondary payer. Although there are several different methods used to coordinate benefits, with the most common method being carve out, beneficiaries still pay some portion of Medicare deductibles and thus have higher cost sharing than true Medigap plans. However, employer supplementary plans are more likely to include coverage for noncovered Medicare benefits, such as chemical dependency treatment, vision coverage, dental coverage, and 'catastrophic expenses' caps, whereby the total out-of-pocket liability is capped.

The main alternative for Medicare beneficiaries without access to group coverage is individually purchased plans, often called 'Medigap' plans. Medigap plans date back to the beginning of Medicare, in the mid-1960s, and have been extensively regulated since the 1990s. During congressional

hearings before the Congressional Select Committee on Aging in 1978, extensive marketing abuses were described. These abuses by the policy sellers included using high-pressure sales tactics, misrepresentation by the issuer of the policy, misrepresentation of policy contents and competitors' policies and 'rollover' of plans, whereby subscribers were forced to change policies to increase policy commissions.

Subsequent to these hearings, two different regulatory reforms of the individual supplemental insurance market were enacted. The first, in 1980, was the Voluntary Certification of Medicare Supplemental Health Insurance Policies, commonly known as the Baucus Amendment (Public Law 96–265, Sec. 507). The Baucus Amendment addressed the abuses of the market by setting minimum coverage standards, outlawing the knowing sale of multiple policies, and requiring higher loss ratios. The amendment was not considered to have successfully achieved its policy goals. Hearing in the 95th Congress suggested that many of the same issues remained, largely due to the voluntary nature of the Baucus Amendment requirements. Thus, a second reform was enacted in the Omnibus Budget Reconciliation Act of 1990 (OBRA-90), Section 1882 of the Social Security Act. Unlike the Baucus Amendment, the OBRA-90 reforms were mandatory and changed the industry significantly.

OBRA-90 increased minimum loss ratio requirements for the plans, prevented the sale of duplicate plans, established consumer counseling programs, limited agents' commissions, and required a 6-month open-enrollment period. This period allowed beneficiaries to purchase Medigap policies without regard to health status, with guaranteed renewal of the policy. The 'enrollment window' opens when the beneficiary initially enrolls in Part B. After expiration of the 6-month window, some policies (although not all policies) become experience rated or medically underwritten. Finally, OBRA-90 required the creation of model policies, which were the only new Medigap policies allowed to be sold after 30 July 1992.

Standardization of 'Medigap' Plans

The standardization requirement limits the Medigap plans that can be sold to the approved plans. These approved plans have precisely the same benefit structure regardless of seller. Initially, the National Association of Insurance Commissioners was charged with the development of 10 model policies. The initial model policies provided a range of options, although all plans were required to cover a set of 'core benefits,' which include coverage for the Part A hospital daily copayments for days 61 through 150, the 20% Part B coinsurance on physician charges, and the first three pints of blood received each year, as well as coverage for an additional 365 days of hospital care. The model policies (typically labeled policies 'A' through 'J') originally featured various combinations of eight benefits: the Part A deductible, the Part B deductible, coverage for the SNF copayment, foreign travel, prescription drug coverage, preventive medical care, and coverage for at-home recovery. Only 2 of the 10 plans covered the Part B deductible (C and F), but together these two plans included more than half the market in the 1990s and 50.9% in 1994. The most common benefits originally were the Part A deductible (included in 9 of the 10 model policies, B through J) and coverage for the SNF copayment and foreign travel (both included in C through J).

Congress has amended the model benefits a number of times, most recently in the Medicare Modernization Act (MMA) of 2003. After MMA, there are eight different benefits plus the 'core' benefits distributed across 10 plans. The core benefits focus on Medicare Part A copayments for extremely long hospitalizations. All of the plans offer some form of coverage for Medicare Part B cost sharing, the first three pints of blood during a hospitalization (uncovered by Medicare), and the Part A hospice care coinsurance. Plans M and N became available for the first time in June of 2010, at which time plans D and G were modified and Plans E, H, I, and J could no longer be sold, although beneficiaries with those plans could continue coverage (Table 1).

Other benefits include the SNF coinsurance (eight plans), the Part A deductible (nine plans), the Part B deductible (two plans), Part B excess charges (two plans), and coverage for emergencies during foreign travel (six plans).

The two most popular plans in 2010 were Plan F (44% of enrollees) and Plan C (14%) . These are also the most comprehensive plans offered. Both Plans C and F will be revised in 2015 to include some cost sharing for Part B services. Participation in the new plans – Plans K–N – is extremely low. Combined, Plans K and L account for less than 1% of plan purchases.

Table 1 Current standardized Medigap plans

Medigap benefits	*Medigap plans*									
	A	*B*	*C*	*D*	*F*	*G*	*K*	*L*	*M*	*N*
Medicare Part A coinsurance and hospital costs up to an additional 365 days after Medicare benefits are used up	X	X	X	X	X	X	X	X	X	X
Medicare Part B coinsurance/copayment	X	X	X	X	X	X	50%	75%	X	X
Blood (first three pints)	X	X	X	X	X	X	50%	75%	X	X
Part A hospice care coinsurance/copayment	X	X	X	X	X	X	50%	75%	X	X
Skilled nursing facility care coinsurance			X	X	X	X	50%	75%	X	X
Medicare Part A deductible		X	X	X	X	X	50%	75%	50%	X
Medicare Part B deductible			X		X					
Medicare Part B excess charges					X	X				
Foreign travel emergency			X	X	X	X			X	X

Premium Levels and Regulation

Federal regulations establish two different time periods when there is 'guaranteed issue' of Medigap premiums, i.e., time periods when insurers may not decline a beneficiary. The first is during the first 6 months after initial enrollment into Medicare Part B. The second time period covers a series of transitional periods between different types of Medicare coverage, such as between Medicare Advantage (MA) and standard fee-for-service. Medical underwriting is allowed, but is generally limited to age, gender, and smoking status.

Premiums are generally regulated at the state level. Seven states required 'community rating' for Medigap policies in 2010. Community rating requires a single premium for all enrollees in the plan. Four states generally use 'issue age' rating, where the premium depends on the age of initial enrollment. The remainder of states use 'attained-age' rating, whereby the premiums depend on the current age of the policy holder. An ASPE analysis of premiums found wide variation across states, with premiums in the most expensive state (New York) nearly double that of the least expensive state (Michigan). During the decade between 2001 and 2010, average premiums increased 3.8% per year. In six of those years, the increase in Medigap premiums was less than that of total Medicare spending. This trend holds within plan types.

Demand for Supplemental Insurance 'Medigap' Plans

Medigap insurance is generally demanded by individuals who are more affluent. Research has found that those buying Medigap plans tend to have higher family income and other financial assets; to be younger, white, and married; and to have more education and a usual source of care. Evidence regarding age and health has been mixed, with some studies finding higher rates of purchase associated with better health and other studies the opposite. It appears that the relationship between the beneficiary's health and supplemental insurance decision is dependent on knowledge of Medicare. In general, Medicare beneficiaries are badly informed about both Medicare design and insurance. However, chronically ill beneficiaries often have enough exposure to the health-care system to become well informed about Medicare's limitations. Thus, the relatively better informed chronically ill beneficiaries are more likely to buy insurance, whereas those without chronic illnesses are not, regardless of self-rated health.

Other Sources of Coverage

There are also several other sources of coverage available to select groups of Medicare beneficiaries. First, some Medicare beneficiaries are also eligible for Veteran's Administration (VA) benefits. In 2004, 13% of Medicare-only beneficiaries (without supplemental insurance) identified a VA facility as their primary source of care. The services provided at VA and military facilities are generally not charged to either the individual receiving the services or to Medicare.

Second, some Medicare beneficiaries are also eligible for Medicaid. There are a number of different Medicaid programs available, depending on income level, but most at least pay for Medicare's cost sharing and some provide coverage beyond the Medicare benefits package ('wrap-around benefits').

Supplemental Insurance and Medicare Advantage Plans

Medigap enrollment has declined markedly since 2006, when Medicare Part D came into effect. Part D provides prescription drug coverage and allows Medicare beneficiaries to select drug plans. The drug plans can be either 'stand-alone' plans or part of a Medicare managed care product, referred to as 'Medicare Advantage' (MA). MA plans are fully capitated health plans that serve much the same purpose as Medigap plans in that they help reduce Medicare cost sharing and provide additional benefits beyond the standard Medicare benefit package.

MA plans have existed in some form since 1983. The most valuable benefit offered by MA plans was prescription drug coverage, the value of which was diluted by MMA and Part D. However, after 2006, Medicare beneficiaries began switching in large numbers to MA plans from Medigap plans. Total Medigap market share declined from 25% in 2003 to less than 20% in 2010, whereas MA enrollment climbed from 11% to 25% in the same time frame.

Effect of Supplemental Insurance Plans on Medicare Spending

There is an extensive literature on the effect of supplemental insurance on Medicare spending. The theory is that by lowering the out-of-pocket price of medical care, the quantity demanded of care will increase. This basic application of demand theory has extensive empirical evidence supporting it, including findings from the RAND health insurance experiment, a randomized controlled trial of the effect of cost sharing on the use of medical services from the 1970s. Over the past 25 years, there have been more than 15 studies on the effect of supplemental insurance on Medicare spending. The results of these studies vary markedly depending on the empirical methodology, data, and approach to controlling for adverse selection.

Adverse selection is a particular problem for the empirical estimation of the effect of supplemental insurance on Medicare spending. Theoretically, one would expect that higher risk individuals would be more likely to buy supplemental insurance because it holds greater value for individuals at greater risk of medical events. Showing a positive relationship between the purchase of supplemental insurance and Medicare spending thus is consistent with both adverse selection (higher cost individuals buy supplemental insurance) and demand (supplemental insurance leads individuals to become higher cost).

Empirically, studies have found effect sizes varying from zero (no effect) to a 33% increase in Part A spending and a 42% increase in Part B spending. The Congressional Budget Office (CBO) estimates that individuals with Medigap policies use 25% more Medicare services than those with no supplemental insurance and 10% more than those with

employee-sponsored insurance (which has higher cost sharing than Medigap plans). CBO bases its estimates both on an analysis of the supplemental insurance literature and on the results of the RAND health insurance experiment.

Further Reading

ASPE (2011). *Variations and trends in Medigap premiums*. Washington, DC: Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services, Office of Health Policy.

Atherly, A. (2001). Medicare supplemental insurance: Medicare's accidental stepchild. *Medical Care Research and Review* **58**(2), 131–161.

Buchmueller, T., Couffinhal, A., Grignon, M. and Perronnin, M. (2004). Access to physician services: Does supplemental insurance matter? Evidence from France. *Health Economics* **13**, 669–687.

Commonwealth Fund (2012). *International profiles of health care systems, 2012*. New York: The Commonwealth Fund.

Finkelstein, A. (2004). Minimum standards, insurance regulation and adverse selection: Evidence from the Medigap market. *Journal of Public Economics* **88**, 2515–2547.

Lemieux, J., Chovan, T. and Heath, K. (2008). Medigap coverage and Medicare spending: A second look. *Health Affairs* **27**(2), 469–477.

Maestas, N., Schroeder, M. and Goldman, D. (2009). Price variation in markets with homegenous goods: The case of Medigap. National Bureau of Economic Research. Working Paper 14679. Available at: http://www.nber.org/papers/w14679 (accessed 06.11.12).

Organization for Economic Co-operation and Development. Private health insurance in OECD countries. Paris: OECD Health Project series 2004.

Survey Sampling and Weighting

RL Williams, RTI International, Raleigh, NC, USA

Introduction

A sample survey is a method for collecting data from or about the members of a population so that inferences about the entire population can be obtained from a subset, or sample, of the population members. As an example, it may be desired to know the average length of stay in a hospital for surgical versus nonsurgical stays in the US and its territories for the 2012 calendar year. In this situation, a sample of hospital discharges would be obtained along with the duration of stay for each discharge. Then estimates of the average length of stays for surgical and nonsurgical discharges would be calculated and compared. A properly conducted sample survey will support inference from the sample that is scientifically valid about the population. This article focuses on probability sampling and weighting to support such inference.

The discussion is organized around four major steps: (1) survey requirements, (2) sampling design, (3) weighting, and (4) design effect. Although these steps are presented as a linear process progressing in order, in practice much iteration between the steps will occur while planning a sample survey. For example, the initial requirements may prove to be financially infeasible when determining the sample design and compromises will need to be made in the requirements.

Survey Requirements

An important first step is to establish the objectives of the survey that will drive the design. Major areas to consider include:

Target Population

The target population is the finite set of all elements or units about which inferences or conclusions are to be drawn. The definition of the target population should be exact in terms of element, place, and time. For the hospital length of stay example, the target population is all hospital discharges (element) in the US and its territories (place) during calendar year 2012 (time). Various subpopulations are also identified which will be important in the subsequent analysis. For example, it may be important to differentiate between urban versus rural hospitals or public versus private hospitals.

Survey Variables

Associated with each element of the target population are the survey variables to be measured. A survey is conducted to gain information about one or more population characteristics or parameters which are defined in terms of the survey variables. For example, the survey variables might be the length of stay for a hospital visit and if it was surgically related or not. The population parameters of interest might be the average length of stay, the median length of stay, or the total number of hospital inpatient days all by type of discharge (surgical vs. nonsurgical).

Objectives

The objectives of the survey can be either descriptive, analytic, or both. The objectives are stated in terms of the population parameters derived from the survey variables. For the hospital stay example, descriptive objects would include estimating the average length of stay for surgical and nonsurgical stays. Such estimates would be important to planners in determining the number of hospital beds needed in a new hospital or in a service region. The estimates might also be used to determine the anticipated total amount of reimbursement a payer might incur for hospital stays. Alternatively, analytic goals might be to determine factors related to length of hospital stay so that best practices can be established to reduce the average length of stay. For example, average length of stays might be compared between surgical modalities or condition treatment plans.

Precision Requirements

The degree of precision required for the survey objectives are needed to establish the final sample design and the sample sizes. For descriptive objectives, precision is usually stated in terms of the maximum standard error of the estimate or in terms of the maximum length of a confidence interval around an estimate. For example, estimate the average length of stay such that its 95% confidence interval is no longer than plus or minus 0.5 day. Precision for analytic objectives is usually stated in terms of the power, or probability, of rejecting a null hypothesis in favor of an alternative hypothesis for a given value of the population parameters that describe the hypothesis. For example, when testing if average surgical length of stay is the same as the average nonsurgical length of stay, it might be required to have an 80% chance to reject the null hypothesis of equality when the actual, but unknown, population values differed by more than 1 day with a type I error rate of 5%.

Sampling Design

The sampling design consists of the procedures by which elements are selected into the sample from the population. The major attributes of a sampling design are presented next.

Survey Population

The survey population includes any modification to the target population established because of resource limitations or other feasibility factors on the survey. The survey population

usually limits the target population in some way so that the survey is more readily conducted. For example, it might be cost prohibitive to conduct the survey of hospital discharges in the US territories and the survey population would limit the scope of the survey to the 50 states and the District of Columbia of the US.

Sampling Frame

A sampling frame is an important tool in the process of selecting a sample. The sampling frame is the materials or methods which identify and provide access to the elements of the target population. The sampling frame also includes any auxiliary information required to select the sample or to analyze the resulting data. A rule must exist that allows enumeration of all of the elements of the target population. The sampling frame can be a simple listing of all of the members of the target population; for example, a list of all of the hospitals in the US and its territories. More commonly, the sampling frame consists of processes and rules that provide access to the target population. For the hospital length of stay example, a complete listing of all hospital discharges does not exist. However, a multistage approach can be used where lists of hospitals can be used to contact selected hospitals each of which can provide access to a listing of its discharges.

Stratified Sampling

Stratified sampling is one of the major design features used in almost all sample surveys. Stratification is the process of dividing the population into mutually exclusive and exhaustive groups and then selecting a separate independent sample from each stratum. When the observations within each stratum are more homogenous than those between the strata, the variance of the resulting estimate will be reduced. However, stratification is more importantly used to assure that an adequate sample size is obtained for analysis from the various subpopulations included in the survey objectives. For example, it is likely that a nonstratified random sample of hospitals will not contain enough hospitals from rural areas for analysis purposes. In this situation, stratifying the sample of hospitals by urban versus rural areas allows an adequate sample size of rural hospitals to be selected to support the survey's analytic objectives.

Multistage Sampling

Another important and commonly used design feature is multistage sampling. This is a process by which sampling is carried out in two or more stages. At the first, or primary stage, clusters of the sampling elements are formed and a sample of the clusters is selected. At the second stage, a subsample of the elements within each of the selected first-stage clusters is selected. In the hospital length of stay example, the discharges could be clustered by the hospitals where they occurred. A sample of hospitals would then be selected at the first stage followed by a sample of discharges from each of the selected hospitals at the second stage. More than two stages of sampling may be used. For example, a sample of geographic areas, such as US counties, could be selected at the first stage, followed by a second sample of hospitals and then a sample of discharges at the third stage. As noted when discussing sampling frames, multistage sampling is useful when a sampling frame must be constructed in stages. It is also used to control the cost of conducting a survey by concentrating the data collection effort at a limited, and predetermined, number of locations. For example, if a random sample of discharges was selected without clustering, then data collection would occur at a large number of hospitals across the US leading to high data collection costs. However, a two-stage sample of hospitals followed by discharges within selected hospitals would be less expensive as the data collection effort can be concentrated at a smaller number of hospitals.

Probability Sampling

Probability sampling is the mechanism through which inference is extended from the sample to the population. A probability sampling plan associates a nonzero probability of selection with each and every member of the survey population such that the selection probability can be determined for every member of the sample. A random process is used to select the sample so that the desired probabilities of selection are achieved. To demonstrate how probability sampling supports population inference, assume that a probability sample of size n is selected from a survey population of N elements. Then let δ_i be 1 if the i-th element of the survey population is selected into the sample and 0 if it is not selected. The probability that the i-th element is selected into the sample is $\pi_i = \mathrm{E}(\delta_i)$, where the expectation is over the random process used to select the sample. Associated with each survey population element is the value of a survey variable Y_i, with the observed value for each sample member being y_j. The population total is $Y_+ = \sum_{i=1}^{N} Y_i$ with the sample total estimator being $y_+ = \sum_{j=1}^{n} y_j/\pi_j$. It follows that $\mathrm{E}(y_+) = \mathrm{E}[\sum_{j=1}^{n} y_j/\pi_j] = \mathrm{E}[\sum_{i=1}^{N} \delta_i Y_i/\pi_i] = \sum_{i=1}^{N} \mathrm{E}(\delta_i) Y_i/\pi_i = Y_+$ showing that the sample total estimator is an unbiased estimate of its corresponding population total.

Simple random sampling

Simple random sampling is one of the most easily implemented types of probability sampling. The two forms of simple random sampling are with replacement and without replacement. Without replacement sampling assigns the same chance of selection to all $\binom{N}{n}$ possible without replacement samples of size n from a survey population of N elements. With replacement sampling assigns the same chance of selection to all N^n possible with replacement samples. In either case, the selection probability for any member of the sample is n/N. Simple random sampling without replacement is more commonly used and is appropriate when each member of the survey population is of equal interest or importance.

Probability proportional to size sampling

Probability proportional to size (PPS) sampling is commonly used when selecting multistage samples. PPS sampling, as its names implies, results in each sample member having a selection probability proportional to a measure of its size. For

example, the size of a hospital might be measured by its annual number of discharges or its number of beds. Similarly, the size of a geographic unit might be the number of persons living in the unit. The PPS selection probability for a unit is $\pi_i = nS_i/S_+$, where n is the sample size, S_i is the size measure for the i-th unit, and $S_+ = \sum_{i=1}^{N} S_i$ is the total of all size measures for units in the survey population. When a very large sampling unit has a size measure such that $S_i > S_+/n$, then the unit is called a self-representing unit as its PPS selection probability is greater than one. In this situation, all self-representing units are included in the sample with probability one and the remainder of the sample is selected PPS from the survey population excluding the self-representing units.

Equal probability of selection method

Equal probability of selection method (EPSEM) is any sampling design that yields equal selection probabilities for the ultimate sampling elements used in the analysis. Having equal selection probabilities for the analysis units is often a desirable property as it usually reduces the variance of the survey estimates. In multistage sampling, this is achieved by combining PPS and simple random sampling at different stages of sampling. For the hospital length of stay example, assume that a two-stage sample of hospitals followed by discharges within the selected hospitals is planned. A common approach would be to select a PPS sample of n hospitals, where the size measure is the number of discharges from the hospital (S_i). This would then be followed by a sample random sample of m discharges from each selected hospital. The selection probability for the j-th discharge from the i-th hospital is the product of the hospital's selection probability and the conditional selection probability of the discharge from its hospital. This is $\pi_{ij} = \pi_i \times \pi_{j|i} = (nS_i/S_+) \times (m/S_i) = nm/S_+$ and is the same for all discharges regardless from which hospital a discharge is selected. In most situations it is not possible to have a size measure from which exactly equal probabilities of selections are achieved. However, using a measure of size that is proportional to the desired measure of size will yield nearly equal selection probabilities. In the example, the number of hospital beds or the number of discharges from a previous year would usually be good measures of size.

Weighting

As was shown above, probability sampling provides a process for drawing valid inferences from a small sample about the population parameters of a large population. This is done by defining a sampling weight for each sample member that is the inverse of its sample selection probability. Symbolically, the sampling weight for the j-th sample member is $w_j = \pi_j^{-1}$ The sampling weight can be roughly thought of as the number of population members that a sample member represents. The sampling weights are used to expand the sample members up to approximate the population. When all of the selected sample members respond and cooperate with the survey, unbiased estimates of linear population parameters, like population totals, are obtained when the sampling weights are used to expand the sample data. Nonlinear population parameters are consistently estimated through functions of weighted estimates of totals.

To illustrate this process, assume that a probability sample of n hospital discharges, both surgical and nonsurgical stays, from a total population of N hospital discharges has been selected. Two population parameters of interest are the total number of days spent in hospital and average length of stay, both for surgical hospital stays. Let Y_i be the length of stay associated with the i-th discharge in the population and let γ_i be 1 if the discharge is for a surgical stay and 0 otherwise. The population total number of days spent in the hospital for surgical stays is $Y_{S+} = \sum_{i=1}^{N} \gamma_i Y_i$ and the population total number of surgical stays is $N_{S+} = \sum_{i=1}^{N} \gamma_i$. Thus, the population average length of surgical stay is $A_S = Y_{S+}/N_{S+}$. The unbiased estimators of Y_{S+} and N_{S+} are the weighted sample values $y_{S+} = \sum_{j=1}^{n} w_j \gamma_j y_j$ and $n_{S+} = \sum_{j=1}^{n} w_j \gamma_j$, respectively. A consistent estimator of the population average length of surgical stay, A_S, is the ratio of the two weighted sample values $a_s = y_{s+}/n_{s+}$. This example demonstrates the general process of using the sampling weights to expand the survey values associated with the sample members to unbiasedly estimate the population totals. The estimate totals are then combined to consistently estimate other population parameters such as means, percentages, and regression coefficients.

In almost all surveys some selected sample member will not respond and their data will be missing. Simply leaving out the missing data from the sample nonrespondents will bias the resulting estimates. To mitigate the effect of the missing data, adjustments to the sampling weights are used to create analysis weights, which compensate for the nonrespondents in the analyses. Weight adjustment methods are beyond the scope of this chapter.

Design Effect

Complex sample surveys rarely result in a set of independent and identically distributed observations because of sample design features such as stratification, multistage sampling, and unequal weighting. Such features affect the variance of survey estimates and specialized software is needed for the analysis that allows the sample design to be used when estimating the variances. For example, survey data analysis software is available in SUDAAN®, SAS®, and Stata®.

To understand the effect of the design features, the concept of a design effect is used. The design effect is the ratio of the variance under the sample design used to collect the data to the variance of a simple random sample selected with replacement of the same sample size. Symbolically, the design effect of the mean is $DEFF = Var(\bar{y})/(S^2/n)$, where S^2 is the population variance of the variable in question, and $Var(\bar{y})$ and n are the variance of the estimate and the sample size under the sample design used to collect the data.

The sample design feature that usually most affects the variance is multistage sampling. When clusters of observations are selected together, the variance of an estimate is usually increased because the observations within a cluster are most often positively correlated. In a two-stage sample design, where clusters are sampled first followed by individual observations within each cluster, the amount of increase in the

variance of the estimated mean is approximately $DEFF = 1 + (m - 1)\rho_y$ where m is the average number of observations selected per cluster from the analysis domain and ρ_y is the intracluster correlation between two observations in a cluster. In the hospital length of stay example, the clusters are the hospitals and it would be expected that the length of stays for discharges from the sample hospital are positively correlated. For regression coefficients, the inflation, or possible deflation, in variance is approximately $DEFF = 1 + (m - 1)\rho_y\rho_x$, where ρ_y and ρ_x are the intracluster correlation coefficients for the dependent variable and the independent variable, respectively. For certain designs and regression models it is possible for ρ_x to be negative, resulting in a decrease in the variance of the estimated coefficient.

A related concept is the effective sample size which is given by $n_e = n/DEFF$. The effective sample size is the sample size for a simple random sample selected with replacement that yields the same variance of an estimate as that obtained from the sample design used to collect the data. An enlightening example for the mean estimated from a two-stage design illustrates the interpretation of the effective sample size. Consider a two-stage design where 10 $(= m)$ sampling units are selected from each of the 50 sampled clusters for a total sample size of 500. If $\rho_y = 1$, then $DEFF = 10$ and $n_e = 50$, the number of clusters. This is the situation where the observations within a cluster are perfectly related and no further information is gained by selecting more than one observation from each cluster. Thus, the effective sample size is the number of clusters. However, if $\rho_y = 0$, then the observations within each cluster are unrelated, and $DEFF = 1$ and $n_e = 500$. This is the situation of independent observations all of which contribute equal information to the estimate. In most situations, ρ_y is between 0 and 1, and the effective sample in this example is between 50 and 500.

The effective sample size can be used to estimate power or precision when planning a survey. The effective sample size can be approximated using the relationships described above using information from previous studies to approximate $n_e = n/DEFF$ and then used in a power/precision formula or software package to determine the approximate power or precision.

See also: Missing Data: Weighting and Imputation

Further Reading

Kish, L. (1965). *Survey sampling*. New York: Wiley.
Lavrakas, P. J. (ed.) (2008). *Encyclopedia of survey research methods*, vol. 2. Los Angles: Sage.
Levy, P. S. and Lemeshow, S. (2008). *Sampling of populations: Methods and applications*, 4th ed. New York: Wiley.
Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Switching Costs in Competitive Health Insurance Markets

K Lamiraud, ESSEC Business School, Cergy, France

Glossary

Community-rated premiums Premiums are community-rated when, within a given insurance company and for a given type of insurance contract, they are uniform across enrollees with different health statuses.

Risk selection practices in health insurance markets Three main types of risk selection practices exist. 'Dumping' refers to deliberately losing the sickest clients (e.g., based on medical information). 'Cream skimming' refers to attracting healthy individuals, for example, through advertisement campaigns targeted at young people. 'Stinting' occurs when a company initially tries to and then continues to keep high-risk individuals away, for example, by not sending application forms to older patients who ask for them. In Switzerland, risk selection may occur because of community rating: Companies expect to make profits on good risks and losses on bad risks. Dumping is forbidden in Switzerland, but cream skimming and stinting may occur.

Switching costs (barriers) Costs incurred by consumers when switching from one supplier to another. Categories of switching costs include those caused by: Transaction costs of switching suppliers (e.g., monetary costs and time lost due to necessary paperwork when switching to a new provider); uncertainty about the quality of untested brands; psychological costs such as 'brand loyalty'; discount strategies implemented by the current provider; shopping costs when consumers buy several products from the same provider. Switching costs may also be related to search costs.

Introduction

Many European countries have social health insurance where citizens cannot choose between different providers for basic coverage. While Germany and the Netherlands have only recently introduced policies giving citizens the freedom to choose their own health plan, this has been a long-standing feature of the Swiss health care model. The assumption is that competition to obtain consumers puts insurance providers under pressure to increase service quality and/or decrease premiums. As for any market, however, competition in health insurance only works if the threat of consumers 'voting with their feet' is credible. In other words, it only works if enough consumers switch to more competitive insurers.

In this article, the possible presence of switching costs when consumers are offered the opportunity to change their basic health insurance carrier is investigated. It is focused on the specific case of Switzerland as this country, through its pure form of competition and the period of time which has elapsed since this system was implemented, offers one of the best settings to study competition in basic health insurance markets.

The article is organized as follows. Following this introduction, Section 'Switching Costs' provides some insights into the general theory on switching costs. Section 'Managed Competition in Switzerland: The Regulatory Framework' describes the features of managed competition in basic health insurance in Switzerland. Section 'Stylized Facts: The Ineffectiveness of Competition to Date' highlights the persistence of huge premium differences within Swiss cantons, which may be explained by low-switching rates. Section 'Possible Barriers to Switching Behaviors' explores possible switching barriers. Section 'Conclusions' concludes and suggests ways of improving the current system through reducing switching costs.

Switching Costs

In many markets, consumers incur costs when switching from one supplier to another. These costs are called switching costs (barriers).

Categories of switching costs include those caused by: Transaction costs of switching suppliers (e.g., monetary costs and time lost due to necessary paperwork when switching to a new provider); uncertainty about the quality of untested brands; psychological costs such as 'brand loyalty'; discount strategies implemented by the current provider; shopping costs when consumers buy several products from the same provider. Switching costs may also be related to search costs.

In a market with switching costs, the rational consumer will not switch to the supplier offering the lowest price if the switching costs (in terms of monetary cost, effort, time, uncertainty, and other elements) outweigh the price differential between their current supplier and the new one. If this happens, the consumer is said to be locked-in to the current supplier. If a supplier manages to lock-in consumers in this way, it may raise prices to a certain point without fear of losing these customers. However, the incentive to do so must be balanced with the incentive to set a lower price to attract new customers. Despite this second incentive, the first situation is expected to dominate. Switching costs often do raise average prices in competitive markets compared with competitive markets without switching costs. The possible consequence of this is that consumers may be worse-off. Accordingly, policy intervention to reduce switching costs may be appropriate.

Empirical studies highlight the importance of switching costs for a wide range of markets including credit cards, cigarettes, computer software, supermarkets, air travel, phone services, online brokerage services, electricity suppliers, and automobile insurance. In this article, the presence of switching costs in basic health insurance markets is investigated.

Empirical evidence from the Swiss context serves to illustrate several types of switching costs presented above.

Managed Competition in Switzerland: The Regulatory Framework

Switzerland (population 7.8 million in 2009) is divided into 26 cantons, each canton being responsible for the organization of its own health care system. Overall health care is regulated by the Federal Law on Social Health Insurance (LAMal), which has been in force since 1996 after its ratification in a popular referendum in 1994.

The main regulatory features of Swiss health insurance markets are described below.

1. An 'individual mandate' requires all residents to have health insurance coverage. Individuals must take up insurance within their canton of residence. Each family member must contract on an individual basis. Health insurance cannot be provided by an employer as a fringe benefit and so the premium is paid in full by the insurance enrollee, a situation which should make the latter very reactive to differences in premium. Cantons are given the responsibility for ensuring that every resident receives coverage. The threat of lawsuits ultimately enforces the individual mandate in cases of noncompliance.
2. The law defines a standardized benefit package in order to avoid competition on content of coverage. Hence, all insurance companies must reimburse the same basket of goods. Although small variations may exist in the quality of services provided (e.g., different reimbursement time frames), these are minimal in nature and do not call into question the characteristics that the same product has to be offered by various providers. The level of cost sharing is also defined by law and is invariable across insurers (see **Box 1**).
3. The law authorizes full freedom in terms of choosing one's primary physician as well as unlimited access to specialists. Physicians are paid on a fee-for-service basis. However, enrollees can voluntarily opt for contracts with limited choice of physicians and those physicians who provide services within such contracts are paid on a per capita basis (see point 4 below).
4. Premiums charged by companies to consumers are community-rated (see Glossary). This means that although they can differ between health plans, an insurer must offer uniform premiums to people who meet all three of the following criteria: same age group (0–18, 19–25, and >25 years), same geographic area, and same type of coverage. With regard to geographic areas, there are 78 pricing areas, i.e., three per canton. Nevertheless, for a given company, prices turn out to be very similar between the three price areas within the same canton. Hence it can be considered that there are effectively 26 areas of price competition. With regard to the type of coverage, three types of basic health insurance coverage are available: all companies must offer a contract with a low deductible which guarantees access to any physician; they can also offer contracts with higher deductibles (see **Box 1**) and/or contracts with a limited choice of physicians. In 2008, the most frequent choice by enrollees was a 300 CHF deductible health insurance policy (38.7%), followed by plans with higher deductibles (31.2%). Insurance covering a limited choice of providers (Health Maintenance Organizations – HMO contracts) accounted for 30.0% of enrollees. This latter figure reflects HMOs recently increasing market share, given that only 8.2% of enrollees held HMO contracts in 2003.
5. Note that premiums paid by enrollees are neither rosk- nor income-related. Clients on low incomes receive subsidies from their canton of residence. In 2008, the mean yearly subsidy was 1511 CHF per subsidized enrollee.
6. A risk-equalization mechanism is enforced at the cantonal level (see **Box 1**) so that funds with a higher percentage of bad risks are compensated in comparison with those with a higher percentage of good risks and in order to avoid risk selection practices by health insurers (see Glossary).
7. Health insurers must accept every application for basic insurance.
8. Enrollees can switch companies twice a year, in June and December.

Finally, there is clear regulatory separation between basic and supplementary coverage. For health care services not included in the basic benefit package, an individual may subscribe to contracts for supplementary coverage, which cover, for example, dental care, private or semiprivate hospital rooms, cross-border care, and alternative medicine. Supplementary insurance is regulated by the Insurance Contract Law, which allows risk selection by companies and does not impose any constraint on the coverage supplied. Basic and supplementary insurance can be purchased from two different insurers or from the same insurer.

These features suggest that freedom of choice in terms of choosing one's insurer is very much encouraged by the regulatory framework and in particular that changing health insurers for basic coverage involves very low quality-related or transaction-type switching costs. Indeed, basic insurance coverage is virtually identical from one health insurer to the next, and generally, the enrollee can remain with the same physician or hospital regardless of insurer. Furthermore, the switching procedure is simple: the individual must write a letter to their health insurer, the templates for which are freely available on well-known websites. Also, search costs are low. All premiums are published officially every year by the Federal Office for Public Health and distributed to households that request them. Furthermore, the most competitive premiums can be easily found on the Internet and in newspapers.

If one looks at the market structure, it can be seen that enrollees have a great deal of choice. Although the number of health insurers (all nonprofit) offering mandatory health care insurance in Switzerland decreased between 1996 and 2008 (145 and 86 authorized health insurers, respectively), the choice set faced by each consumer has increased since the LAMal was implemented. In 1996, the mean number of health plans per canton was 39. Consumers could choose from more than 40 health plans in only two cantons. The mean number of health plans per canton rose to 57 in 2006, varying between 50 and 69 choices.

Box 1 The Swiss Basic Health Insurance Regulatory Features in detail

Cost-Sharing Arrangements

All contracts include a deductible on yearly expenditures. Enrollees can choose from 6 possible deductible levels (300, 600, 1000, 1500, 2000, 2500 CHF). Once the deductible level has been reached, enrollees pay a 10% coinsurance rate up to a maximum of 700 CHF. Hence, if the enrollee chooses a 300 CHF deductible, then the maximum out-of-pocket amount that they may have to pay is 1000 CHF.

Risk Equalization (or Risk-Adjustment)

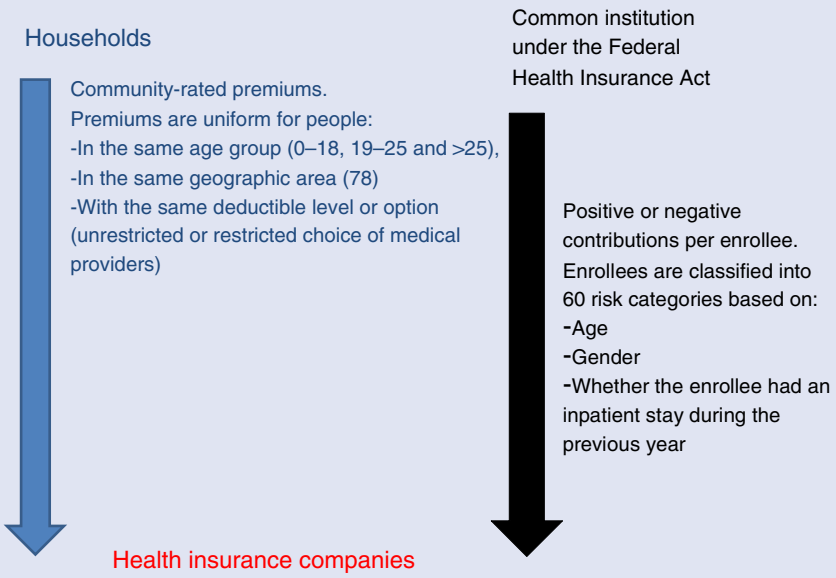
A retrospective risk-equalization mechanism is enforced by the law at the cantonal level. The 'Common Institution under the Federal Health Insurance Act,' a federal office, regulates the system. Risk adjustment consists in adjusting the premium by redistribution.

Until 2011, only age and gender were used as risk-adjusters. Adult policyholders were classified according to 15 age categories (18–25, 26–30, …, >91) and gender. Hence there were a total of 30 risk categories. The average value of costs in a given risk category was computed at the end of the calendar year. The average value of costs within each canton was also determined. The difference between these two averages indicated whether health insurance funds had to pay a contribution (if the value was <0) or receive a contribution (if the value was >0). Sickness insurance funds had to pay (or receive) contributions for each policyholder belonging to the risk category concerned. For example, if the difference between the average of costs for females aged 66–70 and the average of total costs amounted to 1400 CHF over a year in a given canton, then each insurer in this given canton would receive 1400 CHF for each of its female enrollees aged 66–70. Note that the costs used to compute contributions are total health care costs incurred by the patient minus out-of-pocket health expenditures directly borne by the patient. Also note that the value of money transfers obtained/given back for each enrollee does not depend on whether the enrollee has opted for a low or a high deductible level.

A third risk adjuster, based on whether the enrollee had an inpatient stay (of at least three days) during the previous year, was adopted by parliament on the 21 December 2007, hence increasing the number of risk-adjusters to three and the number of risk categories to 60. This change came into effect on 1 January 2012 and will be implemented at the end of 2012.

Hence an enrollee brings their community-rated premium to the insurer plus the risk-adjusted money transfer from the 'Common Institution under the Federal Health Insurance Act.' The chart below illustrates this.

Money flows to health insurance companies:



Sources of Data used in this Article

Data concerning the Swiss health insurance markets come from various sources. Premiums are published officially every year by the Federal Office for Public Health. Information on consumer health plan choices was collected by two surveys. The Federal Social Insurance Office survey was carried out in 2001 and can be obtained from the Swiss Information and Data Archive Service. A follow-up survey was carried out in 2007 by the University of Lausanne under the supervision of Brigitte Dormont, Pierre-Yves Geoffard, and Karine Lamiraud who wrote the questionnaire.

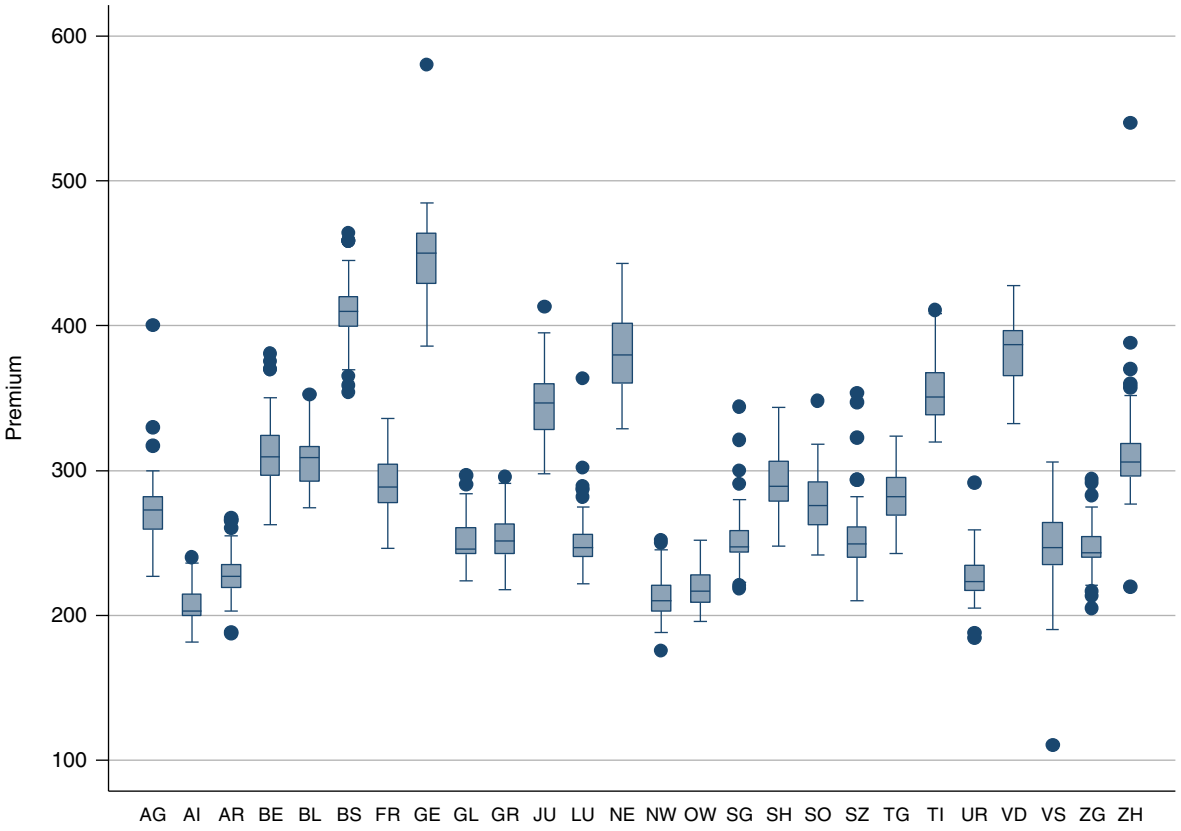


Figure 1 Box plot of adult monthly premium (for a 300 CHF deductible contract).

Stylized Facts: The Ineffectiveness of Competition to Date

In a health insurance market like the Swiss one, with its community-rated premiums for each health plan, homogenous benefits, open enrollment, and a large choice of insurers, one would expect strong price competition within each area of competition, resulting in small premium differences across plans and, in turn, cost-containment or at least a moderation in premium increases.

However, the observed facts are very different. Premiums have continued to rise. In 1997 the price index for an adult basic health insurance contract was 92.3. It increased to 100 in 1999 and 142.1 in 2008. Furthermore, there is a great deal of variability across firms in premiums within a given canton, as suggested by the box plots of monthly premiums displayed in Figure 1. For example, in 2011, in the Geneva canton, the difference between the least and the most expensive yearly premiums equaled 1665 Swiss Francs, a difference of 39%, for an adult contract with the lowest deductible, which is very large, considering it is only for one family member. Moreover, the within-canton variance has remained quite stable over time. Hence, moderation in premium increases has not occurred and premiums have not converged.

Competitive health care plans cannot be held responsible for the increase in premiums. The rise in premiums mostly mirrors the rise in health care costs. Indeed health care costs represent a very high percentage of premiums (administrative costs are quite low and represent approximately 6% of collected premiums) and premiums are highly correlated to health care costs in each canton, with Pearson and Spearman correlation coefficients close to 1. The increase in health care expenditures in Switzerland can be largely accounted for by an increase in wealth and in technological developments. It may also be related to patient behaviors including the moral hazard effect (i.e., overconsumption because of guaranteed coverage) and physician behaviors (e.g., incentives to increase the volume of care in the fee-for-service payment system). However, the moral hazard effect is not specific to private health insurance systems, such as the Swiss one. Public insurance systems can also be affected. Furthermore, as suggested above, competitive health plans in Switzerland do not have the means to control health care expenditures from the supply side. In particular they cannot define modes of payments for physicians or implement selective contracting except in HMO options, which have recently started gaining market share. Hence, competitive health care plans cannot be held responsible for the increase in health care expenditures.

In contrast, the lack of premium convergence may be related to the ineffectiveness of competition, and one important factor for this is low-switching rates. In Switzerland, annual switching percentages are low despite existing price differentials for identical benefit packages. Health plan switching

rates only averaged approximately 3% between 1997 and 2007, ranging from 2% to 5%.

Now the barriers to switching in the Swiss basic health insurance market are investigated.

Possible Barriers to Switching Behaviors

How can low-switching rates be explained?

Given the fact that price information is widely available and that there would seem to be substantial opportunities to take advantage of lower premiums for what appear to be homogeneous health plans, the authors begin by considering how the standard market model addresses consumer choice. The model assumes that, under uncertainty, the expected utility of customers will be maximized. Given a set of N choices, an individual will choose a health plan if its expected utility is greater than that of each of the alternatives. After choosing a plan, a consumer may experience a change in health state or other personal circumstances (e.g., reduced income) or face a new set of premium choices due to new health plans entering the market. The new circumstances may cause the individual to reassess the expected utility of their current health plan in light of alternatives, with the result that they may switch.

Since the implementation of LAMal in Switzerland, the size of the choice set in local markets has grown (see Section 'Switching Costs') and the set of plans offering the cheapest premiums for basic insurance has continued to change. Hence, low-switching rates may be explained by switching costs.

As already mentioned, the most obvious type of switching costs, i.e., transaction-type and quality-related switching costs, can be ruled out in the Swiss context. Consequently others need to be considered. Now the following three main types of switching costs are investigated: psychological switching costs, switching costs related to the multiproduct environment, and switching costs related to regulatory features.

Psychological Switching Costs

Choice overload

Research in economics and psychology consistently brings up the issue of whether greater choice is always in the consumer's interest, the argument being that too much choice may inhibit consumers from making any choice.

Two underlying forces may explain such a phenomenon. First, the information or cognitive overload theory argues that, as the choice set grows, the cost of the individual's information processing increases. This happens if individuals continue to consider all alternatives as the choice set expands. Even if consumers use shortcuts (e.g., eliminating the worst alternative), information processing costs grow with the choice set. This leads to the hypothesis that consumers can be overwhelmed by 'too great a choice.' The result is an inverted U relationship between the size of the choice set and the quality of decision making.

The second psychological force concerns the fear of making an incorrect choice or subsequent regret in situations where decision making is complex, consequential, and uncertain. One response to such decision-making circumstances, observed in both experimental and observational studies, is a tendency toward decision avoidance either by opting for the status quo or by walking away from the decision entirely.

In the Swiss health insurance context, the large number of competing health plans may result in information overload, even though individual health plans can be easily assessed. Frank and Lamiraud (2009) provide some support for this phenomenon. A survey focusing on individuals' health plan choices in Switzerland together with market price data were used to study the factors associated with the probability of switching and, in particular, to investigate the impact of the number of available plans on the probability of an individual switching. The results showed a monotonically decreasing likelihood of switching with increasing choice. Cantons with more choices had significantly lower switching rates *ceteris paribus*. Furthermore, it was demonstrated that consumers consider all health insurance companies, including fringe players (i.e., companies with small market shares) when deciding about insurance cover. These results are consistent with the inertia in decision making associated with choice overload.

Status quo bias

Consumer attachment to the status quo (status quo bias) could also account for low-switching rates. This has been associated with a tendency to exaggerate the disadvantages of leaving one's current situation and to understate the potential gains of switching, in an environment of uncertainty and complex decision making.

Three results highlighted by Frank and Lamiraud (2009) provide further evidence of the existence of a status quo bias in Swiss health insurance markets. First, people with longer periods of attachment to a particular health plan were less likely to express their intention to switch plans. Second, people making new health plan choices (switchers and those new to the market) chose to enroll in a different set of health plans from those who had not switched for some time. Third, survey respondents explicitly reported that their decision not to change their health plan was out of habit or because they were satisfied with their policy.

Switching Costs Generated by the Multiproduct Environment

Another possible barrier to switching behavior is the relationship between basic and supplementary insurance.

Although a clear regulatory separation exists between basic and supplementary insurance in Switzerland, in reality, both types of insurance coverage are strongly linked: Companies are allowed to operate both in basic and supplementary markets and most individuals subscribe to the same provider for both. Of the 88% of enrollees who took out supplementary coverage in 2007, only 9% subscribed to different companies for their basic and supplementary contracts.

To analyze the interaction between basic and supplementary insurance, two characteristics of the Swiss health insurance market have to be considered. First, there are additional costs when a client's basic and supplementary contracts are with different companies (e.g., separately mailed bills). Second, risk selection is authorized for supplementary insurance.

Switching costs generated by the relationship between basic and supplementary insurance have been shown to originate both from the consumer (Supplementary Health insurance as a barrier to switching Basic Health Insurance provider) and the firm (Pricing Strategies).

Supplementary health insurance as a barrier to switching basic health insurance provider

Dormont *et al.* (2009) investigated four possible mechanisms through which holding a supplementary insurance contract may act as a barrier to switching one's basic insurance policy to another insurer.

1. The first mechanism relies on a 'pure switching cost effect.' Given that subscribing to basic and supplementary contracts with two different insurers induces administrative costs, those planning to switch may have to consider moving both their basic and supplementary contracts: This is more burdensome than a single switch.
2. The second mechanism refers to selection practices in the supplementary insurance market, and to consumer beliefs about the existence of such a policy. Take for example a customer who thinks that they are a 'bad risk' and believes that insurers reject applications for supplementary insurance contracts from individuals considered as such. Currently holding a supplementary insurance contract would then act as a barrier to them switching basic insurance. Indeed, the new insurer may reject the application for a supplementary contract or propose an intentionally unacceptable offer (e.g., very high premiums).

   The third and fourth mechanisms ((3) and (4) below) refer to selection practices in basic health insurance markets (see Glossary). In such cases, there is an incentive for insurers to retain enrollees who hold supplementary contracts and drop the others.
3. Mechanism (3) is simply based on the fact that regulation for supplementary insurance is less restrictive. Lack of contract standardization may lead to less competition and profits may be realized from selling supplementary insurance contracts. In this context, profit-maximizing insurers would have an incentive to retain supplementary contract purchasers.
4. Mechanism (4) is based on the assumption that holding a supplementary insurance contract might be correlated with being a 'good risk' vis-à-vis basic insurance, i.e., having a lower probability of consumption in basic insurance for a given illness. This conjecture might be relevant for supplementary insurance covering alternative medicine: Individuals who subscribe to such contracts may be more reluctant to consume 'standard' health care, especially drugs, covered by the basic insurance. This might also be true for other kinds of supplementary contracts, perhaps indicating the subscriber's greater attention to prevention. An insurance company can observe the use of health services by its enrollees, but the econometrician cannot: only self-assessed health is observed. If we suppose that supplementary insurance indicates that the individual is a good risk, then discovering that those with supplementary insurance and/or good health are less likely to switch, would in turn suggest that insurance providers try and succeed in retaining good risks.

Given the prohibition of risk selection in the Swiss market for basic insurance, mechanisms (3) and (4) raise the question of what indirect tools are available to insurers in order to retain certain enrollees. Anecdotal evidence regularly reported in newspapers suggests that some insurance companies rely on such commercial practices as offering discounts on sports items or events.

To disentangle these four possible mechanisms, Dormont *et al.* (2009) assessed to what extent the influence of supplementary contracts on switching rates depends on the enrollee's health status. For (1) and (3) to be true, the effect of supplementary contracts on switching rates would have to be unrelated to the individual's self-assessed health. For (2) to be true, holding supplementary insurance would have to act as a barrier to switching for those in poor health. For (4) to be true, holding supplementary insurance would have to act as a barrier to switching for those in good health.

Controlling for relevant covariates, Dormont *et al.* (2009) show that holding a supplementary contract reduces the probability of switching in basic insurance for those in poor self-assessed health, but has no effect on the switching behavior of enrollees in good/very good health. These empirical findings suggest that the main mechanism at work is (2): If the customers think they are a bad risk and believe that insurers reject applications for supplementary contracts from individuals considered as such, they might refrain from switching basic insurance provider. This effect, identified through survey data covering the period following the implementation of the reform (1997–2000), was confirmed by behaviors observed over 2003–07 (Dormont *et al.*, 2013).

Pricing strategies

Lamiraud and Stadelmann (2011) examine the relationship between basic and supplementary health insurance from a different angle. They analyze firms' pricing strategies (i.e., pricing of basic and supplementary products) as a way of reinforcing consumer inertia. In particular, they investigate whether firms use bundling strategies or supplementary products as low-price products in order to capture consumers. Bundling is the sale of two or more products in a package (i.e., one basic contract and one supplementary contract). The bundle comes at a discount with respect to the total price of the individual goods when sold separately. Another strategy consists in establishing a low price for a product in order to attract customers who are likely to buy other products at regular or high prices.

Lamiraud and Stadelmann (2011) do not show any evidence of bundling in the Swiss setting. They do however show that firms use low-price supplementary products to lock-in consumers. A majority of firms price one of their products at a low price. None offer cheap products overall (i.e., in both basic and supplementary markets). Low-price insurance products differ across companies. When buying a low-price supplementary product, consumers always buy their basic contract from the same firm. Furthermore, those who opt for low-price supplementary products are less likely to declare an intention to switch basic insurance companies in the near future. The latter result is true for each level of risk category.

Hence, pricing strategies seem to generate additional barriers to switching basic insurance provider, thereby reinforcing consumer inertia.

Switching Costs Generated by the Regulatory Features

Another possible barrier to switching behaviors is that some regulatory features may tend to attenuate competition in the basic health insurance market. In this article it is focused on the rules for building reserves and the risk-equalization mechanism.

Health insurers are legally obliged to build reserves in order to protect against unpredictable financial risks associated with unforeseen catastrophic events, such as epidemics. Depending on their size, insurance companies are required to keep between 10% and 20% of collected premiums in reserve. Reserves per enrollee are not transferable to another company. Hence, if a subscriber leaves one insurance provider, their reserves stay with that provider. Consequently that fund becomes richer, whereas the new one is impoverished by the new entrant. The result is that a firm which attracts many new enrollees during a given year (because of low prices) would mechanically have to increase its premium the following year in order to start creating reserves. If a client expects that switching plans will ultimately induce a subsequent rise in premiums (due to the building of reserves as the individual expects that the low-price plan will attract a lot of new consumers), then the expected utility of switching becomes lower and this may induce consumer inertia. Hence the rules obliging reserves creation may result in market failure.

Poor regulation in terms of the previous risk-equalization mechanism (see **Box 1**) may have also induced low-switching rates. When only age and gender were used as risk adjusters, the incentives for companies to practice risk selection were very strong. There was an incentive for insurance firms to try to avoid insuring unhealthy young people, whereas retaining healthy older enrollees. If such risk selection practices existed, they may have represented an additional barrier to switching and consequently may have prevented price competition from working properly.

Conclusion

Switzerland has implemented a relatively pure form of health care competition in which a great degree of free choice has been provided to the consumer. Nevertheless, the persistence of great variations within cantons in terms of insurance premiums, together with low-switching rates, raises the question about the effectiveness of competition in Switzerland in the basic insurance market. Several barriers to switching were identified, namely choice overload, status quo bias, the possession of supplementary contracts for enrollees in bad health, firm's pricing strategies based on providing low-price supplementary products, poor regulation of reserves, and the limitations of the previous risk-equalization system, which left room for profitable risk selection practices.

Do such inefficiencies imply that competition should be replaced by a single health insurance scheme? A referendum was held on such a proposal in 2007. It called for a merger of the existing 87 health insurance companies. Final results showed that 71% of voters opposed the reform. This vote demonstrated that the Swiss preferred the current system.

The analysis provided in Section 'Possible Barriers to Switching Behaviors' suggests some ways of improving the current system. First, evaluating an optimal number of insurance companies in the market may be useful, as the results suggest that having too great a choice effectively inhibits switching between health plans. Furthermore, economic analysis tells us that having a limited number of firms may be enough to achieve effective competition. Nevertheless, one would first need to assess whether, from a supply point of view, having a high number of firms induces each insurance company to achieve better efficiency.

Second, reforming the regulation of supplementary insurance could be an option. Our analysis illustrates that consumer choices for basic and supplementary health plans are not independent from each other. Although both types of insurance markets are regulated by two different laws and supervised by two different institutions, they are closely linked. Managed competition in the basic insurance market may suffer from a lack of adequate regulation in its supplementary counterpart. The two main policy options are either to separate these two markets more effectively (i.e., preventing firms from being active in both markets) or to regulate the supplementary insurance market differently, in particular, by preventing risk selection.

Third, there is probably room for improvement in the current regulation. Reserves could follow the individual when they switch.

In this article the focus is on Switzerland. Large premium variation and consumer inertia rates have also been highlighted in the Netherlands, which implemented a system sharing many features of the Swiss one in 2006. Since 2006, switching rates have been reported to be low in the Netherlands. It has also been found that the possibility for switching for bad-risk individuals in the basic insurance market is substantially reduced by the presence of supplementary insurance.

See also: Private Insurance System Concerns. Risk Equalization and Risk Adjustment, the European Perspective

References

Dormont, B., Geoffard, P. Y. and Lamiraud, K. (2009). The influence of supplementary health insurance on switching behaviour: Evidence from Swiss data. *Health Economics* **18**(11), 1339–1356.

Dormont, B., Geoffard, P. Y. and Lamiraud, K. (2013). Assurance maladie en Suisse: les assurances supplémentaires nuisent-elles à la concurrence sur l'assurance de base? *Economie et Statistiques* **455–456**, 63–79.

Frank, R. G. and Lamiraud, K. (2009). Choice, price competition and complexity in markets for health insurance. *Journal of Economic Behavior and Organization* **71**(2), 550–562.

Lamiraud, K. and Stadelmann, P. (2011). Strategic pricing behaviors in the presence of consumer inertia: The case of health insurance. Paper presented at the 2011 NBER Summer Institute.

Further Reading

Farrell, J. and Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. In Armstrong, M. and Porter, R. (eds.) *Handbook of industrial organization*, vol. 3, pp. 1967–2072. Amsterdam: Elsevier.

Synthesizing Clinical Evidence for Economic Evaluation

N Hawkins, Icon PLC, Dublin, Ireland, and University of Glasgow, Glasgow, Scotland

Glossary

Adjusted indirect comparison An analysis in which an indirect estimate of the average treatment effect between two treatments is estimated based on the results of randomized clinical trials that directly compare the two treatments to a common comparator.

Assumption of consistency The constraint applied in network meta-analyses such that $\delta_{AB} = \delta_{AC} - \delta_{BC}$ on the scale of analysis, where δ_{AB} is the indirect estimate of the effect of treatment A compared with treatment B, and δ_{Ac} and δ_{BC} are the direct estimates of the effects of treatments A and B compared with the common comparator treatment C.

Network meta-analysis An extension of pairwise meta-analysis that provides estimates of the relative effectiveness of two or more treatments derived from a statistical analysis of a connected network of clinical trial comparisons.

Introduction

As a vehicle for economic evaluation, model-based cost-effectiveness analysis offers major advantages over trial-based analysis. These include the facility of models to widen the set of options under comparison and to incorporate all relevant evidence. To achieve these, appropriate clinical evidence needs to be indentified and synthesized, particularly that relating to treatment effects on relevant endpoints. Meta-analysis is the field of clinical epidemiology, which focuses on evidence synthesis. Recent method development in this field has focused on the particular needs of economic evaluation to incorporate all the relevant evidence relating to all management options.

Limitations of Pairwise Meta-analysis

Meta-analysis is the process of using statistical techniques to synthesize the results from separate but related studies in order to obtain an overall estimate of treatment effect. Traditionally, randomized clinical trials (RCTs) have been meta-analyzed by combining results or data from a series of trials comparing the same two treatments. This has been referred to as pairwise meta-analysis. This form of analysis has a number of limitations. For example, it may exclude trials that potentially provide indirect information regarding the treatment effect of interest. This may lead to it being impossible to estimate a treatment effect or to a potentially important information being excluded from the analysis. In addition, where there are more than two treatments of interest it can be hard to interpret the results and associated uncertainty of a series of separate pairwise comparisons, particularly where the results of the individual analyses appear to be contradictory.

Network Meta-analysis

Network meta-analysis is an extension of pairwise meta-analysis that provides estimates of the relative effectiveness of two or more treatments derived from a statistical analysis of the data from a set of RCTs, where the trial comparisons form a connected network. The estimates of relative effectiveness, alongside estimates of uncertainty, based on a systematic synthesis of the available RCT evidence, are potentially an important component of medical decision-making.

Adjusted Indirect Comparison

The simplest form of network analysis has been referred to as an adjusted indirect comparison (AIC). In an AIC an indirect estimate of the average treatment effect between two treatments is estimated based on the results of RCTs that directly compare the two treatments of interest to a common comparator. It is adjusted in the sense that it allows for differences between trials in the response to the common comparator. Researchers have suggested the term 'anchored indirect comparison' as an alternative to adjusted indirect comparison to differentiate from those analyses in which covariable adjustment is included.

The indirect estimate is obtained by applying the constraint that $\delta_{AB} = \delta_{AC} - \delta_{BC}$ on the scale of analysis, where δ_{AB} is the indirect estimate of the effect of treatment A compared with treatment B, and δ_{Ac} and δ_{BC} are the direct estimates of the effects of treatments A and B, respectively, compared with the common comparator treatment C.

For example, consider a binary outcome with the constraint that $\delta_{AB} = \delta_{AC} - \delta_{BC}$ on the log-odds scale. An indirect estimate of the log-odds ratio for treatment A compared with B (LOR_{AB}) can be estimated as $LOR_{AB} = LOR_{AC} - LOR_{BC}$, where LOR_{AC} and LOR_{BC} are estimates of the log-odds ratios obtained from single RCTs or from a meta-analysis of multiple RCTs. This is equivalent to $OR_{AB} = OR_{AC}/OR_{BC}$, where OR refers to the odds ratios.

Estimating Uncertainty

The uncertainty in an adjusted indirect estimate can be estimated as the sum of the variances for each of the component direct estimates as these can be treated as independent

 doi:10.1016/B978-0-12-375678-7.01404-8

random variables (coming from separate RCTs). For example, $\text{var}(LOR_{AB}) = \text{var}(LOR_{AC}) + \text{var}(LOR_{BC})$. The standard error and 95% confidence interval can then be derived from the variance estimate. It should be noted that estimates of uncertainty calculated in this way only reflect sampling error and do not take into account the uncertainty as to whether the constraint $LOR_{AB} = LOR_{AC} - LOR_{BC}$ holds for the particular set of trial data. Estimates of uncertainty estimated in this way could be seen as representing the lower bound of uncertainty associated with an indirect comparison.

More Complex Networks

The constraint imposed in an adjusted indirect comparison can be applied to more complex connected networks of trial comparisons in order to obtain consistent estimates of treatment effects. A network is connected if all treatments are connected via direct or indirect comparisons. For example, trials comparing treatments A and B, B and C, and C and D form a connected network whereas trials comparing A and B, and C and D do not.

In a network meta-analysis, estimates of treatment effect are made that comply with the constraint $\delta_{AB} = \delta_{AC} - \delta_{BC}$ on the scale being used for analysis and best fit the observed trial data. These estimates may be obtained using the Bayesian Markov chain Monte Carlo or maximum likelihood methods.

Assumption of Consistency

The constraint that $\delta_{AB} = \delta_{AC} - \delta_{BC}$ has been referred to as an assumption of consistency (direct and indirect estimates are consistent), exchangeability (e.g., if treatments were exchanged between trials the estimated treatment effects would be the same, allowing for random variation), or transitivity (the relationship $\delta_{AB} = \delta_{AC} - \delta_{BC}$ is transitive, e.g., $\delta_{AC} = \delta_{AB} + \delta_{BC}$).

Although it is common to refer to an 'assumption' of consistency in an indirect comparison, it is unlikely that it is believed that the relationship $\delta_{AB} = \delta_{AC} - \delta_{BC}$ holds perfectly across a given set of trial data and that direct estimates comparing treatments A and B, if they became available, would conform perfectly (allowing for random variation) to indirect estimates. Rather, the consistency relationship is an approximation that allows us to make useful predictions based on the available data.

Choice of Scale

The assumption of consistency can be applied on different scales for the measurement of treatment effect. For example, on the log relative risk (LRR) scale ($LRR_{AB} = LRR_{AC} - LRR_{BC}$) or the risk difference (RD) scale ($RD_{AB} = RD_{AC} - RD_{BC}$). These assumptions of consistency are mathematical identities for a single RCT that include treatment A, B, and C (see Figure 1). However, they are not for independent trials comparing A and B, B and C, and A and C. For a set of independent trials, the assumption of consistency may hold on one scale but not another reflect (see Figure 2).

Network Geometry and Testing Consistency

In some networks both direct and indirect estimates of treatment effects may be possible for one or more comparisons. Such a network may be referred to as including loops. The analysis of a network including such loops may be referred to as a mixed treatment comparison (including a mixture of both direct and indirect evidence) and an analysis of a network without such loops as an indirect comparison. The term network meta-analysis is used to refer to any analysis of a connected network of trials evidence including both adjusted indirect comparisons and mixed treatment comparisons.

If the network of trial evidence includes loops it is possible to compare the direct and indirect treatment effect estimates within a network and hence 'test' the reliability of the

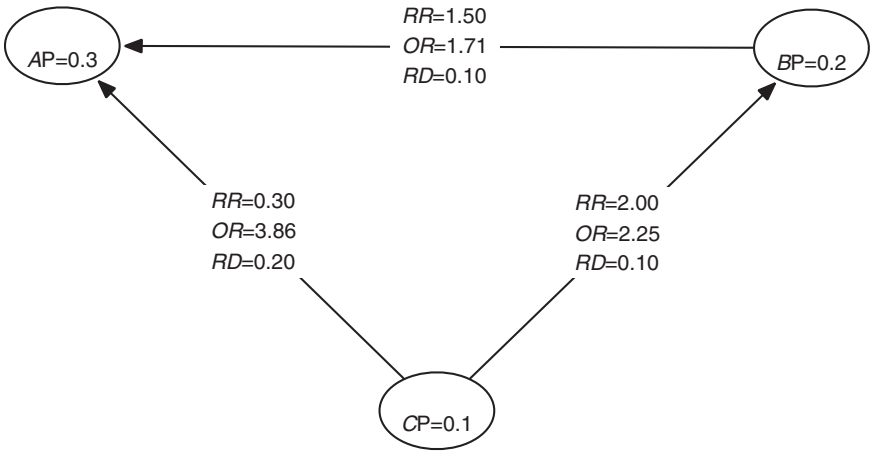


Figure 1 Consistency of a three-arm trial on various scales. Within this three-arm trial, treatment effects are, by definition, consistent on the relative risk: $\left(RR_{AB} = \frac{RR_{AC}}{RR_{BC}}; 1.5 = \frac{3}{2}\right)$, odds ratio $\left(OR_{AB} = \frac{OR_{AC}}{OR_{BC}}; 1.71 = \frac{3.86}{2.25}\right)$, and risk difference $(RD_{AB} = RD_{AC} - RD_{BC}; 0.1 = 0.2 - 0.1)$ scales.

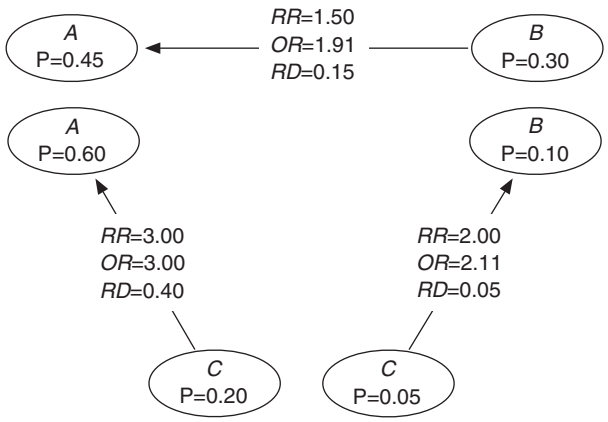


Figure 2 The treatment effect estimates from these three trials are consistent on the relative risk scale $\left(RR_{AB} = \frac{RR_{AC}}{RR_{BC}}; 1.5 = \frac{3}{2}\right)$, but not on the odds ratio $\left(OR_{AB} = \frac{OR_{AC}}{OR_{BC}}; 1.91 \neq \frac{6.00}{2.11}\right)$ and risk difference $(RD_{AB} = RD_{AC} - RD_{BC}; 0.15 \neq 0.3 - 0.05)$ scales.

consistency constraint – at least with respect to those comparisons for which direct and indirect evidence is available. The direct and indirect estimates can be compared in using formal hypothesis tests. It should be noted, however, that even where the direct and indirect estimates for a given comparison are consistent, the component treatment effect estimates may not all be exchangeable; the effect of treatment effect modifiers may cancel out across the network.

Network Meta-analysis in Practice

Any connected network can be analyzed using the techniques of network meta-analysis. The confidence in the utility of analysis will depend on the empirical evidence regarding the likely deviation of true values of treatment effects from the consistency constraint. If there are material differences between trials in terms of study design (including endpoint definition) or subject characteristics this will reduce the confidence in the analysis.

If there is evidence from subgroup or regression analysis within individual trials that those factors that vary between trials modify treatment effects, this will further reduce confidence in an analysis. Conversely, if these factors do not appear to modify treatment effects within trials this will increase confidence in an analysis.

If there is sufficient overlap in the range of subject characteristics between trials, it may be possible to adjust for the effects of treatment effect modifiers based on an analysis of within-trial variation of response. This may take the form of stratified or regression analysis. Propensity score methods have also been used where individual patient level data analysis is available for some trials and aggregate data for others.

If there are sufficient trials meta-regression analysis of the variation in treatment effects between trials and study characteristics or subject characteristics aggregated at the study level can also be used to adjust for heterogeneity between trials.

If empirical adjustment for heterogeneity between trials is possible, this may increase the confidence in the results in an analysis based on indirect comparisons.

Alternatives

Finally, when considering the utility of a network meta-analysis, it is important to consider the alternatives. Decisions could be based solely on analyses of direct evidence. However, the available direct evidence may be contradictory and difficult to interpret in a piecewise manner. In contrast, a network meta-analysis will provide estimates that are consistent with readily interpretable estimates of uncertainty.

In many cases direct evidence may not be available, in this case the decision could be deferred until direct evidence becomes available (although when it does it may be inconsistent with the existing indirect evidence); some assumption of equivalent effectiveness might be made; or the response to treatment in individual trial arms compared. The latter is termed a naive indirect comparison.

Whereas a naïve indirect comparison will be confounded by factors that affect the response to treatment in individual trial arms, an adjusted indirect comparison or network meta-analysis based on treatment effect estimates from RCTS analyzed as randomized will only be confounded by factors that act as treatment effect modifiers. Factors that affect response in individual treatments arms, but do not alter the average treatment effect on the scale used for analysis, will not confound network meta-analyses. This has led several commentators to recommend network meta-analysis or adjusted indirect comparisons over naïve indirect comparisons.

It should also be noted that any use of RCT evidence for decision-making infers some form of exchangeability between the subjects within the trial and those patients who are the object of the decision-making process. If multiple trials are viewed as being relevant to the decision-making process for an individual patient, this infers some degree of exchangeability between these trials. Network meta-analysis could be seen as the formalization of this exchangeability.

Conclusions

Network meta-analysis is increasingly used as a framework for estimating the effects of the full range of relevant options. This article has considered the principles of these methods and, in particular, the underlying assumptions needed for reliable estimates.

Further Reading

Caldwell, D. M., Ades, A. E. and Higgins, J. P. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *British Medical Journal* **331**(7521), 897–900.

Caldwell, D. M., Welton, N. J., Dias, S. and Ades, A. E. (2012). Selecting the best scale for measuring treatment effect in a network meta-analysis: A case study in childhood nocturnal enuresis. *Research Synthesis Methods* **3**(2), 126–141.

Higgins, J. P. T., Jackson, D., Barrett, J. K., et al. (2012). Consistency and inconsistency in network meta-analysis: Concepts and models for multi-arm studies. *Research Synthesis Methods* **3**(2), 98–110.

Hoaglin, D. C., Hawkins, N., Jansen, J. P., et al. (2011). Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR Task Force on indirect treatment comparisons good research practices: Part 2. *Value in Health* **14**(4), 429–437.

Jansen, J. P., Fleurence, R., Devine, B., et al. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR Task Force on indirect treatment comparisons good research practices: Part 1. *Value in Health* **14**(4), 417–428.

Jones, B., Roger, J., Lane, P. W., et al. (2011). Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* **10**(6), 523–531.

Lu, G. and Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**(20), 3105–3124.

Lu, G. and Ades, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**(474), 447–459.

National Institute for Health and Care Excellence Technical Support Documents Evidence Synthesis Series. Available at: http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series(2391675).htm (accessed 28.06.13).

Salanti, G., Marinho, V. and Higgins, J. P. T. (2009). A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of Clinical Epidemiology* **62**(8), 857–864.

Signorovitch, J. E. J., Sikirica, V., Erder, M. H., et al. (2012). Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research. *Value in Health* **15**(6), 940–947.

Welton, N. J., Sutton, A. J., Cooper, N. J. and Abrams, K. R. (2012). *Evidence synthesis for decision making in healthcare*. London: Wiley.

Theory of System Level Efficiency in Health Care

I Papanicolas, London School of Economics, London, UK
PC Smith, Imperial College, London, UK

Introduction

In recent years there has been an increased interest in the notion of the health 'system,' the ultimate goal of which is to protect and improve the health of its population. The definition of the health system is contested, but a frequently invoked starting point is the World Health Report in 2000, which "... defines a health system to include all the activities whose primary purpose is to promote, restore or maintain health" (World Health Organization (WHO), 2000, p. 5). Systems level efficiency is concerned with understanding how well a specific system is using the resources at its disposal to improve health and secure related objectives. Identifying inefficiencies in either the system or in its component parts is important as it allows the same objectives to be attained with fewer resources (or alternatively it enables the system to produce more with the same resources). As spending on health care continues to rise remorselessly in all developed countries, this issue has become increasingly relevant for policy makers seeking ways to pursue health objectives at the same time as containing cost pressures.

Although the core idea of efficiency is easy to understand in principle – maximizing valued outputs relative to inputs – it becomes more difficult to make operational when applied to a concrete situation, particularly at the system level. It is, for example, quite conceivable that there are efficiently functioning components operating within an inefficient broader health system. Furthermore, efficiency is not strictly determined by the relationship between the physical quantities of inputs and outputs, but by the value attached to those outputs. Indeed, by analogy with the first law of thermodynamics, which postulates that energy cannot be created or lost, the ratio of all physical outputs to all physical inputs will necessarily remain unchanged. What will differ, according to the structures, processes, and systems in place, is the ratio of valued outputs to inputs.

The general assumption in a great deal of the health economics literature is that the objective of the health system and its component parts is to increase the length and health-related quality of life of the population. This is most famously embodied in the notion of a 'quality-adjusted life-year' (QALY). However, there may be other very important objectives attached to a health system, such as reducing disparities in health, protecting citizens from the financial consequences of illness, and improving the responsiveness of health services to personal preferences.

Valuations attached to different health system outputs can vary because of variations in individual preferences, the decision-making perspective being used, or even because of the level of analysis being applied. As a result, a number of sometimes conflicting definitions for 'efficiency' exist in the economics and policy literature, and even within health economics itself (**Table 1**).

Although each definition attempts to clarify the nature of inputs and outputs, the variety of perspectives illustrates that there is no consistent approach. In particular, there is considerable variation as to what the valued outputs are, including: Volume of care, quality of care, levels of quality, 'performance' and health improvement; and this reflects the lack of clarity as to the concept of 'valued outputs.' Throughout this article the authors generally adopt the assumption that 'health improvement' is the valued health system output. However, on occasions reference to some of the other legitimate objectives that society may attach to the system shall be made.

One terminological issue needs to be addressed: The productivity literature usually refers to the products of a production process as 'outputs.' In health care, it has become conventional to refer to the physical products (such as an episode of hospital care) as an output, but to the health benefits achieved as 'outcomes.' Thus, health outcomes can be thought of as the 'value' attached to an output. Throughout this article the authors use the term 'outputs' and where necessary 'valued outputs' unless it is specifically necessary to refer to the health benefits achieved.

Although there appears to be more consistency regarding system inputs, specifications are often quite vague, referring simply to 'costs' or 'resources.' Only a few of the definitions in **Table 1** identify particular inputs, such as "The relative quantity, mix and cost of clinical resources" (Pacific Business Group on Health, 2006, p. 2). The lack of clarity about the concepts of 'valued outputs' and inputs reflects a wider ambiguity about the organizations and production processes of the health system. In particular, it is often not clear which resources and health outcomes are considered to fall within the responsibility of the health system. This article therefore begins by discussing the health system definition and its relationship to health system efficiency. Key concepts relating to efficiency such as productivity, technical and allocative efficiency, and their relationship to the production frontier are considered. These theoretical concepts are then related back to health policy to indicate the types of questions that might be considered in any analysis of technical or allocative efficiency.

What is the 'Health System'?

To understand and measure efficiency, it is first necessary to define the scope of the entity under scrutiny. Health is the product of numerous determinants, some that can be directly influenced in the short term by factors in the health services (e.g., improving medical care), others that require long-term action of factors not directly associated with health services (e.g., environmental policy), and yet others that depend primarily on the actions of individuals and their families (e.g., diet). If the health system is assumed to be comprised only of

 doi:10.1016/B978-0-12-375678-7.00203-0

Table 1 Alternative definitions of efficiency in a health care context

Key concept	*Definition*
Efficiency of care	A measure of the cost of care associated with a specified level of quality of care (AQA Alliance, 2006)
Efficiency of care	A measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality (Institute of Medicine (IOM), 2001)
Efficiency of care	A measurement construct of cost of care or resource utilization associated with a specified level of quality of care (National Quality Forum, 2007)
Efficiency	The relative quantity, mix and cost of clinical resources used to achieve a measured level of quality (Pacific Business Group on Health, 2006)
Efficiency	An attribution of performance that is measured by examining the relationship between a specific production of the healthcare system (also called output) and the resources used to create that product (also called inputs) (RAND, 2008)
Efficiency	A measure of the cost at which any given improvement in health is achieved. If two strategies of care are equally efficacious or effective, the less costly one is more efficient (Donabedian, 1990)
Efficiency	Technical efficiency (or production efficiency), getting the maximum output for money, and allocative efficiency, producing the right collection of outputs to achieve goals, or being on the production possibility frontier (Roberts et al., 2008)
Efficient (not wasteful) care	Delivery and insurance administration, delivered at the right time and right setting and where new innovations can be evaluated for both effectiveness and value (Commonwealth Fund, 2006)
Health system efficiency	Production efficiency, the combination of inputs required to produce care and related services at the lowest costs, and allocative efficiency, the combination of inputs that produce the greatest health improvements given the available resources (Aday et al., 2004)
Health system efficiency	Microeconomic efficiency, measured health system productivity as compared to its maximum attainable, Macroeconomic efficiency, what effect a change in the level of resources would have on the desired level of health outcomes and responsiveness compared to other goods and services (Hurst and Jee-Hughes, 2001)
Health system efficiency	Actual (health system) goal attainment achieved related to what could be achieved given the resources available (WHO 2000, 2007)

Source: Adapted with permission from Chung, J., Kaleba, E. and Wozinak, G. (2008). A framework for measuring healthcare efficiency and value. *Working Paper Prepared for the Physician Consortium for Performance Improvement. Work Group on Efficiency and Cost of Care*; and Papanicolas, I. (2013). Frameworks for international comparisons. In Papanicolas, I. and Smith P. C. (eds.) *Performance comparisons for health system improvement*, pp. 31–75. Maidenhead: Open University Press.

health services, then actions that may have a greater impact on health are excluded (such as education or employment). Thus under a narrow definition of the health system, confined to health services, an analysis of system level efficiency may not consider the possible efficiency gains that could be secured by allocating resources differently between areas such as health care, education, and housing. However, a broader analysis can rapidly lead to lack of clarity of the boundary of the system, and associated difficulties with measurement and attribution.

For example, **Figure 1** illustrates the potential production process of a health system, with examples of costs and physical inputs put into the system and a selection of outputs and consequent outcomes that are produced. The different shades represent different boundaries of the health system, starting from a consideration of only medical care and extended to consider all factors that influence health. Across these boundaries, many of the outcomes of the system do not change – for example, health improvement and risk protection are outcomes arising from medical care, public health and health promotion, intersectoral action, as well as from economic growth and public sector investment. However, the physical inputs that contribute to the attainment of these outcomes will differ markedly depending on the choice of system boundaries.

In an evaluation of health system efficiency it is important not only to consider the physical inputs that correspond to the health system as defined, but also to ensure that the outcomes being assessed also represent only the contribution attributable to those particular inputs. For example, if the efficiency of medical care were to be assessed it would be crucial to isolate the contribution of medical care to health improvement, and to adjust for any contribution of other activities such as public health, development, and education.

A particular issue that sometimes arises is whether to scrutinize only publicly owned or financed institutions, or to include also the private and not-for-profit sector in an analysis of health system efficiency. The position of the World Health Organization is clear on this – they assign accountability to the government for the entire health sector, however, organized (World Health Organization (WHO), 2000). Under this formulation, the regulation and performance of privately funded healthcare should be included in any analysis. This reflects the WHO emphasis on governments as 'stewards' of population health.

In defining the health system, it is therefore important for analysts to be clear about what allocation of resources needs to be considered, what parties will be affected, and who will be making the decisions regarding allocation. Ideally the definition should be aligned with the factors under the control or influence of a responsible person or organization, such as the health minister. The scope of this accountable entity may therefore vary depending on a country's institutional arrangements. In other words, the definition of the health system should reflect a country's accountability arrangements. Whatever definition is chosen should then determine the scope and perspective of the analysis. It is then crucial that all definitions of inputs, outputs, and exogenous influences on attainment are aligned with that choice. This may lead to methodological

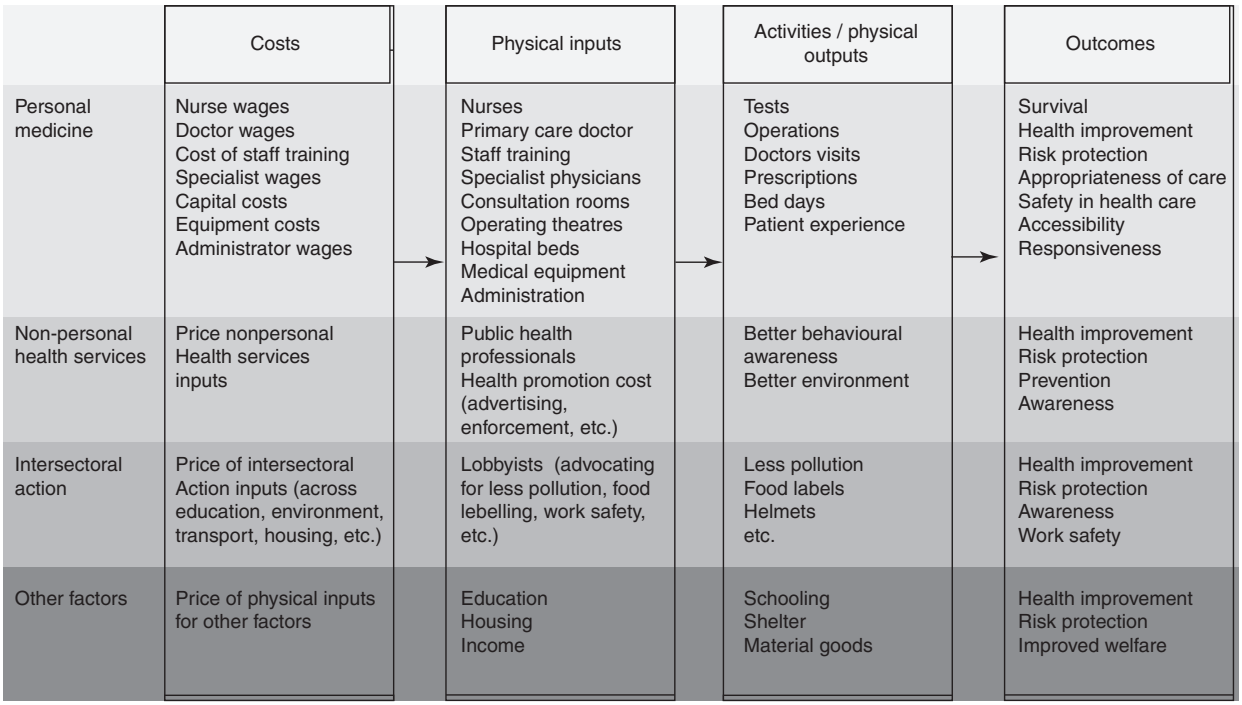


Figure 1 Selected inputs and outputs of the health system at different boundaries.

challenges when attempting to compare one country with another where there exists no universal agreed categorization of diseases, health care procedures, health care organizations, or even health systems (Cylus and Smith, 2013).

Efficiency at Different Levels

Reflecting the wide range of potential perspectives, economists and policy makers have adopted different conceptualizations of efficiency when analyzing different levels of the health system. At the formal organizational level, definitions usually refer to the extent to which health service objectives have been achieved compared to the maximum that could be attained, given the resources available and the external constraints on attainment. The conventional concepts of allocative, technical, and economic efficiency are then used to describe efficiency at this level (see section The Elements of Efficiency).

However, the concept of efficiency may change if the perspective changes. It is possible to take a broader view of sector level resource allocation, for example, by considering the level of resources devoted to the health system relative to other sectors that can also provide a positive contribution to health or indeed to broader welfare. The Organization for Economic Cooperation and Development refers to this broader conceptualization of efficiency as macroeconomic efficiency (Hurst and Jee-Hughes, 2001). In principle, the size of the health sector should be determined by its marginal contribution to welfare, relative to other sectors. Work on macroeconomic efficiency might therefore examine whether healthcare expenditure has reached levels where marginal spending on medical services contributes less to welfare than if it were directed at other sectors, such as education, housing, the environment, or private consumption.

At the other extreme, resource allocation can be considered at the very micro level, for example, by guiding the decisions of individual clinicians on how to distribute healthcare resources across treatment options in order to maximize valued outputs. Study of this type of efficiency often takes the form of a systematic analysis of the effects and costs of alternative methods or programs for achieving the same objective (e.g., improving quality of live, extending year of life lived, or providing services). These methods include cost-effectiveness analysis (CEA), comparative effectiveness analysis, cost-benefit analysis, and cost utility analysis. In principle, these techniques can be crafted to reflect different personal preferences, although in practice they usually produce guidance on the basis of a uniform set of preferences.

Thus, health system efficiency can be characterized in very broad terms at different levels of analysis, as adapted from Chung *et al.* (2008):

- Physician level: How to distribute resources between interventions in order to choose a treatment strategy that maximizes the patient's value or utility given existing resources.
- Organizational level: How to distribute resources within a healthcare organization and across health care processes in order to maximize aggregate health gain and satisfaction.
- Systems level: How to distribute resources within the broad components of the health system (such as prevention, primary care, secondary care) so that production of healthcare maximizes citizen health and welfare.
- Societal level: How to distribute resources between sectors of the economy (education, health, public goods, etc.)

such that health and broader welfare within society are maximized.

These issues are addressed in the section Allocative and Technical Efficiency as Applied to Health Policy.

The Components of the Health System

The different levels of the health system are profoundly interdependent, as actions at one level will often influence behavior and outcomes elsewhere. Bringing together findings from studies at the different levels can provide fundamental tools for understanding how an entire system is managed and sustained. The interconnectedness of the separate areas of study of a health system can be illustrated by the distinct areas of health economics set out in Alan Williams' 'plumbing diagram' reproduced in **Figure 2** (Williams, 1987).

Each box in the diagram represents one of the subdisciplines of health economics. The various fields are connected to one another by 'pipes' with the arrows indicating the direction of the relationship. Each of the boxes, discussed briefly in **Figure 2** involves both positive and normative issues and different societal preferences (Maynard, 2007). In relation to the discussion of efficiency, Boxes A, B, C, and D represent the key factors that determine the initial allocation of resources by society and what inputs and outputs are available and valued within health systems. These factors create the framework for the analysis of efficiency at the patient, organization, and system level represented in Boxes E, F, G, and H.

Box A models the determinants of health, focusing particularly on social and behavioral determinants that lie beyond the immediate control of the health system. These will be important when considering the system boundary issues discussed in the sections What is the 'Health System'? and Efficiency at Different Levels. It is important to understand what is meant by 'health' and how it is valued by individuals and society. Thus, Box B represents the study of perceptions of health as well as the valuations of its different states and life. These are important to the study of efficiency in order to be able to assign value to the outputs being produced at the patient, organization, system, and society levels, and inform decisions on the allocation of resources.

Boxes C and D consider the demand for and supply of health care respectively. In any market, supply and demand will determine how the market allocates resources from producers to consumers. In health services, demand is driven by patients seeking health care to improve their health status and longevity. This can be influenced by information asymmetries between patients and providers, or providers and third-party payers, barriers to care such as cost, distance or culture, and externalities of care. Supply of health care considers a wide

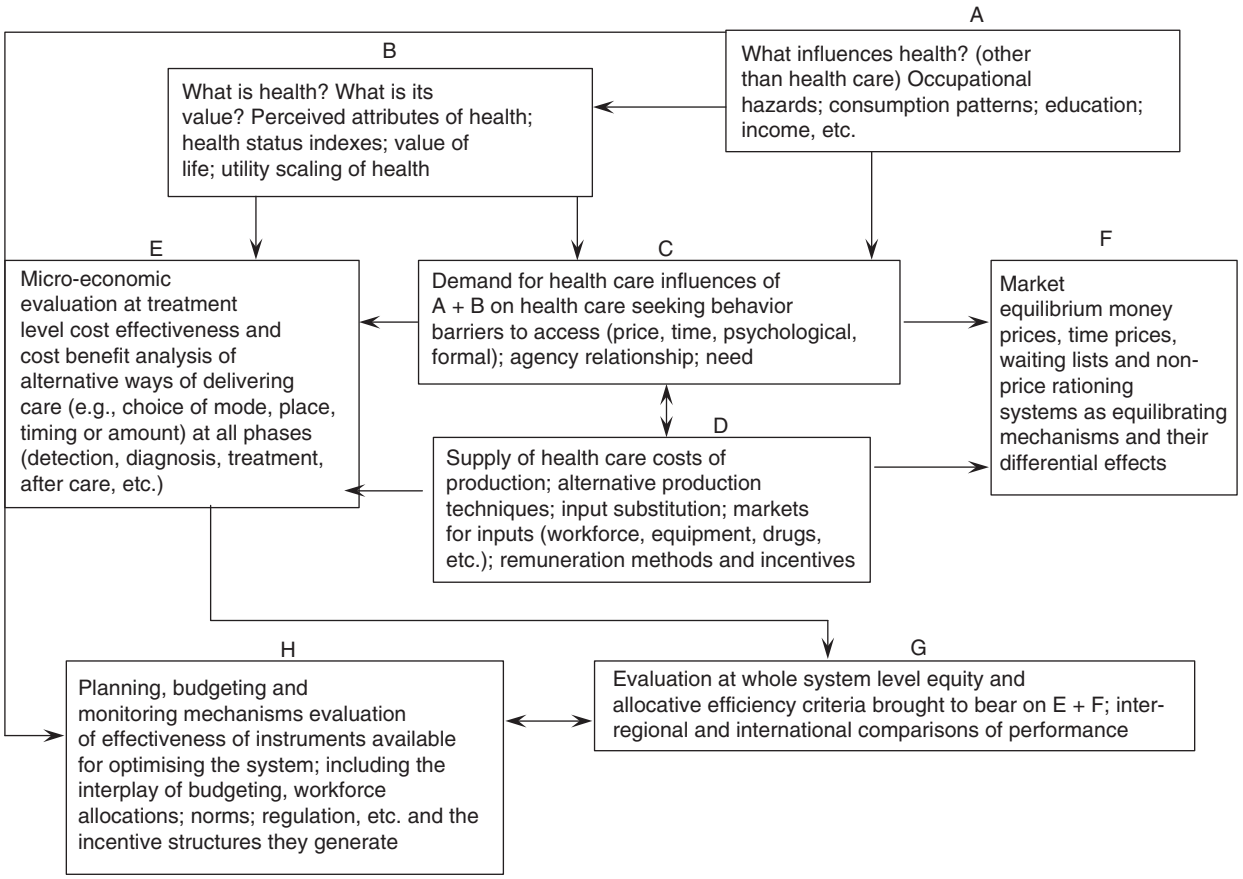


Figure 2 Williams' Health Economics Plumbing Diagram. Reproduced with permission from Williams, A. (1987). Health economics: The cheerful face of a dismal science. In Williams, A. (ed.) *Health and economics*, pp. 1–11. London: Macmillan

range of factors including the use of private and public suppliers, the behavior of institutions and providers, the skill mix and structures available to provide health, the financing structures in place to provide funds, as well as the organization of service delivery. It is important to understand the factors that influence demand and supply in order to be able to interpret how the market allocates resources and how this can be influenced to correct inherent market failures and associated inefficiencies.

Boxes E, F, G, and H represent the application of core economic principles, including the study of efficiency at different levels. Box E is concerned with the microeconomic evaluation at treatment level, and reflects the impact on efficiency of 'physician level' decisions outlined in the section Efficiency at Different Levels. This area of health services research involves conducting microevaluations such as CEA or cost utility analysis to inform rationing and reimbursement choices. They require information on clinical effectiveness of treatments, as well as the measurement of costs and value added.

Box F addresses the issue of market equilibrium, where market demand equals market supply. Private and public markets will in general clear at different levels, depending on factors such as pricing of goods and services, restrictions of benefit packages, rationing, and waiting times. This area of research encompasses issues of organizational level efficiency, where different production processes, input prices, and willingness to pay influence decisions of what and how much of different goods and services should be produced.

Finally, Box G represents the evaluation of the whole system, concerned with understanding how inputs are used to achieve the key objectives of the health system. This includes the area of system level efficiency. Box G will be influenced by planning, budgeting and monitoring mechanisms represented in Box H. All three of boxes F, G, and H can embrace regulatory mechanisms intended to correct market failures and maximize the optimum allocation of resources in order to meet society's health system goals.

Although each box constitutes a separate area of study, the 'pipes' between the eight boxes show some of the linkages between the different areas, indicating their interdependence. For example, the operation of the hospital would be best considered in Box F. Yet the hospital's market equilibrium in Box F would be influenced by the supply and demand for health care (Boxes D and C) as well as the severity of patients being treated (Box A), and the acceptable forms of treatment administered (Box B). Moreover, although the hospital could be perfectly technically efficient at the organizational level, it could be operating in a very inefficient health system, represented by Box G, if the overall health improvements made at the system level are below what could be achieved if current health expenditures were reallocated – say between hospital and preventive services. This may be related to the mechanisms in place for planning and budgeting, represented in Box H, for example, if too much money is being allocated to secondary care as opposed to primary care.

The health system is complex to analyze, because of people's heterogeneous health needs, the enormous scope for market failures, and the interdependencies illustrated in Figure 2. The assessment of efficiency is therefore also challenging. In the next section the authors set out some basic building blocks needed for the successful examination of health system efficiency.

The Elements of Efficiency

The underlying aim of efficiency analysis is to understand how inputs are translated into valued outputs. In this section the authors seek to clarify some of the different general concepts of efficiency as they apply to the health system.

Productivity and efficiency are often used interchangeably, but they refer to slightly different concepts. Productivity refers to the ratio of a (possibly partial) measure of output to a (possibly partial) measure of input. In contrast, efficiency seeks to assess the attained level of output in relation to the maximum that can be produced, given the inputs used, system constraints and available technology. Efficiency will often be calculated taking into account constraints (such as scale) that inhibit improved productivity. Ideally, efficiency will express the outputs being produced in terms of their value to consumers or society. Thus, there is an implication that efficiency is a more comprehensive and normative tool than productivity.

Fundamental to the study of efficiency is the concept of production function, which models the maximum possible level of outputs for given levels of inputs, given current technology. Alternatively, it is sometimes convenient to model production possibilities in the form of a cost function, which models the minimum feasible cost of producing a given set of outputs. In reality, for most production processes, there are both multiple inputs and multiple outputs, and it is more accurate to think in terms of a production possibility frontier, which maps the maximum levels of output attainment for any mix of inputs. Whether a production function or cost function perspective is adopted usually depends on the specific focus of the study. In what follows the focus is mainly on the production function.

Whatever perspective is adopted, efficiency analysis can be considered broadly as the study of two main questions:

1. Are resources being used so that the maximum level of chosen outputs is produced given the available inputs? (or are resources being used so that the minimum level of inputs are used to produce the chosen outputs?) That is, is the entity located on the production frontier, rather than inside it?
2. Is the 'right,' or most valued, mix of outputs being produced, given society's valuation of those outputs. Or conversely, is the 'right,' or minimum cost, mix of inputs being used, given the chosen outputs. That is, is the entity located at the maximum value (or minimum cost) point on the production frontier?

These two questions relate respectively to technical and allocative efficiency, which are now considered in turn.

Technical Efficiency

Technical efficiency indicates the extent to which an entity is producing the maximum level of output for a given level of inputs under the prevailing technological process, therefore

addressing the first question of efficiency analysis posed in the section The Elements of Efficiency. To identify whether the health system is technically efficient, it is thus necessary to determine four key characteristics:

1. What are the inputs of the health system?
2. What are the outputs of the health system?
3. What is the maximum level of health system output that can be produced for different levels of input?
4. What are the external constraints that may limit the ability of the health system to be technically efficient?

The discussion in the section Efficiency at Different Levels illustrated the challenges involved in defining the boundaries of the system under scrutiny and identifying relevant inputs and outputs. In particular, the chosen definition of the health system will determine whether factors such as public health, health promotion, and socio-economic determinants of health are considered in the analysis. The choice of boundaries will be crucial in determining first what resources make up the health system inputs, and second what are considered uncontrollable exogenous constraints on attainment.

With regard to health system outputs, a key challenge relates to the differences in stakeholder perceptions about what the health system should be producing. International and national health system frameworks identify a number of potential health system objectives, some of which are almost universally recognized (such as health improvement) and others where views differ (such as the extent to which a system offers patients choice of provider). Many definitions of efficiency require some measure of quality as well as volume of outputs (**Table 1**), yet notions of what constitutes quality of care are sometimes vague and conflicting (Papanicolas, 2013).

Once inputs and outputs have been identified, it is necessary to identify the maximum level of health system output that can be produced for different levels of inputs, using the concept of the production frontier. Any unit lying on the production frontier will be technically efficient. Inefficient units lie strictly within the frontier. In the context of health systems, therefore, a technically efficient system will be one that produces the maximum achievable level of valued outputs (in the form perhaps of health outcomes) given its inputs (or uses minimum level of inputs, given its outputs). Although this tool serves well in understanding the theory of systems level efficiency, it poses many practical challenges in the empirical estimation of the production function.

A central concern in estimating the production function is the identification of external constraints – the uncontrollable influences on attainment that limit the production of desired outputs. At the health system level such constraints might include: the underlying health of the population, the configuration of provider organizations, and the skills and size of the workforce. In the short term, many of these factors can be considered genuinely exogenous influences on levels of attainment, and so should be included as constraints in the analysis. In the longer term, it might be expected that the health system should be responsible for addressing some of the constraints. So, for example, an inefficient scale of providers might be an acknowledged handicap in the short run, but should not be considered as an 'excuse' for poor attainment in the longer run. However, some constraints, such as the physical terrain of a geographical area, might be considered truly exogenous.

A final key consideration that is rarely modeled satisfactorily is the dynamic nature of the health system. Outputs measured at one point in time will have been influenced by inputs of a previous time period, and similarly inputs in a current time period will to some extent influence outputs in a future time period. Ignoring dynamic aspects of efficiency may incorrectly attribute all current performance to current actions and hold stakeholders accountable for past actions for which they may not have been responsible. In principle, the proper approach to longer term investments in (say) disease prevention is to treat them as a capital investment, the benefits of which accrue over several time periods. They can either be treated in the same way as more conventional investments in physical capital, or by using proxies for the future benefits of current investments (e.g., expected QALYs gained per person immunized). By properly including a time dimension, analysis of efficiency may even provide improved incentives for policies with long-term effects, as future benefits will be recognized even in the short term.

Allocative efficiency

Technical efficiency examines the extent to which the unit is failing to reach the production frontier, as expressed in the cost or production function. In contrast, allocative efficiency examines whether production is allocated across either inputs or outputs so as to maximize the value to society. This principle can be interpreted in a number of ways. However, the common theme is that allocative efficiency refers to the extent to which a socially optimal point on the production frontier has been reached. In a conventional market, market prices can be indicative of the value of goods and services according to the trade-off consumers or society are willing to make between them. In input space this is readily transferred to health systems, where allocative efficiency can be interpreted as the extent to which the minimum cost of inputs is being used, given the market prices of those inputs. For example, to what extent is the right mix of clinical skills, physical inputs and medical products being used to secure system objectives, given prevailing wage rates, property prices, product prices, and so on.

However, in health systems, there are rarely market prices for outputs. It therefore becomes more difficult to define the relative value of outputs, as a guide to identifying the 'right' mix of outputs. As a result, in output space a number of definitions of allocative efficiency have arisen. Roberts *et al.* (2008) are, however, representative in defining allocative efficiency as whether a nation is producing the right mix of outputs to maximize attainment of its overall goals.

Thus, technical efficiency refers to the question of how goods are produced given certain inputs, meaning that a technically efficient point is one that lies anywhere on the production possibility frontier. At such a point a provider can produce more of one output only by reducing production of another. Allocative efficiency, however, refers to the question of what inputs are used or what outputs are produced, and suggests that there is a unique point on the production frontier that maximizes societal values relative to all other

attainable sets. To do this it is (in principle) necessary to specify a 'social welfare function,' which aggregates societal preferences into a single measure of the benefits to society of a social program.

One therefore needs information on the relative value to society of different health system outputs. There are many possible approaches to identifying those values, based on competing theories of justice. Various types of market mechanism, technical analysis, and political process seek to address this challenge. Moreover, even if there is agreement on which approach to adopt, individuals will always hold different values about what are the 'right' outputs to be produced. This diversity in normative perspectives and individual values relating to the allocation of health resources often makes it difficult to agree on a common starting point.

Nevertheless, if policy makers had knowledge of individual utility functions and preferences, they could in principle specify a social welfare function that aggregates the utilities across members of society. One could then examine the problem of efficiency as one of welfare maximization – that is, finding the feasible allocation of resources that maximizes a chosen concept of social welfare. This approach has been labeled 'welfarism.'

However, in practice construction of a social welfare function is extremely challenging, and systematic aggregation of individual preferences is infeasible. Yet – given the importance of health care in society – policy makers must make allocation decisions, and they have therefore used criteria other than traditional welfarism to inform resource allocation, such as potential health gain, need for treatment, demand for treatment, or simply cost. This approach is often referred to as extra-welfarism. This school of thought rejects the notion of using only utility as the outcome of interest, but seeks to consider broader factors such as an individual's capabilities: the goods and services that enable individuals to flourish. The dominant application of extra-welfarism in health economics has been concerned with aggregate health maximization, an approach that allows the aggregation and comparison of individual benefits.

Allocative and Technical Efficiency as Applied to Health Policy

To illustrate the points raised in the section The Elements of Efficiency, Table 2 considers the outputs, inputs, and trade-offs that might be considered in an analysis of technical or allocative efficiency within the health system. The provider (micro) level considers an individual seeking treatment where the inputs being considered are the money spent on treating the patient, and the output is the heath gain. The most efficient point of provision hinges on the choice of treatment. The intention is to treat the patient with the most cost-effective treatment (allocative efficiency), with maximum effectiveness (technical efficiency), within budget constraints. The choice of the allocatively efficient point should in principle consider the patient's preferences for types of treatment (e.g., attitudes toward pain and risk aversion), and offer the treatment that (subject to expenditure constraints) maximizes the expected value of the treatment to the patient. Given the lack of information about potential outcomes for many (if not most) treatments and the practical difficulty of crafting treatment choices specific to each patient, lack of information, and knowledge of medical care and types of treatment, clinical guidelines and payment mechanisms set at a higher organizational level may be very important in ensuring a technically and allocatively efficient distribution of resources.

At the organization (meso) level, assuming the only goal of the health system is to maximize health, the technically efficient point would be where expected health gain is maximized within each organization, by offering an optimal mix of cost-effective treatments to patients with different conditions. To achieve the allocatively efficient point mix of outputs, there needs to be a way to measure health gain across different potential treatments that the organization could provide, perhaps in the form of the QALY. Note that the 'mission' and design of organizations may constrain the range of output options available – for example, a hospital may not be capable of delivering a health promotion program. To overcome organizational boundaries, a correct allocation of resources must be made at the higher (system wide) level.

Macro concerns are represented in Table 2 by the system level and the societal level. The system level considers how resource allocation decisions are made to maximize health within the health system, and the societal level considers how resources are allocated between health and other sectors of the economy. For the system, a technically efficient point might be one that allocates resources across health services so as to maximize health gain. The allocatively efficient point is the one that provides the combination of health services that maximizes aggregate health gain across all

Table 2 Examples of allocative and technical efficiency in health systems

Level of analysis	*Technical efficiency: How are inputs used to produced outputs*	*Allocative efficiency: What mix of outputs is produced*	*Decision mechanisms*
Provider level	Maximizing personal outcomes, subject to cost constraint	Medical or surgical treatment of cancer	Patient consultation
Organizational level	Maximizing health gain in relation to expenditure	Providing hip replacements or cataract surgeries	Priority setting
System level	Minimizing avoidable mortality in relation to health system expenditure	Investing in curative or preventative care	Budgetary processes
Societal level	Maximizing life expectancy in relation to gross domestic product	Investing in health or education	Elections

services. At the societal level, the problem is analogous, but the health gain can be from different sectors of the economy, such as education or housing. Addressing this intersectoral allocation problem in principle requires consideration of the relative value of nonhealth objectives of other sectors, especially if these are produced jointly with health-enhancing programs.

Finally, Table 2 also considers possible decision mechanisms that can be used to determine what mix of outputs will be produced. For example, at the provider level the optimal mix of outputs will be informed by cost constraints within the system and individual preferences for treatment. The decision processes at the organizational level, such as a hospital, will be informed by the priority setting system in place. This priority setting system may differ across health systems, and may be based on factors such as the needs of the population being treated, the costs of inputs, the cost-effectiveness of treatments, or historical trends in purchasing. At the macro level, the resource allocation decision mechanisms in common use attempt to reflect the preferences of society, often through political processes. At the systems level, where the budget must be allocated to resources across the entire health system, the decision mechanism will often be made by bureaucratic processes overseen by elected representatives. At the societal level, decisions arise from a mix of private markets and political decisions about how to spend public finances, typically arising from some sort of election process.

Although Table 2 may help to conceptualize what seem like abstract notions, it takes a highly simplified view. It assumes an extra-welfarist perspective, in the sense that the only valued output of the health system is health gain. In practice, the health system often has broader objectives, and in principle a comprehensive analysis would examine the impact of allocations on broader social welfare. The market for health care makes it inherently prone to inefficiency if left to its own devices. Key market failures include: The lack of competition, especially for the provision of health services; the information asymmetries between patients and physicians, and between physicians and third-party payers; and externalities that are not reflected in prices (such as the spillover benefits offered by vaccinations). These market characteristics may, in the absence of corrective mechanisms, result in highly inefficient allocations of resources. The last column of Table 2 therefore indicates the sort of mechanisms necessary to promote allocative and technical efficiency at the different levels of analysis, given such market failures. This is particularly important for allocative efficiency, where the structure of a particular health system will determine what potential there is at each level of analysis for the provider to decide what mix of outputs should be produced.

Conclusions

There is increasing awareness that the design of the health system has a fundamental impact on the health and broader welfare of the population. Improved system efficiency is an important consideration because it enhances the capacity to produce valued outputs and the consequent sustainability of the system. However, the conceptualization of efficiency in health systems is far from straightforward. The first fundamental challenge is assigning 'value' to outputs (and also possibly inputs). Definitions of efficiency differ across institutions with no consistent reference to valued outputs and inputs. In practice, different definitions in use cover a range of valued outputs such as 'overall' performance, quality of care, health gain, or volume of treatment.

The challenge of identifying a set of valued outputs has implications for the conceptualization of both technical and allocative efficiency. To determine the technically efficient points of production, it is necessary to identify the outputs of the production process. Similarly, in order to determine what bundle of health services to provide, and thus identify the 'allocatively efficient' point of production, it is necessary to understand the preferences of the population being served. This will require consideration of whose preferences to consider, whether they should reflect the utilities associated with consumption of health care or the capabilities created by the consumption of health care, and the aggregation of preferences across society.

The second conceptual challenge refers to the difficulties associated with defining the boundaries of the entities under scrutiny. Determining the boundaries of a health system is one of the key areas of debate in health services research. The central point of discussion arises from the recognition that health outcomes are the result of numerous determinants, many of which might lie outside the direct influence of health policy makers. How narrowly or broadly the boundaries are set will influence judgments about the causal responsibility for improving health, thus influencing assessments of the level of 'efficiency' of the defined health system. A clear definition of relevant boundaries will facilitate the specification of objectives and the valued outputs and inputs for different areas of the health system.

The third challenge in conceptualization of efficiency for health systems relates to the intertemporal nature of the health system. All health systems are dynamic entities; performance in one period will influence performance in later periods. Health outcomes are the result not only of factors in the time period being measured, but also a product of behavior over the lifecycle as well as previous efforts of the health system. The physical resources such as hospitals and medication available in a current period are a result of investments made in previous years, and will in part contribute to future attainment. Any definition and metric of efficiency should in principle attempt to capture the dynamic processes that make up the health system.

In conclusion, the above discussion summarizes a burgeoning literature and policy debate on the theory of health systems efficiency. The challenges at the system level identified amount to an extensive research agenda, the purpose of which should be to create a clearer understanding of the health system and how it can be used to the best effect in line with societal objectives. Although these challenges may appear daunting, considerable progress has been made in addressing many of these issues at the micro level, through the literature of CEA. Research in such analyses has secured major progress both in the theory and use of economic thinking, and has had a fundamental impact on health policy. It is to be hoped that similar progress can now be made at the other levels of analysis.

See also: Cost–Value Analysis. Efficiency in Health Care, Concepts of. Evaluating Efficiency of a Health Care System in the Developed World. Quality-Adjusted Life-Years. Resource Allocation Funding Formulae, Efficiency of. Welfarism and Extra-Welfarism

References

Aday, L. A., Begley, C. E., Lairson, D. R. and Balkrishnan, R. (2004). *Evaluating the healthcare system: Effectiveness, efficiency, and equity.* Chicago, IL: Health Administration Press.

AQA Alliance (2006). Principles of 'efficiency' measures. Available at: http://www.aqaalliance.org/files/PrinciplesofEfficiencyMeasurement.pdf (accessed 23.06.11).

Chung, J., Kaleba, E., and Wozinak, G. (2008). A framework for measuring healthcare efficiency and value. *Working Paper Prepared for the Physician Consortium for Performance Improvement. Work Group on Efficiency and Cost of Care*. Available at: http://www.ama-assn.org/ama1/pub/upload/mm/370/empirical_applications.pdf (accessed 21.04.11).

Commonwealth Fund (2006). *Framework for a high performance health system for the United States*. New York: The Commonwealth Fund.

Cylus, J. and Smith, P. C. (2013). Comparative measures of health system efficiency. In Papanicolas, I. and Smith, PC. (eds.) *Health system performance comparison: An agenda for policy, information and research*. Maidenhead: Open University Press.

Donabedian, A. (1990). The seven pillars of quality. *Archives of Pathology and Laboratory Medicine* **114**, 1115–1118.

Hurst, J. and Jee-Hughes, M. (2001). Performance measurement and performance management in OECD health systems. *OECD Labour Market and Social Policy Occasional Papers, No. 47*. Paris: OECD Publishing.

Institute of Medicine (IOM) (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Maynard, A. (2007). Health Economics in the past, present and future. In Suvanto, A. and Vartiainen, H. (eds.) *Finance and incentives of the health care system*. Available at: http://www.vatt.fi/file/vatt_publication_pdf/j45.pdf#page=21 (accessed 05.03.13).

National Quality Forum (2007). Measurement framework: Evaluating efficiency across episodes of care. Available at: http://www.qualityforum.org/Projects/Episodes_of_Care_Framework.aspx (accessed 23.06.11).

Pacific Business Group on Health (2006). Hospital cost efficiency measurement: Methodological approaches. Available at: http://www.pbgh.org/storage/documents/reports/PBGHHospEfficiencyMeas_01-2006_22p.pdf (accessed 23.06.11).

Papanicolas, I. (2013). Frameworks for International comparisons. In Papanicolas, I. and Smith, P. C. (eds.) *Performance comparisons for health system improvement*, pp. 31–75. Maidenhead: Open University Press.

RAND (2008). Identifying, categorizing and evaluating health care efficiency measures. *AHRQ Publication No. 08-0030*. Santa Monica, CA: RAND. Available at: http://www.ahrq.gov/research/findings/final-reports/efficiency/efficiency.pdf (accessed 23.06.11).

Roberts, M. J., Hsiao, W., Berman, P. and Reich, M. R. (2008). *Getting health reform right: A guide to improving performance and equity*. Oxford: Oxford University Press.

Williams, A. (1987). Health economics: The cheerful face of a dismal science. In Williams, A. (ed.) *Health and economics*, pp. 1–11. London: Macmillan.

World Health Organization (WHO) (2000). *The world health report 2000: Health systems: Improving performance*. Geneva: WHO Publications.

World Health Organization (WHO) (2007). *Everybody's business: Strengthening health systems to improve health outcomes. WHO's framework for action*. Geneva: WHO Document Production Services.

Further Reading

Allin, S., Hernandez-Quevedo, C. and Masseria, C. (2009). Measuring equity of access to health care. In Smith, P. C., Mossialos, E., Papanicolas, I. and Leatherman, S. (eds.) *Performance measurement for health system improvement: Experiences, challenges and prospects*, pp. 187–221. Cambridge: Cambridge University Press.

Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New York: John Wiley & Sons.

Brouwer, W. F. B., Culyer, A. J., van Exel, N. J. A. and Rutten, F. F. H. (2008). Welfarism vs. extra-welfarism. *Journal of Health Economics* **27**, 325–338.

Culyer, A. J. (1990). Commodities, characteristics of commodities, characteristics of people and the quality of life. In Baldwin, S., Godfrey, C. and Propper, C. (eds.) *The quality of life: Perspectives and policies*, pp. 9–27. London: Routledge.

Culyer, A. J. (1995). Need: The idea won't do – but we still need it. *Social Science and Medicine* **40**, 231–238.

Hernandez-Quevedo, C. and Papaicolas, I. (2013). Conceptualizing and comparing equity across nations. In Papanicolas, I. and Smith, P. C. (eds.) *Performance comparisons for health system improvement*. Maidenhead: Open University Press.

Jacobs, R., Street, A. and Smith, P. C. (2006). *Measuring efficiency in health care*. Cambridge: Cambridge University Press.

Nolte, E., Bain, C. and Mckee, M. (2009). Population health. In Smith, P. C., Mossialos, E., Papanicolas, I. and Leatherman, S. (eds.) *Performance measurement for health system improvement: Experiences, challenges and prospects*, pp. 27–6272. Cambridge: Cambridge University Press.

Office for National Statistics (2011). *Public service output, inputs and productivity: Healthcare*. Newport: ONS. Available at: http://www.ons.gov.uk/ons/rel/psa/public-service-productivity/healthcare-2011/index.html (accessed 05.03.13).

Street, A. and Hakkinen, U. (2010). Health system efficiency and productivity. In Smith, PC., Mossialos, E., Papanicolas, I. and Leatherman, S. (eds.) *Performance measurement for health system improvement: Experiences, challenges and prospects*. Cambridge: Cambridge University Press.

Tandon, A., Lauer, J. A., Evans, D. B. and Murray, C. J. L. (2003). Health system efficeincy: Concepts. In Murray, C. J. L. and Evans, D. B. (eds.) *Health systems performance assessment: Debates, methods and empiricism*. Geneva: World Health Organization.

Time Preference and Discounting

M Paulden, University of Toronto, Toronto, ON, Canada

Introduction

Decision making is a fact of life, both for individuals and for policy makers acting on behalf of society. All decisions result in costs and benefits. The purpose of economic evaluation is to compare these and determine whether the costs of a policy are justified by the benefits.

Economic evaluations comprise two steps. First, a specific intervention is identified (e.g., a vaccination program), along with any relevant alternative uses of the same resources (e.g., other public health activities). The most desirable of these alternatives is referred to as the opportunity cost, since it represents the opportunity forgone by implementing the intervention. Second, the intervention and its opportunity cost are each assigned a value. The intervention is considered desirable only if it has greater value than its opportunity cost.

The opportunity cost of an intervention is usually dependent upon the point in time at which the costs of the intervention are incurred. For example, a rational decision maker would prefer to incur a cost of $100 next year rather than a cost of $100 this year, if this allows the $100 to be invested for an year and yield a positive return.

Meanwhile, the value assigned to the benefits of the intervention and its opportunity cost generally depends upon the point in time at which these benefits will be experienced. For example, the benefits from a vaccination program due to be experienced 10 years from now may be assigned less value by the policy maker than the benefits realized from an alternative health intervention this year, even if these benefits are otherwise identical. This is known as time preference.

To allow for comparability between costs and benefits experienced across different points in time, it is conventional to represent all future costs and benefits in terms of their present value (i.e., their value today), regardless of when they are actually experienced. Typically, the present value of future costs and benefits is less than their value at the time they are experienced. Calculating present values therefore requires discounting of future costs and benefits. Although in everyday language the term 'discounting' is used in a variety of contexts, in economics it is used almost exclusively to describe the steps taken to calculate the present value of future costs and benefits.

Discounting is conducted as follows. First the costs and benefits are disaggregated into time periods (usually years). Then the costs and benefits in each time period are assigned weights, with lower weights assigned to more distant time periods. These weights are usually given by

$$1/(1+d)^{t-1} \quad [1]$$

where t represents the time period and d represents the discount rate. Costs and benefits incurred in the current time period ($t=1$) are not discounted. As a result of compounding, the discount applied to costs and benefits in the distant future can be substantial. For example, since 2003 the UK Treasury has recommended that most future costs and benefits be discounted at an annual rate of 3.5%. This means that a cost of £1000 incurred 30 years from now has a present value of £356 $[1000/(1+0.035)^{29}]$. Before 2003, the UK Treasury recommended a higher discount rate of 6%, which would have resulted in a present value of just £174 $[1000/(1+0.06)^{29}]$. The effect of this compounding is demonstrated in **Table 1**.

In some cases there are further complications to consider. First, costs and benefits might not be discounted at the same rate. For example, the Dutch Health Care Insurance Board recommends discounting costs at 4% per year and benefits at 1.5% per year. This practice is known as differential discounting, and has recently been the subject of considerable debate. Second, the discount rate itself might be conditional upon the point in time at which costs and benefits occur. For example, the UK Treasury recommends that a lower discount rate be applied to costs and benefits more than 30 years in the future, with this rate falling progressively from 3.5% at 30 years to 1.0% beyond 300 years. This is known as non-constant discounting, of which hyperbolic discounting is a common form.

Discounting can tip the balance between an intervention appearing desirable or not. It can also impact upon the relative desirability of interventions. Interventions with upfront costs but long term benefits (including many public health activities) generally appear less desirable following discounting, whereas interventions with long-term costs sometimes appear more desirable.

Given the complexities of discounting and its potential importance, it is the subject of considerable controversy. In the following article, the rationales for discounting are explained in greater detail. Some of the more contentious issues in discounting, such as the merits of differential or non-constant

Table 1 The present value of £1000 incurred at different times using alternative discount rates

Discount rate (per annum)	*Time from the present*						
	1 year	*2 years*	*3 years*	*5 years*	*10 years*	*30 years*	*100 years*
1.5%	£985	£971	£956	£928	£862	£640	£226
3.5%	£966	£934	£902	£842	£709	£356	£32
6%	£943	£890	£840	£747	£558	£174	£3

 doi:10.1016/B978-0-12-375678-7.00506-X

discounting, are discussed. Recent papers that have attempted to determine the appropriate discount rates for health policy making are then reviewed. The article ends with some suggested further reading.

Conventional Approaches to Discounting

Discounting is founded on considerations of opportunity cost and time preference. A common way of considering the opportunity cost of an intervention is to estimate the 'marginal social opportunity cost of capital,' which represents the amount of private investment foregone by adopting the intervention. Meanwhile, time preference is often considered by estimating the 'social rate of time preference,' which represents society's time preference for consumption.

Opportunity Cost

In cases where the resources used for a publicly funded intervention can otherwise be used in the private sector, the opportunity cost of the intervention is the value of the best possible forgone private investment opportunity. In this case, the intervention should only be funded if its benefits exceed those of the forgone private investment. In technical terms, this requires that the rate of return on the intervention exceeds the marginal pretax rate of return on the forgone private investment, otherwise known as the marginal social opportunity cost of capital.

Many economists have proposed methods for estimating the marginal social opportunity cost of capital. Under strict assumptions, it is equivalent to the marginal pretax rate of return on riskless private investments. This is sometimes approximated by the average real pretax rate of return on top-rated corporate bonds. However, this may be an over-estimate if the average rate of return is higher than the marginal rate of return, or if market distortions affect the rate of return.

Time Preference

It has been widely observed that individuals exhibit time preference: they prefer to receive benefits sooner rather than later. Economists have three standard explanations for individual time preference, each of which may be extended to explain time preference at the level of society.

Individual time preference

The first explanation for individual time preference is that individuals may expect their incomes to increase over time, allowing them to consume more in the future than they do today. However, the extra utility that individuals receive from any additional consumption tends to decline as consumption increases. As a result, individuals may prefer to consume more today, at the expense of future consumption, in order to smooth their lifetime consumption and increase their lifetime utility.

Second, every individual faces some risk of death or some other catastrophe that would prevent them from consuming in the future. Offered the choice between consumption today and identical consumption in 30 years' time, many individuals would prefer to consume today on the grounds that they are not guaranteed to be alive for 30 years' time.

Third, individuals might prefer to consume sooner rather than later regardless of their expectations of future consumption. This is referred to as pure time preference, and reflects the fact that individuals often exhibit myopic or impatient preferences. Individuals might not appreciate that they are forfeiting future consumption by consuming more sooner, or they might regard their future utility as being less important than their utility today.

Societal time preference

Extending the first explanation for individual time preference to the societal level is relatively uncontroversial. As the incomes of individuals increase over time, the aggregate consumption of society also increases. Policy makers, acting on behalf of society, may prefer to enact policies which not only smooth the consumption of individuals over their lifetimes but also smooth the aggregate consumption of society over time, so as to maximize aggregate utility (i.e., social welfare) across generations.

Extending the second explanation to the societal level is more problematic. Although individuals face a non-negligible risk of an event, such as death, which prevents them from consuming in future years, the risk of catastrophe faced by society is much smaller. Although every year some members of society will die (or emigrate), others will be born (or immigrate) and society can be expected to carry on regardless. Only a truly catastrophic event would prevent society from consuming in future years. Consequently, society has much less justification than individuals for preferring consumption sooner rather than later.

The final explanation for individual time preference – that individuals often have myopic or impatient preferences – is particularly controversial in a societal context. Although there is considerable empirical evidence that individuals exhibit these preferences, many economists, philosophers, and other thinkers have argued against considering these pure time preferences in societal decision making. This is discussed further in the next section.

The social rate of time preference represents the rate at which society is willing to postpone current consumption in exchange for future consumption.

Under strict assumptions this may be approximated by the after-tax rate of return on risk-free securities (e.g., government bonds). However, these assumptions require that individuals express all their preferences within the market. In addition, the approximation presupposes that individuals do not change preferences when faced with decisions that affect society rather than just themselves. Where these assumptions do not hold, the social rate of time preference is likely to be lower than that implied by market rates.

An alternative means of estimating the social rate of time preference is to use a formula attributed to British mathematician Frank Ramsey. According to the Ramsey formula, the social rate of time preference is given by

$$\mu g + L + \delta \qquad [2]$$

The formula assigns a separate rate to each of the three standard explanations for time preference given above – the

diminishing marginal utility of consumption (μg), the risk of a catastrophic event (L), and pure time preference (δ) – and sums these to give the social rate of time preference. The rate assigned to the diminishing marginal utility of consumption is calculated by multiplying an estimate of the elasticity of the marginal utility of consumption (μ) by the growth rate of real per capita consumption (g). The UK Treasury used this methodology in 2003 to derive its current 3.5% discount rate: citing various sources, it estimated that $\mu = 1$, $g = 2\%$, $L = 1\%$, and $\delta = 0.5\%$, implying a social rate of time preference of 3.5% per annum.

Deriving a Social Discount Rate

Conventionally, attempts to derive a social discount rate – used to discount future costs and benefits in economic evaluations of public interventions – have focused on reconciling the marginal social opportunity cost of capital with the social rate of time preference. Under very unrealistic assumptions, including complete and undistorted markets, the marginal rate at which present consumption opportunities can be transformed into future consumption opportunities is equal to the marginal rate at which society would choose to substitute present for future consumption. Under such a scenario the marginal social opportunity cost of capital and the social rate of time preference are equivalent. Where these assumptions do not hold, some economists have advocated using one or the other as the social discount rate, whereas others have proposed ways of reconciling the two. These include the 'weighted average' and the 'shadow price of capital' approaches.

The weighted average approach holds that the social discount rate should be a weighted average of the marginal social opportunity cost of capital, the social rate of time preference, and, in the case of an open economy, the cost of borrowing on international markets. These weights should reflect the proportion of funds obtained from each source, implying a different social discount rate for each intervention. A limitation of the weighted average approach is that the benefits of the intervention are assumed to be consumed immediately. If these benefits are instead reinvested in the private sector, the weighted average approach will overestimate the social discount rate.

The shadow price of capital approach addresses a key limitation of the weighted average approach by recognizing that the benefits of an intervention may be reinvested in the private sector. The benefit of an intervention is given by the sum of the consumption resulting from the intervention and any future consumption generated from reinvestment of the benefit. The cost is given by the sum of the consumption directly displaced by the intervention and any future consumption forgone because of the displacement of private investment. Although this approach is theoretically attractive, it is more difficult to implement than the weighted average approach.

Discounting in the Context of Health Policy Making

Conventional approaches to deriving a social discount rate can be problematic in the context of health policy making. Health policy makers are often faced with a constrained budget. In this context, the opportunity cost of adopting a specific health intervention is typically not forgone private investment, but rather displaced health care activities elsewhere within the health care system. Furthermore, health policy makers are often concerned specifically with society's health, rather than society's consumption. Since society's time preferences for health typically differ from those for consumption, health policy makers may therefore need to instead consider the 'social rate of time preference for health.' Finally, health policy makers may have reason to adopt different discount rates for costs and health benefits, rather than a single 'social discount rate' for both (this is returned to in the final section). For clarity, the 'social rate of time preference' will hereafter be referred to as the 'social rate of time preference for consumption,' to differentiate it from the 'social rate of time preference for health.'

The social rate of time preference for health

The standard explanations for society's time preference for consumption also apply to society's time preference for health. As society's health improves over time, it may have a preference for earlier health benefits over later health benefits, because of the diminishing marginal utility of health. Society may also prefer earlier health benefits because of catastrophe risk or pure time preference.

The social rate of time preference for health generally differs from the social rate of time preference for consumption. One reason is that the relative value of health and consumption might change over time. Dave Smith and Hugh Gravelle have suggested that the consumption value of health might grow over time, since it is positively correlated with increasing incomes.

The social rate of time preference for health may be estimated using the Ramsey formula. It may also be implicitly revealed by the allocation of health budgets across time (this is returned to in the final section).

Implications for discounting

Where a health policy maker has a specific concern for society's health, and is faced with a fixed budget constraint, it follows that the appropriate discount rate(s) to adopt for economic evaluations of health interventions cannot be derived from either the marginal social opportunity cost of capital or the social rate of time preference for consumption, but rather by considering the social rate of time preference for health and the specific opportunity cost of adopting the health intervention in question (i.e., the health forgone elsewhere as a result of displaced health care activities). The final section of this article reviews recent work demonstrating how the discount rate(s) should be derived in this context.

Contentious Issues in Discounting

Should Benefits be Discounted at All?

Although there is substantial empirical evidence that individuals prefer earlier benefits to later benefits, there is widespread controversy over whether society should display similar

time preferences. Some authors have expressed frustration that discounting health benefits causes many interventions (particularly public health activities) to appear much less desirable. Others have raised ethical objections on the grounds that discounting benefits discriminates against future generations.

A popular view among contemporary economists is that social welfare should be determined by aggregating the preferences of individuals, specifically those individuals who are members of the currently living generation. It follows that if these individuals prefer earlier benefits to later benefits then society should too. An exception is sometimes made for those aspects of individual time preference resulting from myopia or impatience. Many economists regard economic evaluation as a means of bringing greater rationality into societal decision making, and so oppose the consideration of these aspects of time preference on the basis that they are 'irrational.' This view is not universally shared by economists. For example, the UK Treasury explicitly considered such preferences in the derivation of its 3.5% discount rate.

This focus on the preferences of individuals is a relatively new concept. As Murray Krahn and Amiram Gafni have noted, earlier thinkers, including Jeremy Bentham, David Hume and the early utilitarians, had an objective, interpersonal and intergenerational view of social welfare, in which the subjective preferences of the current generation were given relatively little weight. The time preferences of individuals were viewed as a failing of human reason, as representing intellectual or moral weakness, and potentially harmful to social welfare. According to Arthur Cecil Pigou, the government therefore has a 'duty' to "protect the interests of the future in some degree against the effects of our irrational discounting and of our preference for ourselves over our descendants." More recently, John Rawls argued that the principle of intergenerational justice should guide social decision making, in which the interests of all generations are given equal consideration.

However, there are legitimate reasons for individuals and society to prefer earlier benefits, even if equal regard is given to the welfare of future generations. First, consumption or health may be expected to increase over time. If social welfare is regarded as an aggregation of individual utilities, and if there is diminishing marginal utility to consumption or health, then an equal concern for the welfare of all generations may require that preference be given to improving the consumption or health of earlier generations. Alternatively, if intergenerational justice requires that consumption or health be equalized across generations, then an expectation that consumption or health will increase over time implies that preference should be given to improving the consumption or health of earlier generations. Finally, there is always some risk, however small, of a catastrophe preventing society from enjoying the benefits of consumption or health in the future.

It follows that society's rate of time preference for either consumption or health is most likely positive but lower than that for individuals. For these reasons, future benefits generally should be discounted, regardless of whether society accounts for myopic or impatient preferences in societal decision making.

Should Costs and Health Benefits be Discounted at Different Rates?

Although not a new controversy, this debate was reignited in 2004 by the decision of the UK's National Institute for Health and Clinical Excellence (NICE) to no longer recommend the differential discounting of incremental costs and health benefits (at rates of 6% and 1.5% respectively) but instead recommend that both be discounted at a common rate of 3.5%. In a 2011 paper, Karl Claxton and colleagues brought together the prominent authors from both sides of this debate and clarified the causes of this disagreement.

The authors identified a number of matters of context which must be considered before this question can be answered, including the health policy maker's perspective on social choice, the specific objective adopted by the policy maker, and whether the policy maker faces a fixed budget constraint.

Where the policy maker faces a fixed budget constraint, differential discounting is justified only if the cost–effectiveness threshold is expected to change over time. Alternatively, if the policy maker does not face a budget constraint, and if the policy maker adopts a welfarist or extra-welfarist perspective on social choice, then differential discounting is only justified if the consumption value of health is expected to change over time.

These issues are described in more detail in the next section.

Is Differential Discounting Logically Inconsistent?

A number of arguments have been made that differential discounting is logically inconsistent, and so common discounting is unavoidable. These include Emmett Keeler and Shan Cretin's paradox of indefinite delay, Milton Weinstein and William Stason's chain of logic argument, and William Kip Viscusi's equivalence argument. Karl Claxton and colleagues also criticized the 'illogicality' of differential discounting in a previous paper in the recent debate. In response, Erik Nord has recently argued that all of these 'consistency arguments' are themselves logically inconsistent.

The paradox of indefinite delay

According to Keeler and Cretin's paradox of indefinite delay, if two alternative interventions, X and Y, are identical in every respect, except that X is implemented today and Y is implemented in 10 years' time (Table 2), then discounting benefits at a lower rate than costs will result in Y having a more favorable cost–benefit ratio than X. Unless benefits are discounted at a rate at least as high as costs, this implies that policy makers will always prefer to indefinitely delay every intervention.

A problem with this argument, as noted by Michael Parsonage and Henry Neuburger, is that the policy relevant

Table 2 Keeler and Cretin's paradox of indefinite delay

Intervention	*Current year*	*10 Years' time*
X	$10k, 1 life years	
Y		$10k, 1 life years

question is not where in time to locate an intervention, but rather how to set priorities within a constrained budget in any given year. There is no reason to discount Y to today's present value at all: it can simply be appraised in 10 years' time, when it will have the same cost–benefit ratio as X. Nord argues that a distinction should also be made between start time difference and benefit time difference: policy makers may not wish to indefinitely postpone the start time of an intervention, but they may still have a preference over the timing of benefits in programs with given start times.

The chain of logic argument

Weinstein and Stason's chain of logic argument, cited by the Washington Panel as the 'consistency argument,' runs as follows. Suppose there are two interventions: A costs $10 000 now and saves 1 life year in 40 years' time, while B costs $70 000 in 40 years' time and also saves 1 life year in 40 years' time. Assuming a discount rate on costs of 5%, A and B are equivalent. A third intervention, C, is then considered which costs $70 000 now and saves 1 life year now (Table 3). Assuming a constant value of a life year, C is equivalent to B and hence equivalent to A. Since C costs seven times as much as A, the benefits of C must also be seven times the value of those of A, implying a discount rate on benefits of 5%. Costs and benefits must therefore be discounted at the same rate.

However, Nord argues that this is true only if the value of a life year is constant, which was presupposed in the argument. Indeed, Weinstein and Stason acknowledged that adopting a non-constant value of a life year may justify differential discounting of costs and benefits.

The equivalence argument

Viscusi's equivalence argument considers an intervention that costs $8 million now and saves two lives in 10 years. The value of a life in year 10 is V. Costs are discounted at r and benefits at d. In present value terms, the intervention is worthwhile if $2V/(1+d)^{10} > 8$. Viscusi suggested that one could instead look at 'terminal values,' with the intervention worthwhile if $2V > 8(1+r)^{10}$, which can be rearranged to $2V/(1+r)^{10} > 8$. Since these equations are equivalent only if $d = r$, it follows that costs and benefits should be discounted at the same rate.

However, as Nord notes, Viscusi made only the trivial arithmetic point that if one uses the same discount rate for both costs and benefits then either present values or terminal values may be used. Viscusi presupposed $d = r$ in his argument. The real issue about whether the discount rates should be the same is not addressed.

Further arguments

An earlier paper by Claxton and colleagues made two further arguments against the 'illogicality' of differential discounting: first, that support for differential discounting "must rest on a claim that health, unlike wealth, is not tradable over time"; second, that "the true cost of health gained is health forgone – at whatever date these gains or losses may occur. Put in this fashion, the illogicality of wanting to discount health forgone at a different rate from health gained becomes plain."

The first argument is disputed by Nord, who notes that differential discounting can be justified even if health is tradable over time. The second argument is correct in stating that, faced with a fixed health budget constraint, "the true cost of health gained is health forgone," but does not account for the possibility that the cost–effectiveness threshold might change over time. This possibility was considered in the more recent paper by Claxton and colleagues.

Table 3 Weinstein and Stason's chain of logic argument

Intervention	*Current year*	*40 Years' time*
A	$10k	1 life years
B		$70k, 1 life years
C	$70k, 1 life years	

Is Non-Constant Discounting Appropriate?

Conventional discounting is consistent with Paul Samuelson's discounted utility model, with a key assumption being that the discount rate remains constant over time. For example, if the discount rate is 5% then a benefit today is equivalent to a 5% greater benefit in 1 year, whereas a benefit in 10 years is equivalent to a 5% greater benefit in 11 years. Adopting a constant discount rate results in time consistent decision making. This means that if an intervention with distant costs and benefits appears worthwhile today, then it will also appear worthwhile if reappraised in 10 years' time.

However, empirical studies have demonstrated that individuals rarely have a constant rate of time preference. Individuals often exhibit strong time preferences for benefits in the near future: offered a choice between $10 now or $15 next year, many will prefer $10 now. But this time preference becomes weaker in the distant future: offered a choice between $10 in 20 years' time or $15 in 21 years' time, many of these same individuals will prefer $15 in 21 years. A possible reason is that individuals have difficulty comprehending differences between distant time periods. The result is time inconsistent decision making: although the option of $15 in 21 years appears more attractive today, if asked to reappraise their decision 20 years from now many individuals will regret their decision and, if possible, switch their choice.

The issue of how to deal with time inconsistent decision making remains unresolved. Perhaps as a result, non-constant discounting (including hyperbolic discounting) is rarely adopted in practice. Although constant discounting allows policy makers to avoid this issue, the tradeoff is that society's time preferences cannot be fully reflected. An exception is the UK Treasury, which recommends non-constant discounting for very distant costs and benefits. The discount rate falls progressively from 3.0% (between 30 and 75 years), to 2.5% (76–125 years), to 2.0% (126–200 years), to 1.5% (201–300 years), to 1.0% (beyond 300 years). Even with this declining rate, costs and benefits in 300 years are given just 0.14% of the weight of present costs and benefits (under a constant discount rate of 3.5%, this weight would be 0.003%).

Discounting in the Context of Health Policy Making

Over recent years, health policy makers around the world have made increasing use of cost–effectiveness analysis (CEA) to

guide their decision making around the adoption of new health interventions. Typically a CEA compares the costs and health outcomes associated with a health intervention to each of its comparators, and a judgment made as to whether its adoption would be cost-effective. Policy makers, or the agencies which conduct CEAs on their behalf, generally publish guidance as to the discount rates that should be used. For example, NICE currently specifies that costs and health benefits should both be discounted at 3.5% per year, whereas the Canadian Agency for Drugs and Technology in Health (CADTH) recommends that both be discounted at 5% per year. Such guidance has the advantage of providing consistency and comparability across the variety of CEAs considered by each policy maker. It has also resulted in considerable debate as to the most appropriate discount rates to use, with much of this debate focused on the merits of differential discounting of costs and health benefits.

Recent contributions to this debate have demonstrated that the appropriate discount rates to use depend on the context in which health policy is made. This requires consideration of a number of issues, including the health policy maker's perspective on social choice, the specific objective adopted by the policy maker, and the existence or otherwise of a fixed budget constraint for health.

The Perspective on Social Choice

In defining the policy context, the first consideration is the perspective on social choice adopted by the health policy maker. This is the subject of a vast literature, and there are many possible perspectives that policy makers may reasonably adopt. These can be usefully characterized into two groups: those that regard the primary purpose of policy making to be to improve social welfare, as defined by welfarist or extra-welfarist economics; and those that regard policy making as a means for satisfying specific and explicit objectives, rather than improving social welfare more generally.

A welfarist perspective

Traditional welfarist economics assumes that individuals rationally maximize their utility by ordering the various options available to them and acting according to their preferences. Individuals are regarded as the only judges of what contributes most to their utility. Social welfare is judged to be nothing more than an aggregation of these individual utilities. This notion of social welfare is very restrictive: in particular, it cannot take account of outcomes other than utilities, and it does not permit the use of sources of valuation other than the individuals affected by the policy decision.

An extra-welfarist perspective

Over recent decades, these limitations have resulted in the rise of extra-welfarist economics, in which non-utility information such as the quality of individuals' utilities, equity weights, and individuals' characteristics and capabilities are considered alongside individual utilities. This provides substantially more flexibility in the definition of social welfare. Extra-welfarist economics otherwise retains many of the features of welfarist economics: the purpose of policy making is still to improve social welfare, and individual preferences remain an important consideration.

Problems with the definition of social welfare

To judge whether policy decisions improve social welfare requires the expression of an explicit and complete social welfare function: a ranking over all conceivable social states. However, there are many reasons why the expression of an explicit and complete social welfare function might not be possible or even desirable. The work of Nobel Laureates Kenneth Arrow and Amartya Sen demonstrated that it is impossible to specify an explicit and complete social welfare function that satisfies basic requirements while remaining non-dictatorial and respecting minimal liberty. It is therefore unlikely that any explicit and complete social welfare function could be expressed which would carry social legitimacy. Furthermore, policy makers may express no desire in specifying an explicit social welfare function in any case. This is problematic if improving social welfare is regarded as the primary purpose of policy making.

A social decision-making perspective

In response, many economists have advocated for an alternative approach. The social decision-making perspective identifies a more modest role for policy making: satisfying specific and explicit objectives rather than improving social welfare more generally. Under this perspective, policy making agencies (such as NICE) are seen as agents of a socially legitimate higher authority (in NICE's case the UK's democratically elected parliament). This higher authority does not specify an explicit social welfare function, but nevertheless allocates resources among different sectors (e.g., health, education, etc.) and grants each agent the responsibility to pursue a specific and explicit objective subject to a budget constraint. In NICE's case, this objective may be to improve society's health, subject to the budget for health allocated by the UK parliament. Although the higher authority does not specify an explicit social welfare function, the objectives it delegates to the agents, and its allocation of resources between sectors and within sectors across time, represent a partial expression of some unknown latent social welfare function.

The Objective of the Policy Maker

The second consideration of context is the health policy maker's objective. This is influenced by the perspective on social choice. A recent paper by Karl Claxton and colleagues considers two possible objectives that a health policy maker might reasonably adopt: the first under a social decision-making perspective, the second under a welfarist or extra-welfarist perspective.

A social decision-making perspective

Under a social decision-making perspective, the health policy maker may reasonably seek to improve society's health, subject to the budget constraint set by the higher authority. Society may also have a preference for earlier health benefits, represented by the social rate of time preference for health.

The health policy maker's objective may therefore be to 'maximize the present value of health.'

A welfarist or extra-welfarist perspective

Under a welfarist or extra-welfarist perspective, the health policy maker instead seeks to improve social welfare. Health may be considered in consumption terms by weighting it by the consumption value of health. As Hugh Gravelle and colleagues note, if consumption and health are the only arguments in the social welfare function, or are separable from other arguments, then maximizing the consumption value of health is equivalent to maximizing social welfare. Society's time preferences are represented by the social rate of time preference for consumption. The health policy maker's objective may therefore be to 'maximize the present consumption value of health.'

The Existence or Otherwise of a Budget Constraint

The third consideration of context is the existence or otherwise of a fixed budget constraint for health. This has important implications for the opportunity cost of adopting an intervention.

A constrained budget

Where the health budget is constrained, any additional costs of adopting an intervention fall within this budget. It is inevitable that one or more other health interventions will then be displaced, resulting in forgone health. This represents the opportunity cost of adopting the intervention. A critical part of appraising the cost–effectiveness of the intervention is estimating this opportunity cost. Unfortunately, health policy makers are usually unaware of the specific health interventions displaced, so the extent of forgone health must be estimated in some other way.

One such approach is to estimate the slope of the health production function. This function describes how changes in the health budget affect the aggregate health output of the health system, with health output usually considered to be a positive but diminishing function of the health budget. The reciprocal of the slope of the health production function at the prevailing health budget and health output represents the 'cost–effectiveness threshold,' denoted as k in Figure 1.

The cost–effectiveness threshold reveals how much health output is expected to be forgone following a marginal reduction in the existing health budget. For example, suppose that reducing the health budget by $50 000 reduces aggregate health output by 1 quality-adjusted life-year (QALY). The expected opportunity cost of adopting a new intervention would therefore be 1 QALY for every additional $50 000 spent, implying a cost–effectiveness threshold of $50 000 per QALY.

Since the cost–effectiveness threshold represents a matter of fact – how much health output is forgone, rather than the value of this health output – its estimation is an empirical matter. All else equal, the cost–effectiveness threshold will grow with increases in the health budget and fall with improvements in the marginal productivity of the health system. The possibility of the cost–effectiveness threshold changing over time must therefore be considered by health policy makers.

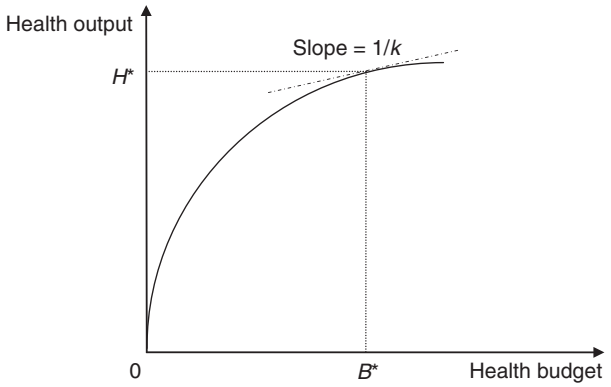


Figure 1 The 'cost–effectiveness threshold' (k) is determined by the reciprocal of the slope of the health production function at the prevailing health budget (B^*) and health output (H^*).

A non-constrained budget

Where the health budget is not constrained, any additional costs associated with adopting an intervention generally fall on other sectors or taxpayers. Under a welfarist or extra-welfarist perspective, the resulting opportunity cost may be regarded in terms of forgone consumption. To determine the cost–effectiveness of the intervention, the health benefits can be weighted by the consumption value of health so that a comparison may be made in consumption terms. Since adopting interventions in this context does not displace health, the cost–effectiveness threshold is redundant.

The Appropriate Discount Rates to Adopt

Recent work by Karl Claxton and colleagues demonstrated how these matters of context determine the appropriate discount rates for the health policy maker to adopt when appraising the cost–effectiveness of health interventions.

In cases where the health policy maker is faced with a constrained health budget, the authors assume that the policy maker determines whether an intervention is cost-effective by comparing its incremental cost–effectiveness ratio (ICER) to the current estimate of the cost–effectiveness threshold. Alternatively, where the health budget is not constrained, it is assumed that this ICER is compared to the current estimate of the consumption value of health.

According to Claxton and colleagues, where the health budget is constrained and the policy maker adopts a social decision-making perspective on social choice:

- Incremental costs should be discounted at approximately the social rate of time preference for health plus the expected growth rate of the cost–effectiveness threshold.
- Incremental health benefits should be discounted at the social rate of time preference for health.

Alternatively, where the health policy maker adopts a welfarist or extra-welfarist perspective:

- If the health budget is constrained, incremental costs should be discounted at approximately the social rate of time preference for consumption minus the expected

growth rate of the consumption value of health plus the expected growth rate of the cost–effectiveness threshold.

- If the health budget is not constrained, incremental costs should be discounted at the social rate of time preference for consumption.
- Regardless of whether or not the health budget is constrained, incremental health benefits should be discounted at approximately the social rate of time preference for consumption minus the expected growth rate of the consumption value of health.

In a subsequent paper, Mike Paulden and Karl Claxton provide an alternative specification for the appropriate discount rates to adopt under a social decision-making perspective. The authors argue that, in societies with a single-payer health system funded by a socially legitimate higher authority (such as a democratically elected parliament), the social rate of time preference for health is implicitly revealed by the allocation of health budgets across time. In this context, the social rate of time preference for health is shown to be approximately equal to the real interest rate faced by the higher authority which finances the health system minus the expected growth rate of the cost–effectiveness threshold.

Combining this result with the findings of Claxton and colleagues, it follows that:

- Incremental costs should be discounted at the real interest rate faced by the higher authority which finances the health system.
- Incremental health benefits should be discounted at approximately the real interest rate faced by the higher authority that finances the health system minus the expected growth rate of the cost–effectiveness threshold.

Intuition and policy implications

Where the budget is constrained, expected growth in the cost–effectiveness threshold must be accounted for, since the opportunity cost of adopting interventions also changes (for higher thresholds, incremental costs result in less health forgone, and vice versa). However, when the ICER of the intervention is compared to the current estimate of the cost–effectiveness threshold, this growth is not accounted for. The only practical way to account for this growth is to adjust the discount rate used for incremental costs. This results in differential discounting.

Under a welfarist or extra-welfarist perspective, if the consumption value of health is expected to change over time, an adjustment must be applied to the discount rate for incremental health benefits. If the health budget is not constrained, this results in differential discounting. However, if the health budget is constrained, the same adjustment must also be made to the discount rate used for incremental costs. This is because incremental costs fall on the health budget and result in forgone health, and any change in the consumption value of health also applies to health forgone. Under a constrained budget, change in the consumption value of health does not therefore justify differential discounting, but rather a lower discount rate for both incremental costs and health benefits. In this case, differential discounting is only appropriate if the cost–effectiveness threshold is expected to change over time.

Estimating the growth rate of the cost–effectiveness threshold requires extensive empirical research. With the exception of recent work in the UK, this research has not yet been undertaken in any jurisdiction. Theoretically, the cost–effectiveness threshold should grow with increases in the health budget but shrink with improvements in marginal productivity. As such, it may not be obvious in many jurisdictions whether the cost–effectiveness threshold is growing or shrinking. As Mike Paulden and Karl Claxton note, it may therefore be reasonable to assume that the growth rate of the cost–effectiveness threshold is zero (implying common discounting of incremental costs and health benefits) until a reliable empirical estimate of the growth rate of the cost–effectiveness threshold is available.

The real interest rate faced by a higher authority may be approximated by the real yield on its long term bonds. As of November 2012, the real yield on long term bonds issued by the UK and Canadian governments was in the region of 0.5–1.5% per annum. It follows that, under a social decision-making perspective, health policy making agencies such as NICE and CADTH should discount incremental costs and health benefits at a lower common rate than currently recommended.

See also: Adoption of New Technologies, Using Economic Evaluation. Analysing Heterogeneity to Support Decision Making. Budget-Impact Analysis. Cost-Effectiveness Modeling Using Health State Utility Values. Decision Analysis: Eliciting Experts' Beliefs to Characterize Uncertainties. Economic Evaluation of Public Health Interventions: Methodological Challenges. Economic Evaluation, Uncertainty in. Ethics and Social Value Judgments in Public Health. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Infectious Disease Modeling. Information Analysis, Value of. Observational Studies in Economic Evaluation. Policy Responses to Uncertainty in Healthcare Resource Allocation Decision Processes. Priority Setting in Public Health. Problem Structuring for Health Economic Model Development. Quality Assessment in Modeling in Decision Analytic Models for Economic Evaluation. Searching and Reviewing Nonclinical Evidence for Economic Evaluation. Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies. Statistical Issues in Economic Evaluations. Synthesizing Clinical Evidence for Economic Evaluation. Value of Information Methods to Prioritize Research. Valuing Informal Care for Economic Evaluation. What Is the Impact of Health on Economic Growth – and of Growth on Health?. Willingness to Pay for Health

Further Reading

Brouwer, W., Niessen, L., Postma, M. and Rutten, F. (2005). Need for differential discounting of costs and health effects in cost effectiveness analyses. *British Medical Journal* **331**(7514), 446–448.

Brouwer, W., Culyer, A., Van Exel, N. and Rutten, F. (2008). Welfarism vs. extra-welfarism. *Journal of Health Economics* **27**(2), 325–338.

Claxton, K., Paulden, M., Gravelle, H., Brouwer, W. and Culyer, A. (2011). Discounting and decision making in the economic evaluation of health-care technologies. *Health Economics* **20**(1), 2–15.

Claxton, K., Sculpher, M., Culyer, A., et al. (2006). Discounting and cost–effectiveness in NICE – Stepping back to sort out a confusion. *Health Economics* **15**(1), 1–4.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature* **40**(2), 351–401.

Gravelle, H., Brouwer, W., Niessen, L., Postma, M. and Rutten, F. (2007). Discounting in economic evaluations: Stepping forward towards optimal decision rules. *Health Economics* **16**(3), 307–317.

Krahn, M. and Gafni, A. (1993). Discounting in the evaluation of health care interventions. *Medical Care* **31**, 403–418.

Nord, E. (2011). Discounting future health benefits: The poverty of consistency arguments. *Health Economics* **20**(1), 16–26.

Paulden, M. and Claxton, K. (2012). Budget allocation and the revealed social rate of time preference for health. *Health Economics* **21**(5), 612–618.

Smith, D. and Gravelle, H. (2001). The practice of discounting in economic evaluations of healthcare interventions. *International Journal of Technology Assessment in Health Care* **17**(2), 236–243.

Treasury, H. M. (2003). *The Green Book: Appraisal and Evaluation in Central Government*. Available at: http://www.hm-treasury.gov.uk/data_greenbook_index.htm (accessed 09.11.12).

Zhuang J., Liang, Z. and Lin, T. (2007). *Theory and Practice in the Choice of Social Discount Rate for Cost–Benefit Analysis: A Survey*. Asian Development Bank, Economics Working Papers Series. Available at: http://www.adb.org/publications/theory-and-practice-choice-social-discount-rate-cost-benefit-analysis-survey/ (accessed 09.11.12).

Understanding Medical Tourism

G Gupte, Boston University, Boston, MA, USA
A Panjamapirom, The Advisory Board Company, Washington, DC, USA

Glossary

Competitive advantage An advantage that a firm has over its competitors.
Coronary artery bypass graft (CABG) A type of surgery that improves blood flow to the heart. Surgeons use CABG to treat people who have severe coronary heart disease.
Gestational surrogacy Surrogacy is an arrangement in which a woman carries and delivers a child for another couple or person and is biologically unrelated to the child.
Intracytoplasmic sperm injection One of the *in vitro* fertilization procedures (see the next term) in which a single sperm is injected directly into an egg.
***In vitro* fertilization (IVF)** Commonly referred to as IVF. IVF is the process of fertilization by manually combining an egg and sperm in a laboratory dish.
Medical tourism The travel for healthcare services outside the main local healthcare coverage area.
Organization for Economic Cooperation and Development (OECD) An international economic organization of 34 countries founded in 1961 to stimulate economic progress and world trade.
Synergies Where the whole becomes greater than the sum of the individual parts.

Growth of Medical Tourism

Globalization has inevitably become part of all industries, as we observe the economy stretch worldwide due to cheaper travel, better communication methods, and common solutions to problems. Over the past decade a healthcare practice that has emerged with an unimagined scale of magnitude as a multibillion dollar industry is the global medical tourism industry. Privatization and commercialization of conventional home country or region-based healthcare has been redefined in a global market. Medical tourism (also often referred to as international medical tourism) is generally defined as the travel of patients seeking healthcare services outside the main local healthcare coverage area.

According to sources, the number of medical tourists will rise from approximately 10.5 million in 2011 to 23.2 million by 2017. Some predict global revenue approximately between $40 and $60 billion, with different growth rates estimated over the next 10 years, some at 20% annual growth. Some reports estimate specific area growth, such as worldwide surgical volume at 60 000 patients every year. The most obvious benefit of medical tourism is cost savings for consumers, which typically range from 40% to 90%. For example, heart bypass that may cost approximately $180 000 in the US can be rendered at $10 000 in India or Thailand. Including first class air fare and four-star hotel accommodations for recovery, the savings are routinely more than 60%.

However, one needs to remember that medical tourism is not a new phenomenon, ancient civilizations and historians have recorded the travel of patients across regions and countries searching for appropriate treatment and better care. For example, Greeks traveled to spas known as 'Asklepia' in the Mediterranean for purification and spiritual healing; for over 2000 years foreign patients have traveled to the Aquae Sulis reservoir built by the Romans in what is now the British town of Bath; and Chinese and Indian scholars traveled across countries to seek more knowledge of diseases and conditions bringing patients with them. In the early nineteenth century, sanatoriums attracted tuberculosis patients to pristine mountain air such as Davos, Switzerland.

For the past few decades, outstanding facilities, such as Cleveland Clinic and Mayo Clinic, have attracted medical tourists to developed countries such as US, especially wealthy patrons from the developing world. However, worth noting is the new pattern of reversed care destination where patients from developed countries travel to seek medical services in developing countries such as India, Thailand, Malaysia, Chile, Argentina, Philippines, Jordan, South Africa, and others (Figure 1). These countries offer state-of-the-art technology and facilities, employ US-trained physicians, and concierge healthcare services at a fraction of costs that would have otherwise been incurred in the developed countries (low labor costs and overheads making it affordable). Other factors contributing to the growth of medical tourism include, but are not limited to, financial mobility, free trade, technological advances, cheap transportation, more resources, and rapid communications. Patients unable to gain prompt access to services due to a number of reasons, such as restricted insurance policies, long waiting time, and unavailable treatment options have started traveling beyond their borders to receive the care of preferences. Additionally, 'word of mouth' promotion by recent medical tourists, careful marketing, and interests from health insurance companies are some other drivers of medical tourism.

To attract patients from developed countries, medical tourism packages include procedures (Table 1) with preestablished prices, air fare, accommodation, ground transportation, concierge treatment, food, recuperation therapy, and supplementary trips to popular destinations. Often these packages are well coordinated by medical tourism companies representing care delivery organizations in the host countries.

 doi:10.1016/B978-0-12-375678-7.01007-5

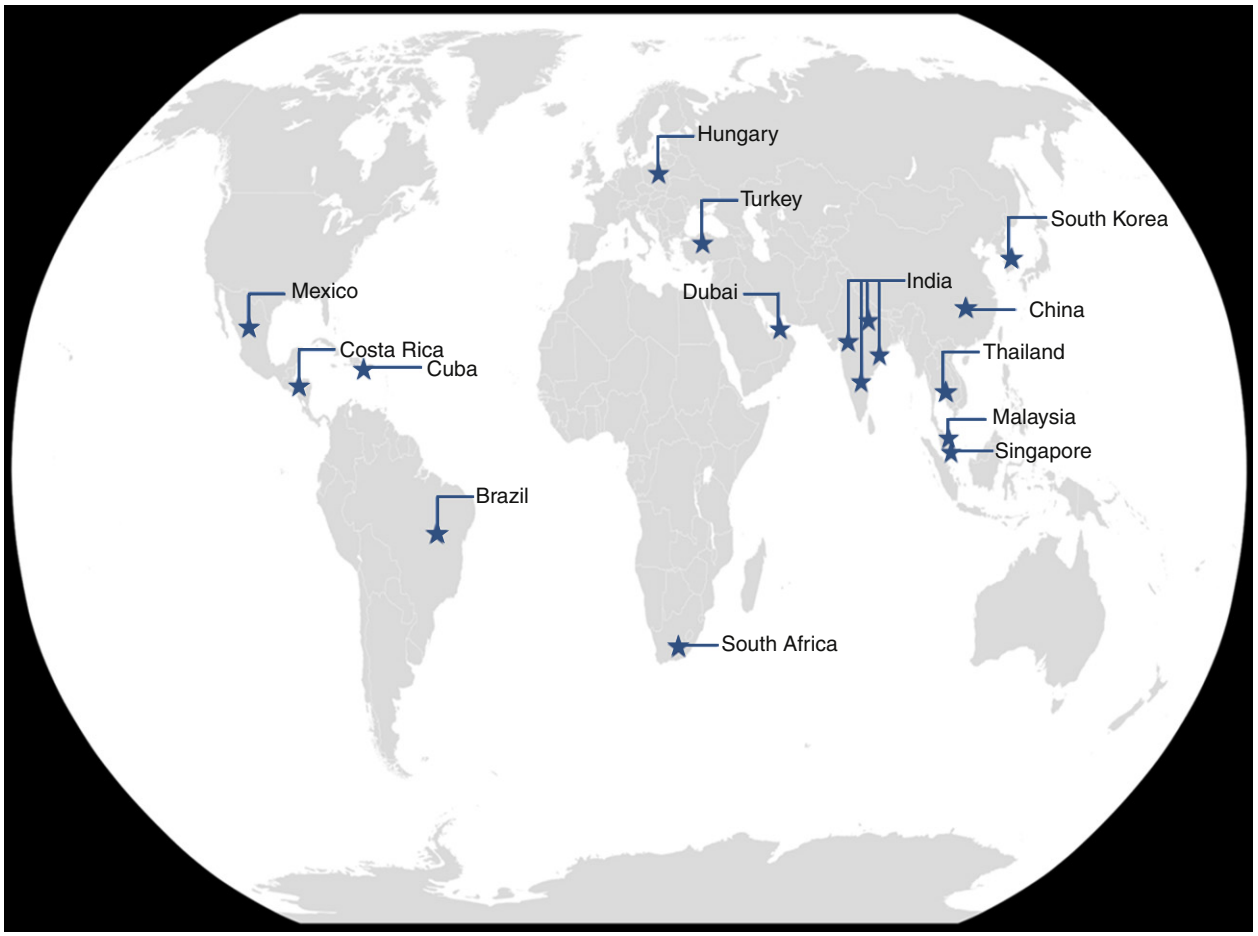


Figure 1 Medical tourism destinations.

Table 1 Commonly conducted procedures/treatments for medical tourism and their countries

Common medical tourism procedures	*Countries*
Dental tourism	Mexico, Bulgaria, Croatia, Argentina, Thailand, Hungary, and Poland
Transplant tourism (kidney, heart, lung, and liver transplant)	Thailand, India, and China
Reproduction tourism (includes *in vitro* fertilization, gestational surrogacy, and intracytoplasmic sperm injection)	India, Barbados, and UK
Cardiac procedures (coronary artery bypass graft and bypass surgery with heart valve replacement)	India, Thailand, Costa Rica, Singapore, Malaysia, and South Korea
Knee surgery	India, Thailand, Singapore, and Malaysia
Hip replacement	India, Thailand, Singapore, Malaysia, and Turkey
Ophthalmic surgery (cataract surgery, cornea alteration procedures, and glaucoma treatments)	Mexico, Bulgaria, Croatia, Argentina, India, Thailand, Singapore, Malaysia, and Turkey

Medical Tourism through a Systems Thinking Perspective

Systems thinking is an approach that reveals the underlying characteristics and relationships of systems. It uses comprehensive suite of tools and approaches to map, measure, and understand a system and its dynamics with the environment. Its utility in understanding and integrating complex, real-world settings makes it the right strategy for explaining the growth in medical tourism industry. To understand the medical tourism industry, it is imperative to understand the complex effects, synergies, and emergent behaviors of various stakeholders affecting this economic explosion. Additionally, researchers have discussed the benefits of using the systems thinking approach in understanding healthcare systems as these systems are self-organizing, constantly changing, tightly linked, governed by feedback, nonlinear, history dependent, counter-intuitive, and resistant to change.

The main stakeholders in the global healthcare marketplace include hospitals (private and public), patients, stand-alone clinics, governments, medical tourism companies, airlines, hotels, health administrators, and healthcare providers (Figure 2). However, first, it is important to understand how the medical tourism industry fits in the healthcare system. World Health Organization (WHO) defines a health system as, 'consists

of all organizations, people and actions whose primary intent is to promote restore or maintain health.' The goal is 'improving health and health equity in ways that are responsive, financially fair, and make the best, or most efficient, use of available resources.' WHO further identifies six systems building blocks that are used as a framework to apply systems thinking to medical tourism. These blocks serve as framework for understanding the medical tourism and effect of stakeholders in the service. The blocks have been identified as: Service delivery, health workforce, health information, medical technologies, health financing, and leadership and governance. These blocks form an integrated relationship that creates interactions in the medical tourism industry. Thus, using this framework and systems thinking approach an attempt is made to understand the world of medical tourism (Figure 3) (Box 1).

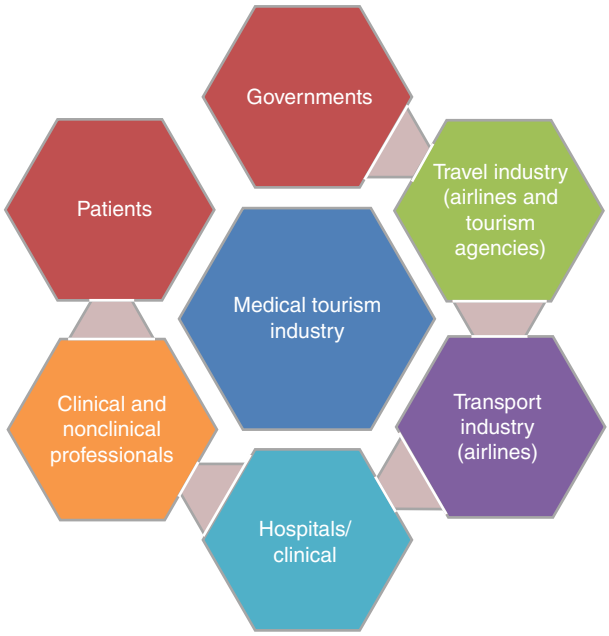


Figure 2 Main stakeholders in the medical tourism industry.

- Service delivery: The most effective medical tourism hospitals are the ones that provide services that are superior quality, reliable, timely, accessible, cost effective, and globally identical to some of the best in the world.
- Health workforce: The workforce is well trained, highly efficient, and reliable. Additionally, the workforce needs to be affordable to maintain the low cost structure. The combination of these attributes is commonly found in healthcare professionals, providing care services in the medical tourism industry.
- Health information: These hospitals ensure that the consumers are provided with current, relevant, and non-asymmetric information. Gaining trust of patients through transparency is what these hospitals aim to achieve at the point of interest initiation. The advancement of information and communications technologies allows consumers to access health information from various electronic sources at their fingertips.
- Medical technologies: To build competitiveness and illustrate advanced medical capabilities, medical tourism providers have invested in medical technologies that are most advanced, cost effective, scientifically sound, and safety and quality focused.
- Health financing: The health financing system is entitled to providing affordable, necessary medical services to patients. Evidently, medical tourism 'provides incentives for providers and users to be efficient' simultaneously.
- Leadership and governance: Strong leadership and effective governance drive strategic thinking to create new care delivery models that meet consumer demands and to generate an accountable care environment via policy and regulation oversight.

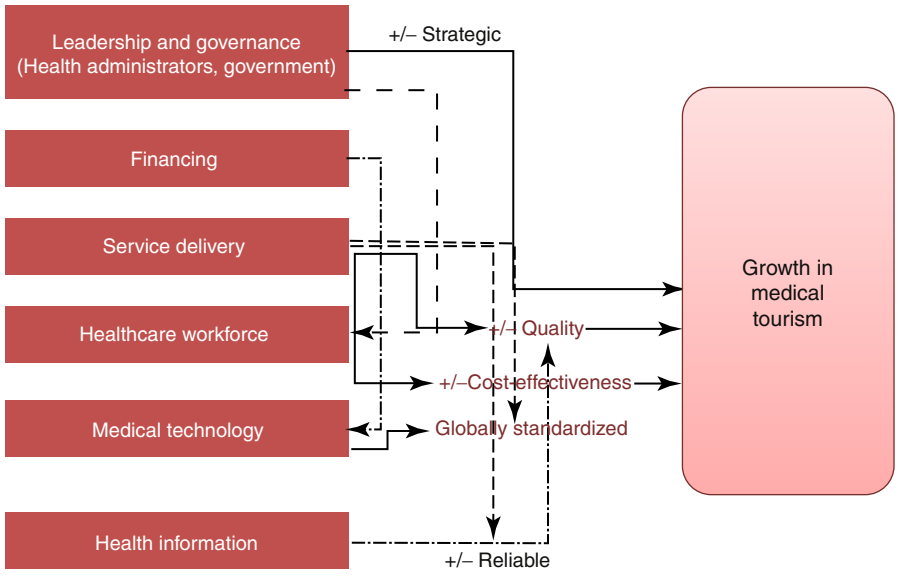


Figure 3 Medical tourism through systems thinking. Reproduced from De Savigny, D. and Adam, T. (eds) (2009). *Systems thinking for health systems strengthening*. Alliance for Health Policy and Systems Research, WHO.

Box 1 Perspective

Perspective on Bumrungrad International Hospital

Bumrungrad International Hospital based in Bangkok, Thailand, was founded in 1980 as a 200-bed facility. Since its inception, Bumrungrad has expanded to a 554-beds facility that delivers over 30 specialty services. It is the largest hospital in southeast Asia that meets international standards. Bumrungrad is the first hospital in Asia accredited by the Joint Commission International since 2002 as well as the first hospital in Thailand accredited by the Institute of Hospital Quality Improvement and Accreditation since1990. Bumrungrad opened its outpatient clinic building in 2008. Bumrungrad is a for-profit hospital traded on the Stock Exchanges of Thailand.

Bumrungrad's vision is to offer 'world class medicine' and 'world class service' and its mission is 'to provide world class healthcare with care and compassion.' These vision and mission statements have been the compass directing Bumrungrad in its venture toward being the top international hospital of Asia and probably of the world. As an international hospital, Bumrungrad annually receives an average of 400 000 patients from more than 190 countries, which comprise 40% of all its patients. It is able to do so thanks to various factors such as its workforce, office support, infrastructure, and technologies. The majority of Bumrungrad medical and managerial workforce pursued education and completed training from developed countries such as the UK, Germany, Japan, Australia, or the US. Bumrungrad employs over 1100 physicians and dentists, 200 of whom are US board certified. Bumrungrad employs over 150 interpreters who facilitate the communication between foreign patients and the clinical staff. To better support its foreign patients, Bumrungrad established International Referral Offices comprising 16 international offices based in 16 countries, overseas offices through its liaison between Bumrungrad and patients outside Thailand.

Bumrungrad was designed as a hospital that delivers its services in a magnificent hotel infrastructure. Its main lobby, which contains five shops and a bank, is reminiscent of a luxurious and spacious five-star hotel lobby; its hospital rooms are equipped with bed-side laptop, Wi-Fi, and interactive television with on-demand movies option. Even the hospital rooms are classified according to hotel standards such as single deluxe rooms, premier suites, premier atrium suites, and premier royal suites. Its 51 hospitality suites are self-contained apartments with inside pool and fitness facility. The top floors of the hospital building contain six international restaurants as well as a premium member lounge. These kind of amenities appeal to patients as they feel like they are in a hotel instead of being sick in a hospital.

Furthermore, Bumrungrad's infrastructure is equipped with state-of-the-art medical, pharmaceutical, laboratory, and information technologies. All these technologies were adopted mainly to improve healthcare quality, patient safety, and efficiency. Thus, medical staff and technicians spend less time on routine activities and devote more time on direct patient care. Bumrungrad uses advanced medical technologies such as digital mammography and image-guided radiotherapy. In addition, Bumrungrad owns a pharmacy robot, which is a fully automated drug management system. Bumrungrad' s medical lab automation is a comprehensive system that performs medication packaging, storage, and dispensing.

Bumrungrad has partnered with Microsoft for the implementation of Microsoft Amalga Health Information System, a fully integrated hospital information system that has streamlined its information processing activities. Although the Veteran Health Administration has the most advanced electronic medical record system in the US, its health information technology specialists were sent to Bangkok to study the Bumrungrad information system. In 2008, Bumrungrad received the 'Best Wireless Project South East Asia' award from Motorola's Enterprise Mobility Business.

In addition to the provision of healthcare, Bumrungrad conducts clinical research in the Bumrungrad International Clinical Research Center, which was established in 2001. Its research activities focus on clinical trials on prescription drugs and biomedical and social science research.

Although Bumrungrad caters for the wealthy locals and foreigners and is being propelled to its status as a world-class leader in healthcare, it has not forgotten those who cannot afford to pay for high-cost services. It created a foundation, the Bumrungrad Hospital Foundation to provide charitable services for the underprivileged Thai population. Such services included the provision of 122 pediatric heart surgeries and health education.

Thanks to Bumrungrad' s continuous strive for hospital wide quality and excellence, it has been ranked sixth of the top ten Thai companies, named among the six most admired Thai companies in terms of corporate reputation, quality, and innovation, and considered one of the best destinations for medical tourists.

Author: Zo Ramamonjiarivelo, Ph.D., Assistant Professor, Governor State University, Chicago, US.

Economic Drivers of Medical Tourism

As mentioned in the section Growth of Medical Tourism, globalization of markets and services serve as a strong enabler of healthcare trades between developed countries and developing countries. Specifically, it forces the global healthcare markets to be smaller and interconnected and opens opportunities for the movement of consumers and resources (i.e., healthcare professionals, medical and information technology, pharmaceutical supplies, capital funding, and international laws and regulations) across national borders. As a result, medical tourism has influenced the economic discourse of healthcare services consumption and production. Like other economic goods, the growth of medical tourism inherently rests on the fundamental economic principles of demand, supply, price, and value.

Because a transaction of medical tourism involves a number of relevant stakeholders, such as patients, providers, insurance companies, governments of the patient's countries, and government of the hosting countries (Figure 2), it is complicated to describe the economic drivers by a stakeholder. This article will discuss the economic drivers through the demand and supply lenses in which the multifaceted perspectives of different stakeholders can be well incorporated. Note that the following discussions will focus on the outbound medical tourism from developed countries to developing or hosting countries. Healthcare systems across countries around the world have confronted different unique challenges, which fundamentally result from how the systems are set up.

Multiple US federal mandates and initiatives have been enacted with the aims of sustainably bending the healthcare cost curve and improving access and quality of care. Meanwhile, a number of Americans are struggling with their chronic or acute illnesses every day, requiring necessary medical treatments, many of whom are uninsured or underinsured. Healthcare providers, especially hospitals, are under a lot of pressures caused by reimbursement cuts and new payment models. The result is increasing operating expenses while being required by law to stabilize any patients present at their facilities regardless of their insurance status. Hospitals that keep taking on the burdens caused by unpaid services are in financial distress causing some to discontinue their operations. This results in a negative domino effect on the other

existing hospitals. In addition, underinsured patients are not authorized for certain procedures and cannot afford the difference between the charges and amount their payers reimburse the provider. As a result, price plays a major role for uninsured and underinsured Americans to receive proper treatments, forcing them to search for options overseas where a similar procedure plus travel and accommodation expenses cost much less. Lack of cost transparency for the medical care provided and uncontrolled escalation of medical service and technology costs add more challenges within the healthcare market.

Canada and England are samples of other Organization for Economic Cooperation and Development (OECD) countries confronting their own issues of waiting time. Although the government will take care of sick patients, they are required to wait a number of weeks or months for certain medical procedures causing unbearable pains and discomforts. In addition, some medical services such as fertility and stem-cell treatments have yet to be legalized or approved in many western countries, leaving patients with no choice but to seek the treatments available elsewhere. Given that these medical tourism facilities are well equipped with the state-of-the-art technologies and staffed with western-trained providers, some patients perceive them as providing higher quality of treatments when compared to local providers. The abundance of information technology also allows patients to gather information on medical treatment options from various sources, which assist their decision making process and help protect themselves from information asymmetry and provider-induced demand. As a result, there exist unmet demands of medical services among citizens of these developed countries. Price, unavailable treatment options, long waiting time, and ability to make a decision based on alternative preferences are some of the major drivers on the demand of medical tourism. In summary, price and value in medical tourism can generally be reflected through the notion that world-class quality care is rendered to patients from developed countries at third world prices.

When there is a demand, there will always be an opportunity for a new market, and those who recognize it and have capabilities to supply the needs can take the pie. The supply side of this medical tourism equation is obviously reflected through the hosting or developing countries that provide medical services to the patients from developed countries with unmet needs. Healthcare delivery organizations in developing countries have quickly grasped on the concept of medical tourism and these untapped demands, emphasizing on quality of care and world-class customer services. Some perceive it as a new business development strategy when domestic private patients have turned to publicly funded healthcare systems such as academic health centers where high quality and advanced care is rendered at a lower price. Some countries have seen the great economic potential of medical tourism turned it into the national agenda and supported the provider organizations both financially and policy-wise (e.g., India and Thailand). For example, opening of more health tourist visa approvals, more subsidies on ventures encouraging tourists in the country, and legalization of certain medical procedures such as surrogacy.

The main competitive advantage that these provider organizations have over their counterparts in the developed countries is the cost of labor and supplies. Labor is a major operating cost for healthcare delivery organizations. Because labor costs in these developing countries, even for healthcare professionals, are much lower than those in developed countries, these organizations are operating at a fraction of the cost incurred by their western counterparts, allowing them to develop attractive pricing strategies. In addition, the medical and pharmaceutical supplies in developing countries are at a much lower negotiated price, part of which is due to international trade agreements.

Owing to the dramatic success in the medical tourism industry within a relatively short timeframe, there is an increasing trend in partnership and investment between health systems in the developed countries and medical tourism providers in developing countries, resulting in the flow of capital funding to accelerate the growth of medical tourism. Over the years, medical tourism has expanded its client base from only out-of-pocket individuals to patients with private or employer insurance because both sides share a win-win situation. The potential revenue streams from these groups of patients have led to different types of arrangements between insurance companies in developed countries and providers in developing countries. As a result, low operating costs and increasing capital are two main factors driving the growth of medical tourism on the supply side.

Health Policy Issues in Medical Tourism

Although medical tourism has created tremendous economic gains to the hosting countries and health benefits to the patients receiving care, it at the same time has undergone a number of legal, social, and ethical criticisms and forfeiting challenges. The emerging trade of medical services brings forth a number of regulatory and policy implications at international and country levels affecting both developed and developing countries. Given that medical tourism is relatively in its infancy and merely merits dramatic attention in recent years, there is no existing law or regulation that specifically aims to enforce appropriate actions, penalize misconducts in medical treatments, and prevent future wrongdoing around this new pattern of medical practice.

Certain regulations and policies aim at supporting the medical tourism industry. There exist some international accreditation organizations such as the Joint Commission International (based in the US) and Trent Accreditation Scheme (based in UK–Europe) from which medical tourism organizations seek accreditation to demonstrate their commitment to high-quality care and patient safety. By being recognized by the world-renowned accreditation bodies, these care delivery organizations can not only provide a level of assurance and comfort to patients traveling from developed countries but also leverage the accreditation as a competitive advantage that could also attract local patients.

To support the growth of medical tourism, some developing countries have relaxed their visa process or established a special type of visa for patients seeking medical treatments in the countries. Additionally, with the great attempt to bend the healthcare cost curve, the US Congress

recently held a Senate hearing on the promise of medical tourism titled 'The Globalization of Health Care: Can Medical Tourism Reduce Health Care Costs?' while there is also a push for Medicare and Medicaid to reimburse medical tourism services. Furthermore, insurance companies have caught on the trend and started incorporating medical tourism into their plans, hoping to decrease the cost burdens they would otherwise have borne should the patients were treated in the US.

However, medical tourism has incurred unintended consequences for which no laws and regulations have yet been prepared. Privacy and security of patient information is of great concern. Although the information created or captured on the US soil is protected by the Health Insurance Portability and Accountability Act, healthcare organizations in other countries are not bound by the law, which in turn makes it difficult to handle any violations. Not only patients from developed countries but also the citizens of the hosting countries could be negatively impacted by medical tourism. The economics doctrine was established on the notion of scarcity where limited resources can never fulfill indefinite wants and needs. As a result, when resources are redirected to serve the needs of spillover patients from the developed countries, there will be negative consequences to the unmet needs of basic healthcare services among the consumers in the hosting countries. This poses highly concerned threats to the local patients, especially among the people with financial constraints. Statistics have shown that the ratio of medical providers to patients in developing countries is much lower than that of developed countries. In addition, on top of the existing issue of external brain drain, the heavy recruitment of highly skilled, experienced, western-trained healthcare practitioners into medical tourism facilities has driven a more severe incidence of provider shortage in the public health system of the developing countries (i.e., internal brain drain). Although there is currently no empirical study that shows that the magnitude of medical tourism impacts on provider shortage and healthcare access of consumers in the developing countries, some anecdotal evidence point to the stark increasing disparities.

Without an effective policy set forth by home country governments or international bodies, the issues could have further implications on social gaps and potentially threaten the health-related goals put in place by the countries or international organizations such as WHO and United Nations. To ensure the mitigation of negative effects, home country governments and international regulatory bodies must work independently and cooperatively to prioritize an understanding of these issues and subsequently develop policies that address these arduous obstacles in both short and long term.

The Future of Medical Tourism

Medical tourism has generated global competition in the delivery of healthcare market. In this competition, the more expensive and resourceful countries are increasingly disadvantaged. However, one cannot remain a nonparticipating player in this lucrative opportunity. As a result, a number of frontrunners in the US healthcare have expanded their strategic coverage to compete for market share in medical tourism. For example, Philadelphia International Medicine is building a hospital in Korea; Harvard University has partnered with India's Wockhardt Hospital and the United Arab Emirates to create the Dubai Healthcare City; and The Johns Hopkins University and Tufts University have opened hospitals in India. Such collaboration and partnership have propelled the image of affordable and high-quality care to new heights.

In the long run, medical tourism may also help insurance providers and employers reduce their costs and provide their employees with alternatives for satisfactory coverage. Only a few insurers offer care overseas, but they have already reported remarkable savings. For example, Blue Cross/Blue Shield of South Carolina gives customers a medical tourism option through its Companion Global Healthcare and even offers coordination of foreign care for preexisting and noncovered procedures. Employers large and small are beginning to offer incentives to their employees to have routine care done outside the country, observing considerable savings. For example, Hannaford, a US grocery chain, found that sending its employees to Singapore for knee or hip replacements lowered the price from $43 000 to $9000.

Medical tourism challenges the ability of nation-states and national governments to control affairs within their own jurisdiction. The lack of legal options for patients choosing healthcare overseas, ethical and social justice issues of traveling abroad for organ transplant and reproductive tourism and concerns about large international hospitals providing level of care unaffordable for citizens is a concern raised by opponents. To address these concerns new international steps are taken among a coalition of countries resulting in the Declaration of Istanbul on Organ Trafficking and Transplant Tourism in May, 2008. Nonprofit organizations promoting international medical travel vouching for quality healthcare are also growing; one of such organizations is Medical Tourism Association that currently has 20 000 members and a slick website and magazine. Additionally, organizations such as Joint Commission International – the international division of The Joint Commission, have been working with healthcare organizations, ministries of health, and global organizations in more than 80 countries. Conferences such as the World Medical Tourism and Global Health Conference are being held every year and are attended by representatives from major international medical centers looking for business, Western firms coordinating medical travel and care, and third-party payers and insurance companies seeking ways to contain costs.

On a general note, globalization of healthcare in the form of medical tourism has several advantages:

1. Economies of scale: The production of medical services in an efficient and standardized format encourages value creation in healthcare delivery.
2. Better consumer purchasing power: Apart from quality, the central focus of this industry is cost-efficiency, which gives consumers (patients) more control of their treatment options and destinations.
3. Faster adoption of innovative healthcare treatments and technologies: The presence of innumerable competitors acts as driver for the early adoption of evidence-based

innovative treatment plans and options. It additionally also creates more acceptability of the services that are provided by hospitals.

4. Emergence of global market segments: As the industry grows, more clustering of this industry is being observed. Presently, the Asian region (India, Thailand, Malaysia, and Singapore) seem to be sprouting medical tourism hospitals at an exponential speed.
5. Development of synergies: The global scale of the healthcare industry is also an opportunity for transfer of ideas on healthcare products and treatments from one country to another. Additionally, the global marketing creates more experience in operating in multicultural and diverse environments.

The key to success for medical tourism therefore lies in the strategic global economic growth, focusing on how revenue from worldwide patient travel translates into output, jobs, and income. It will provide an opportunity to both developing and developed countries. Although developing countries will benefit from the travel resulting in revenue generation from the healthcare services provided, developed countries will benefit from the competitiveness of the market as a stop point to reflect on their cost, quality, and accessibility issues in providing their citizens value-driven healthcare. Even though medical tourism is confronting some social and ethical challenges, it without a doubt is an explicit example of international economics in action where the flow of consumers and resources beyond local boundary promotes an affordable, accessible, and quality healthcare environment.

See also: Competition on the Hospital Sector. Emerging Infections, the International Health Regulations, and Macro-Economy. Global Health Initiatives and Financing for Health. Health and Health Care, Need for. Health Insurance and Health. International E-Health and National Health Care Systems. International Movement of Capital in Health Services. International Trade in Health Services and Health Impacts. International Trade in Health Workers. Medical Tourism. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Resource Allocation Funding Formulae, Efficiency of. Theory of System Level Efficiency in Health Care. Willingness to Pay for Health

Further Reading

Bookman, M. Z. and Bookman, K. R. (2007). *Medical tourism in developing countries*. United States: Palgrave Macmillan.

Carrera, P. M. and Bridges, J. F. (2006). Globalization and healthcare: Understanding health and medical tourism. *Expert Review of Pharmacoeconomics and Outcomes Research* **6**(4), 447–454.

Ehrbeck, T., Guevara, C. and Mango, P. D. (2008). Mapping the market for medical travel. *McKinsey Quarterly*. https://www.mckinseyquarterly.com/Mapping_the_market_for_travel_2134 (accessed 17.02.13).

Herrick, D. M. (2007). Medical tourism: Global competition in health care. *NCPA Policy Report*. http://www.ncpa.org/pub/st304 (accessed 17.02.13).

Horowitz, M. D., Rosensweig, J. A. and Jones, C. A. (2007). Medical tourism: Globalization of the healthcare marketplace. *Medscape General Medicine* **9**(4), 33.

Turner, L. (2010). 'Medical tourism' and the global marketplace in health services: US patients, international hospitals, and the search for affordable health care. *International Journal of Health Services* **40**(3), 443–467.

Unti, J. A. (2009). Medical and surgical tourism: The new world of health care globalization and what it means for the practicing surgeon. *Bulletin of American College of Surgery* **94**(4), 18–25.

Woodman, J. (2008). *Patients beyond borders: Everybody's guide to affordable, world-class medical travel*. United States: Catawba Publishing LLC.

York, D. (2008). Medical tourism: The trend toward outsourcing medical procedures to foreign countries. *Journal of Continuing Education in the Health Professions* **28**(2), 99–102.

Unfair Health Inequality

M Fleurbaey, Princeton University, Princeton, NJ, USA
E Schokkaert, KU Leuven, Leuven, Belgium

Glossary

Atkinson inequality index An index of inequality that has explicit weights attaching to the measured variable (income, health, etc.) at various levels. This sensitivity parameter ranges from 0 (indifference about the nature of the distribution) to infinity (concern attaches only to the position of the very lowest group).

Capabilities The set of all possible physical and social functioning for a person.

Concentration index A measure of the degree of income-related inequality in health. Where there is no income-related inequality, the concentration index is zero. A negative value indicates a disproportionate concentration of ill-health among the poor.

Direct unfairness Inequalities in health or health care after one has removed the effect of determinants not considered to be or to lead to unfairness.

Fairness gap The gap between an actual and a hypothetical distribution of health in which all legitimate grounds for inequality have been removed.

Gini coefficient A number between 0 and 1, where 0 corresponds to perfect equality (everyone has the same income, health care, etc.) and 1 is perfect inequality (one person has all the income, health care, etc.).

Lorenz curve A graph showing the cumulative percentage of income, health expenditures, etc. held by successive percentiles of the population.

Standardization The adjustment of raw data to avoid making false inferences arising from confounding factors.

Introduction

A fair society should give individuals equal opportunities to realize their own life project. Health is of utmost importance for the flourishing of individuals. It seems, therefore, self-evident that inequality in health should get an important place on the fairness agenda. Yet, this seemingly obvious statement raises difficult issues. First, is all inequality in health necessarily unfair? Some health inequalities can be seen as 'unavoidable,' because they are due to biological factors or simply reflect bad luck. Should we not rather target those inequalities that are caused by the organization of our society, and in particular the health inequalities that are linked to indicators of socioeconomic status such as income, wealth, education, and social class? Socioeconomic inequalities have indeed been the main focus of the research, both in the public health and in the economic literature, and they also figure most prominently in policy statements. Yet, this raises a second, similar, question: Are all socioeconomic inequalities necessarily unfair? What if they are partly caused by individual behavior, such as smoking and drinking, or by choices about where to live and what kind of work to pursue? Should people not be held responsible for their lifestyle choices? And if so, should our measure of unfairness not in one way or another take into account this element of individual responsibility? Third, no matter how important health is for human flourishing, it is not the only important dimension of well-being. Does it make sense to focus on health only? Should we not integrate health inequality in an overall view of unfair inequality in well-being?

These questions are the main focus of this entry and therefore other important issues are left aside. First, an explicit defense of egalitarianism will not be constructed and it will simply be taken for granted that some form of equality is necessary for fairness. The real question is: Equality of what? Second, the possible trade-off between total health and its distribution will not be considered. If spreading information about healthy lifestyles leads to an increase in average life expectancy but at the same time to growing inequality (e.g., because different cognitive capacities lead to differences in the efficiency of processing this information), a complete evaluation of the policy requires trading off these two effects. The focus here is on the specification of the fairness element in this trade-off. Third, unfairness is not exclusively a matter of health outcomes but has also a procedural element: Many will not accept that unfair health inequalities should be tackled by introducing explicit discriminatory practices into the process of accessing health care. Fourth, health can be measured in many different ways. Mortality is one possible indicator, the number of chronic conditions another; and much work is based on subjective self-assessed health, either on a continuous scale or in discrete categories. Different health concepts may yield different fairness results. Moreover, the level at which these variables are measured will determine the kind of inequality measures that can be used. These measurement issues will be left aside and the focus will be on the conceptual question: What is unfair health inequality?

Pure Health Inequality

The most straightforward approach is of course to consider simply all health inequalities as 'unfair.' Provided that health can be measured on a ratio scale, the degree of unfair health inequality can then be gauged by any of the measures that

have been developed in the literature on income inequality (such as the Gini or Atkinson coefficients), and one can also draw the traditional Lorenz curve with the cumulative share of the population on the horizontal axis and health on the vertical axis. The closer this curve is to the diagonal, the smaller is inequality in health. The only difficult issue in this context is the choice of an adequate measure of health. All the rest is standard.

However, the question is, whether such pure health inequality an interesting concept? Suppose the inter-regional differences are to be checked in the performance of a health care system. It is observed that mortality is higher in region A than it is in region B. Yet, it is also observed that the population is on average older in A than it is in B. In that case it could be highly misleading to derive conclusions about the relative performance of the health care system in different regions from the simple differences in mortality. A correction for age seems necessary. In this spirit, the epidemiological literature has derived different methods of standardization of the raw measures by making use of the information from a reference group (e.g., the overall population in the country). Direct standardization estimates the mortality rate that would be obtained in regions A and B in the hypothetical situation in which they had the same age structure as the reference group. Indirect standardization first calculates the hypothetical mortality rate that would have been observed in regions A and B if the mortality rates for the reference population were applied to the age structure of the respective regions. One then computes standardized mortality rates by taking the ratio of the observed mortality rates with these hypothetical rates.

Although the use of standardization seems justified in this application, it has also been advocated for measuring unfairness in health. The authoritative World Health Organization Commission on Social Determinants of Health has emphasized that health inequity only arises where systematic differences in health are judged to be avoidable by reasonable action. Naively applied, this view implies that age and gender differences in health should not be seen as unfair because they can be largely explained by biological factors that are 'unavoidable.' One may deplore these differences, but nature itself cannot be fair or unfair. Although this is a popular opinion, it is controversial. It is hard to deny that social and technological developments interact with this biological background. This is especially obvious for gender: Health inequalities between men and women are definitely not only caused by biological factors but also by the position of men and women in society, including the way they are treated by the health care system and in the labor market. The rapid increase in male but not female mortality in the 1990s in Russia and other former Soviet Union countries, during the transition from a planned to a market economy, is a striking illustration of how socioeconomic factors can impact on gender inequalities in health. The same point can also be made with respect to age: Everybody dies, but health and mortality among the elderly depend on the way society is organized. Even the effects of different genetic endowments cannot really be seen as 'unavoidable:' Not only will the rapid technological developments in the domain of total genome analysis increase the potential of interventions in the near future (e.g., for eradicating diseases caused by genetic defects) but also it has become clear that phenotypical differences are almost always the product of interactions between the genetic endowment and the socioeconomic and natural environment. The latter can be influenced by policy. Biological differences do exist, but they do not completely determine the resulting health situations. The practice of quasi-automatic standardization for age (and even worse, for gender) may, therefore, hide important aspects of unfairness that follow from the differences in the treatment of women and the elderly in different countries or in different time periods.

Socioeconomic Health Inequalities

Whatever the position taken with respect to the effects of biological factors, most people would agree that health inequalities related to socioeconomic status in terms of income, wealth, education, or social class are particularly unfair. Differences in socioeconomic status are on their own already an indicator of injustice – and things get worse if individuals with a better socioeconomic status also are in better health. Both the public health and the economic literature have by now produced overwhelming evidence for the existence of such socioeconomic inequalities in health, with different health measures, for different countries and in different time periods.

Regression-based measures based on gaps or ratios between two extreme groups have been especially popular in the public health literature. These have the advantage of being very simple to interpret. As an example, Figure 1 shows the ratio of the estimated death rate from any cause for males at the lowest level of education over the estimated death rate from any cause for males at the highest level of education. All measures are standardized for age. The results are striking. In countries like Hungary, the Czech Republic, and Poland, mortality differs between the lower and upper ends of the education scale by a factor of more than 4. Similar results are found with other measures of socioeconomic status (such as income and occupation) and indicators of health (such as self-assessed health).

Economists have studied the same phenomenon using the concentration index. The concentration index is a measure of the area between the concentration curve and the diagonal, where the concentration curve is drawn with the cumulative share of socioeconomic status (e.g., income) on the horizontal axis and the cumulative share of health on the vertical axis. Compared with the extreme group measures often used in the public health literature, the concentration index has the advantage of taking into consideration the whole distribution. Here also, standardization for age (and sex) is quite common.

Although their methods differ, the economic and public health literature concur with each other in finding strong socioeconomic inequalities in health. This is definitely a finding with much relevance for evaluating the fairness of the social arrangements. Yet, from a broader perspective, two questions can be raised. Are all socioeconomic inequalities necessarily unfair? And how do socioeconomic inequalities fit in a broader view on fair health inequality?

Are All Socioeconomic Inequalities 'Unfair?'

After having observed and measured socioeconomic inequalities in health, a logical next step is its explanation. Different

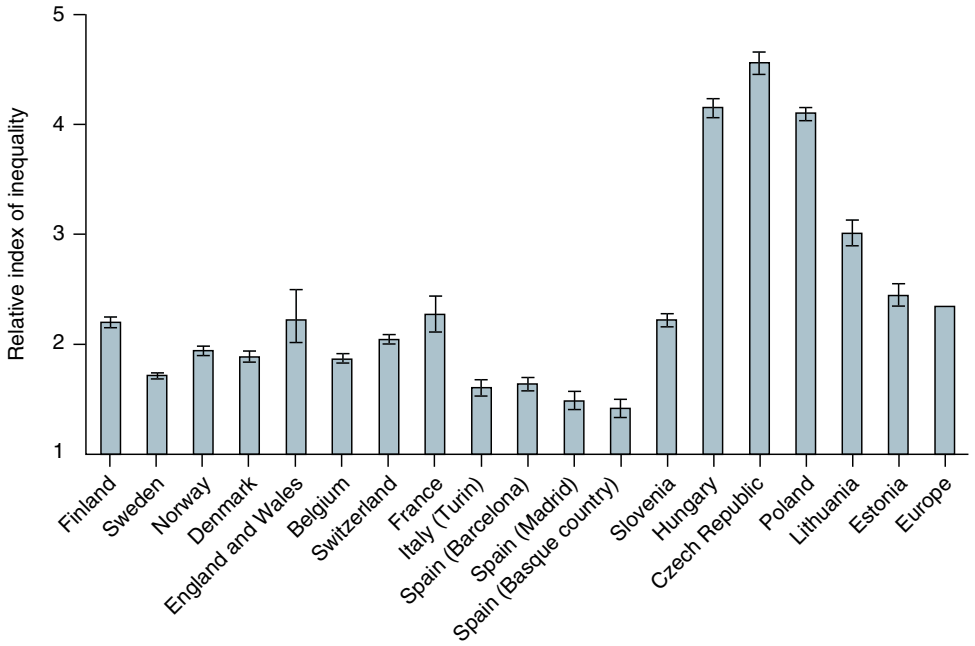


Figure 1 Relative inequality in the rate of death for any cause in different European countries (for men, according to education). Reproduced from Mackenbach, J., Stirbu, I., Roskam, A. J., et al. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine* **358**(23), 2468–2481.

factors have been documented in the literature. Health and mortality inequalities may be caused by differences in working and housing conditions or by differences in access to good quality health care (apparently, an important factor explaining the poor results for the Eastern European countries in Figure 1). Most people will agree that these are indications of unfairness. However, socioeconomic differences in lifestyles are another important cause of the health differences, for example, smoking has a huge effect on health. This empirical finding has led to a sometimes heated debate about the fairness of the resulting inequalities. Should people be held responsible for differences in lifestyles? A positive answer to this question would mean that at least a fraction of the observed socioeconomic inequalities is not unfair.

The debate on the causes of socioeconomic inequalities in health has recently been considerably enriched and deepened by the rapidly growing empirical literature on the effects of childhood circumstances. Childhood circumstances do not only have a direct effect on adult health but also influence adult socioeconomic status and adult lifestyles. Different channels of influence have been documented. First, there is a direct effect of the prenatal (fetal) environment on adult health and lifestyle. As an example, it has been shown that the cohort of children born from mothers who were pregnant during the influenza pandemic of 1918 had a larger chance of being physically disabled in 1970. Second, illness and socioeconomic status in childhood may have lasting effects on adult health and lifestyle. This may be true even if adult income has hardly any effect on adult health. Third, childhood circumstances may affect the socioeconomic status (including the level of human capital) of young adults and this then may influence adult health and lifestyle. The three channels can work together, but opinions differ about their relative importance. Such different beliefs may lead to different ideas about the (degree of) unfairness of the observed socioeconomic health inequalities if the latter are caused by differences in lifestyles. This issue will be discussed in the next Section Unfairness and the Causes of Inequality: A General Framework.

Other 'Unfair' Inequalities

The literature on socioeconomic inequalities in health can also be seen as too narrow from another perspective. Even when disregarding age–gender differences (a position that, as argued, is not beyond criticism), socioeconomic inequalities are not the only cause of unfair health inequalities. Another (obvious) example is provided by regional inequalities. Regional inequalities in health may be linked to differences in economic infrastructure and amenities and the quality of the overall living environment. They may also be caused by differences in the relative performance of the health care system – a natural indicator of unfairness in an National Health Service context where a central government decides about the regional distribution of the available funds. In the US, the debate on health disparities has mainly focused on the effect of race, which can certainly be seen prima facie as a case of unfair health inequalities.

A priori, one might think that it is possible to distinguish explicitly these different examples of unfair inequalities and to analyze each of them separately. Although such a separated approach indeed may generate useful insights, it begs the question of how these different 'unfair inequalities' should be aggregated in order to obtain an overall measure of unfair health inequality in a given country at a given point in time. Moreover, there may be important interactions between the

various 'unfair' inequalities. Focusing only on race in the US context means that one will tend to neglect the important fact that socioeconomic status mediates (at least partly) the relationship between race and health. However, focusing on socioeconomic status only may lead one to forget the fact that there may be (unfair) health differences between people from different races but with the same socioeconomic status. The choice of perspective may reflect a philosophically different view of the world and may have political consequences. Indeed, focusing either on race or on socioeconomic status alone will inspire different policy measures.

Unfairness and the Causes of Inequality: A General Framework

The different questions and remarks raised in the previous sections converge on the same basic idea: there are many different factors leading to health inequalities. Some of these explanatory factors point to unfairness (e.g., socioeconomic status, race, and access to health care); others may reflect individual responsibility (e.g., lifestyles). Inequalities resulting from the former may be seen as ethically illegitimate, inequalities resulting from the latter may not offer a reason for concern. If this general and abstract picture of the world is accepted, a coherent framework is needed that makes it possible to integrate these different aspects.

Suppose, for simplicity, that the health situation of an individual is determined by two variables only: income and lifestyle. Suppose also that the position is taken that health differences due to income are unfair, whereas due to lifestyle are legitimate. Two caveats are in order here. The method is not limited to two variables (it can be applied to any number of explanatory factors) and the method does not presuppose that individuals are responsible for lifestyle (it can be applied for any partitioning of the set of explanatory variables in 'legitimate' and 'illegitimate' factors). This simplistic example is considered only to illustrate a more general method.

A first approach is to calculate for each individual the health status a person would reach with his/her own income but with a reference value for lifestyle. Because this reference lifestyle is kept the same for all individuals, the inequality in the resulting hypothetical values will only reflect income differences and is, therefore, a measure of unfair health inequality. Call this measure direct unfairness. A second approach calculates for each individual the health status a person would obtain with his/her own lifestyle but with a reference value for income. This can be interpreted as the health situation a person would reach in a 'fair' situation, because the (unfair) effect of income differences is removed. The difference (or the ratio) between an individual's actual health status and this hypothetical fair health status can be seen as an individual fairness gap – and the inequality in these fairness gaps is a measure of overall unfair health inequality. Note the very close analogy between direct unfairness and the fairness gap on the one hand and direct and indirect standardization on the other hand.

In general, the two measures (direct unfairness and the fairness gap) do not coincide. How then to choose between them? It seems natural to impose that an adequate measure of unfair inequality should only be zero if there are no illegitimate inequalities left, i.e., if two individuals with the same lifestyle reach the same health outcome. It can be shown that the fairness gap satisfies this so-called compensation requirement, whereas direct unfairness does not. In this sense, the former is to be preferred. There is a price to be paid, however: the compensation requirement implies that if lifestyle affects health differently in different socioeconomic groups, the resulting health differences are interpreted as unfair. If one prefers a stricter position on responsibility and considers all lifestyle effects as fair, one should rather focus on the measure of direct unfairness.

As emphasized before, the methods sketched are general and can be applied to any partitioning of the set of explanatory variables into 'legitimate' and 'illegitimate' causes of health inequalities. Age–gender standardization boils down to considering 'age' and 'gender' as legitimate sources of differences. Focusing on socioeconomic or racial or regional inequality means that one interprets socioeconomic status or race or region as an illegitimate source of inequality, and (implicitly) all other sources as legitimate. Combinations are also possible, of course. Even pure health inequality can be accommodated in this broader framework: it is the extreme case in which all the causes of inequality are seen as unfair. One of the advantages of the general method described here is that it allows for sensitivity analysis, i.e., measures of unfairness can be calculated and compared for different possible partitionings.

Ultimately, the choice between different interpretations of unfairness should be made on philosophical or ethical grounds. Broadly speaking, it is possible to distinguish two approaches in the philosophical and welfare economic literature on the topic. The first defines responsibility as control. In this view, individuals should be held responsible only for those variables that they in one way or other choose themselves. At first sight, this seems a natural approach and it has also become the most popular. Yet, it raises the difficult question of what is really under the control of individuals in a social science perspective with a deterministic view of the world. If the findings on genetic and prenatal influences and on childhood circumstances are taken seriously, there does not seem to be much room left for genuine choice. A second approach holds people responsible for their preferences, i.e., for their own life project, even if this life project is not fully under their control and is (unavoidably) influenced by their education and social environment. This latter approach looks less like a kind of 'disciplining' device and can also be formulated in emancipatory terms as a way to respect the dignity of all individuals by giving them the freedom to choose their own lifestyle.

From a pragmatic point of view, one has to come down from these broad philosophical perspectives to classify the specific empirical variables. This is not a trivial exercise, and different observers will have different opinions. Is level of education a matter of choice? Does smoking behavior under the influence of social pressure and advertising reveal genuine preferences about a life project? One cannot give a convincing answer to these questions – and therefore one cannot construct an adequate measure of unfair health inequality – if one does not first have a good insight into the different channels

through which these specific variables affect health outcomes. Indeed, there is no a priori reason to treat the various channels similarly in terms of fairness. Consider socioeconomic status. Its influence on health may reflect a direct effect of genetic endowment, the prenatal environment, and/or childhood circumstances; it may capture differences in lifestyles because of different capacities of information processing as a result of differences in human capital that for their part may follow from differences in childhood circumstances or from educational choices much later in life; it may reflect differences in health behavior that reflect different ideas about what is important in life; it may follow from differences in working conditions, themselves partly chosen but from a restricted opportunity set. To measure adequately unfair health inequality, a good explanatory framework is first needed that distinguishes between these different channels as well as possible.

A good explanatory framework is not only necessary to measure unfair inequality but also it is essential from a policy point of view: Health and health inequalities are not only influenced by health care arrangements. Quite the contrary, certainly if the interest is in the health of the most vulnerable social groups, labor market status, working conditions, housing, and education are at least as important. This raises immediately the deeper question about how to fit unfair health inequality into the broader picture of overall unfairness.

Health and Well-Being

Consider a policy that lowers income support for the most vulnerable groups in society but makes a huge investment in their access to health care. The result is a marginal decrease in unfair health inequality, but a considerable increase in income inequality. Should it be considered as a move toward a fairer society? Or what about a policy that improves the labor market opportunities for the unskilled, leading to a sharp improvement of their material well-being but at the same time to a slight increase in health inequality because of the increase in stress? In both cases, information about unfair health inequality is insufficient to evaluate the overall unfairness of these policies: Well-being has more than one dimension, and all relevant dimensions have to be taken into account if a global judgment is to be formulated.

This conclusion seems so obvious that one may wonder why the bulk of research and policy attention goes to the partial issue of socioeconomic health inequality. The answer to this puzzle seems to lie in the kind of results that are described in the Section Socioeconomic Health Inequalities: because overwhelming evidence for socioeconomic inequalities in health is found to be at the expense of the poorer groups in society, there is a cumulative effect and the trade-offs sketched in the previous paragraph may seem to be of second-order importance. Yet, this kind of contingent reasoning is not sufficient to defend a normative position in principle. As a matter of fact, tricky issues arise when one considers only these partial inequality measures. The concentration index will always decrease (suggesting a less unfair situation) if health is transferred from someone who is better off in terms of socioeconomic status to someone who is worse off, independently of their own initial health situation. Moreover, it can be shown that more egalitarian countries will do worse on the most popular measures of socioeconomic health inequalities (including the extreme group measures and the concentration index), if there is a causal link from health to income. This is a mechanical effect of the way in which these measures are constructed and it explains partly why, for example, the Scandinavian countries do not do very well in **Figure 1** (and in similar empirical exercises). All this strengthens the conclusion that it is necessary to go beyond such conditional inequality measures and move to overall inequality in well-being – taking into account, of course, that health is an essential element of well-being.

Multidimensional approaches to well-being have grown in popularity recently, as reflected in the success of Sen's capability approach. Techniques for multidimensional inequality measurement have now been firmly established. However, neither of these two approaches offers an attractive solution to the aggregation problem, i.e., the problem of weighting the importance of the various dimensions so as to obtain one overall measure of individual well-being. The capability approach leaves this question largely open, whereas multidimensional inequality measures implicitly 'solve' the problem by imposing a functional form in a rather ad hoc way. Yet, in a democratic society, it seems natural to require that the weighting of the different dimensions should reflect the preferences of the individuals themselves. Well-being does not necessarily coincide with subjective happiness either: it would be strange to claim that a healthy millionaire who feels depressed because he is not successful in having his poems published is worse off than a sick and poor woman who is reasonably satisfied with her life because she has learnt to adapt to her fate by lowering her aspirations. The real challenge consists in formulating a concept of individual well-being that does respect preferences, although at the same time correcting for aspirations. Recent developments have shown that the traditional welfare economics concepts of money-metric utility and equivalent income offer interesting perspectives in this respect.

See also: Equality of Opportunity in Health. Fetal Origins of Lifetime Health. Health and Its Value: Overview. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Intergenerational Effects on Health – *In Utero* and Early Life. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach

Further Reading

Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of *in utero* influenza exposure in the post-1940 US population. *Journal of Political Economy* **114**(4), 672–712.

Bleichrodt, H. and van Doorslaer, E. (2006). A welfare economics foundation for health inequality measurement. *Journal of Health Economics* **25**, 945–957.

Brekke, K. and Kverndokk, S. (2012). Inadequate bivariate measures of health inequality: The impact of income distribution. *Scandinavian Journal of Economics* **114**(2), 323–333.

Case, A., Fertig, A. and Paxson, C. (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics* **24**, 365–389.
Currie, J. (2009). Healthy, wealthy and wise: socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature* **47**(1), 87–122.
van Doorslaer, E. and Van Ourti, T. (2011). Measuring inequality and inequity in health and health care. In Glied, S. and Smith, P. (eds.) *Oxford handbook on health economics*, pp. 837–869. Oxford: Oxford University Press.
Fleurbaey, M., Luchini, S., Muller, C. and Schokkaert, E. (2012). Equivalent income and fair evaluation of health care. *Health Economics*. doi:10.1002/hec.2859.
Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics* **28**(1), 73–90.
Fleurbaey, M. and Schokkaert, E. (2011). Inequity in health and health care. In Barros, P., McGuire, T. and Pauly, M. (eds.) *Handbook of health economics*, vol. 2, pp. 1003–1092. New York: Elsevier.
Hausman, D. (2007). What's wrong with health inequalities? *Journal of Political Philosophy* **15**(1), 46–66.
Kawachi, I., Daniels, N. and Robinson, D. (2005). Health disparities by race and class: Why both matter. *Health Affairs* **24**(2), 343–352.
Mackenbach, J., Stirbu, I., Roskam, A. J., et al. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine* **358**(23), 2468–2481.
Sen, A. (2002). Why health equity? *Health Economics* **11**, 659–666.
WHO Commission on Social Determinants of Health (2008). *Closing the gap in a generation*. Geneva: WHO Press.
Williams, A. and Cookson, R. (2000). Equity in health. In Culyer, A. and Newhouse, J. (eds.) *Handbook of health economics*, vol. 1B, pp. 1863–1910. Amsterdam: Elsevier (North Holland).

Relevant Websites

http://www.econ.kuleuven.be/erik.schokkaert
Erik Schokkaert's page.
http://www.gapminder.org/
Gapminder.
https://sites.google.com/site/marcfleurbaey/
Marc Fleurbaey's page.
http://www.instituteofhealthequity.org/
University College London Institute of Health Equity.
http://www.who.int/social_determinants/en/
World Health Organization.

Utilities for Health States: Whom to Ask

PT Menzel, Pacific Lutheran University, Tacoma, WA, USA

Overview: Two Main Approaches

For use in cost-utility or cost-effectiveness analysis (CUA or CEA), evaluations of health outcomes in terms of quality-adjusted life-years (QALYs) require judgments of the quality of life in different health states. The quality of life of a state is often referred to as its 'individual utility.' There are two main sources for estimating a state's utility. One is people who have experienced it themselves. For example, people with paraplegia can evaluate the state of being paraplegic. Utility thus measured is called 'experience utility.' 'Ex post utility' and 'patient value' are other names for the same. But a score for quality of life in paraplegia can also be obtained by having members of the general public, who mostly have not experienced paraplegia, consider a description of its manifestations and consequences, and imagine what it would be like for them to be paraplegic. Utility assessed in this way is referred to by a variety of labels: 'ex ante utility,' 'hypothetical utility,' 'public value,' 'hypothetical patient value,' or 'nonpatient value.' A further term is 'decision utility,' which can include both ex ante and ex post utilities (see section 'Decision Utility versus Hedonic Experience').

Standard practice in QALY calculations is to use hypothetical (ex ante) utilities, on the grounds that they are more representative of the values and interests of the population at large than values from patient subgroups. But the practice has been challenged by many, and the issue is of more than mere theoretical interest. It takes on practical importance because of the empirical fact that health state utility ratings are typically higher when elicited from patients, particularly those with chronic illness or disability, than from nonpatients who only hypothetically imagine themselves in such conditions. Debate is further fueled by the fact that a major factor accounting for the discrepancy between patients' and nonpatients' values is patient adaptation to diminished health. Expressing time trade-off (TTO) preferences, for example, people with paraplegia, having experienced and adapted to it, may rate their quality of life as 0.95 (they are willing to trade away only 5% of their time alive to regain full function), whereas members of the general public imagining themselves being paraplegic may evaluate the condition as 0.8 (willing to reduce longevity by 20%). If these are the ratings for paraplegia, using patients' adapted values will reduce the value of preventing or curing paraplegia to a quarter of the value it would have if the general public's values were used ($1.0 - 0.95 = \frac{1}{4}(1.0 - 0.8)$).

Ultimately, the questions here are normative. Whose ratings of health-related quality of life should be used? Is it fair to those who do not yet have paraplegia to assess the cost-effectiveness of efforts to prevent this condition by using the higher utility values registered by someone else – patients who have adapted? Such normative discussion will be pursued explicitly in the later section Normative Considerations in Choosing an Approach of this article. The next three sections focus on descriptive and conceptual issues in experience utility.

The Facts of Variation

The so-called 'standard story' of health state valuation data is that patients, particularly people with chronic illness and disability, rate their quality of life more highly than do hypothetical patients who only imagining themselves with those conditions. For instance, of 39 studies reviewed by de Wit *et al.* (2000), 23 reported patient values higher than public ones, 2 reported higher public values, and 11 found no difference. Arnold *et al.* (2009) in a review of 32 studies, found the mean TTO value for all of the disease states evaluated by current patients to be 0.83, compared to a mean value by hypothetical patients of 0.65. Moreover, the difference in ratings does not seem due to cognitive flaws; a mood assessment study by Riis *et al.* (2005) of hemodialysis patients not only produced higher patient values but also found that patients were less flawed in their prior expectations and later recollections than were nonpatients.

Generalization, however, is dangerous. Damschroder *et al.* (2005) not only found some of the most extreme differences in valuing life with new onset of paraplegia, compared to life with preexisting paraplegia, but also found that when the nonpatients engaged in a simple exercise virtually all of the difference with patient values disappeared. In the exercise, they were merely informed about adaptation and encouraged to consider their own ability to adapt.

For some conditions, patient ratings may actually be lower than nonpatient ratings. Traumatic brain injury (TBI) is an outstanding case. The effects of TBI often involve depression, and adaptation to depression is extremely difficult. Moreover, as shown by Wallace and Bogner (2000), nonpatients may have a very incomplete picture of how low one can sink in depression, and many of the symptoms of TBI – anxiety, hostility, distress, etc. – may worsen, not improve, with time and increased awareness.

Another complicating factor, explored in only a few studies, is that former recovered patients sometimes provide lower ratings of quality of life in a given condition than current patients. In a study by Smith *et al.* (2006) of quality of life with colostomy, for example, not only did public representatives provide lower ratings than patients, but those who had their colostomies successfully reversed also provided lower ratings. The finding is notable. Former patients, presumably, are at least as knowledgeable, if not more, about the comparative quality of normal life and life with colostomy. Current patients may not be good judges, repressing or misremembering how good their previous life without the impairment was.

Still, generally, patient values are higher than public ones.

Reasons for Variation: Knowledge and Adaptation

The difference is due to a number of factors. An obvious one is simply that patients directly know life in a particular health state; nonpatients do not. Insofar as nonpatients fear the

condition because it is different – presumably worse than pre-illness, 'normal' life, and in any case relatively unknown – they rate their prospective quality of life low. Even if actual and hypothetical patients were equally knowledgeable about all the facts about a condition, however, and hypothetical patients were no more fearful, their ratings would still likely differ because patients adapt to their condition. Adaptation has thus attracted considerable attention in accounting for higher health state utility ratings by patients.

It is a very broad phenomenon, undoubtedly comprised of many different elements. Although there is no complete agreement on the elements of adaptation, Menzel *et al.* (2002) described eight components:

1. Skill enhancement: People develop skills they previously did not have or had not developed as much. A person with paraplegia, for example, becomes very accomplished in maneuvering a wheelchair.
2. Adjusted choice of activities: Given limitations that make previous activities difficult, one develops new interests. A person with congestive heart failure, for example, gives up the gardening that was her previously quite physically demanding hobby and devotes time to watercolor painting.
3. Revision of substantive goals: Not just particular activities but fundamental goals in life are revised. Instead of ambitious career success, for example, a person may shift his/her most life-defining goals to esthetic appreciation and personal relationships.
4. Heightened stoicism: A person becomes more patient, taking events in life as less within his/her control.
5. Lowered expectations: Without significantly changing activities or fundamental life goals, and in addition to becoming more patient, one does not expect to operate or perform at the same level.
6. Altered conception of health: A person who previously thought of paraplegia as diminished health now looks on it as a limitation not essentially different than previous and continuing limitations (e.g., not being able to run the high hurdles); he/she retains a vigorous conception of health, but one that now does not include some previous physical capacities.
7. Suppressed recognition of full health: One loses sight of how someone can be as healthy as one was before.
8. Cognitive denial of a lowered health state: One refuses to acknowledge that one's health has diminished, not because one has adopted a revised conception of health but by ignoring the pain or limitation. A variant of cognitive denial is focusing illusions: People focus more readily on new things they can do than on the things that they can no longer do.

Several of these elements – lowered expectations, heightened stoicism, revision of substantive goals, and altered conception of health – were included but categorized differently by Schwartz and Sprangers (1999) in a well-known comprehensive framework for analyzing adaptation known as 'response shift.' In the response shift typical of adaptation, persons living for a considerable length of time in an altered state of health make three changes: in their internal standards of measurement ('recalibration'), in their values ('reprioritization'), and in their definitions of essential constructs such as health ('reconceptualization').

Different contexts for adaptation may lead to wide variations in the proportionate influence of these different elements. There is, in any case, little agreement on their relative influence. Discerning the different elements of adaptation and their relative roles may be important to any normative assessment of the proper place of adapted patients' values in CUA and CEA. If adaptation is dominantly constituted, for example, by skill enhancement, adjusted choice of activities, and revision of substantive goals, it will tend to be accorded greater respect. In contrast, suppression and cognitive denial diminish respect for adaptation. In a later section of this entry, Normative Considerations in Choosing an Approach, the normative issues about adaptation will be pursued.

Conceptual Issues in Choosing an Approach

Numerous issues arise in choosing between public and patient values. The two involve distinctions between different concepts and types of value. Articulating these distinctions clearly lays an important background for more explicitly normative aspects of the discussion.

Individual Utility and Social Value

In approaching the choice of whose values to use in CEA, it is important to understand the role that individual utility plays more generally in the framework for health state evaluation in health economics. If individual utility's role is not dominant but more limited, the question of whose values to employ can be pursued with an awareness that other dimensions of the value of health may be available to resolve certain quandaries.

Individual utilities of any sort, including health state utilities, contrast with social values for priority setting (also referred to as 'societal values'). The former concerns individual well being, whereas the latter relationship between persons or the well being of communities, including considerations of fairness. The difference between individual utility and social value lies not in who expresses or holds the values but in the individual versus interpersonal nature of object to which value is attributed. Both individual utility ratings and social values get expressed by individuals.

Within the influential 'welfarist' utilitarian tradition in health economics, people can easily lose sight of the difference between individual utility and social value. In the welfarist view, only individual utilities are needed to build judgments about social value; the highest social value just is maximum aggregate individual utility. In the case of health care procedures and programs specifically, not only CUA but also conventional CEA are conducted by measuring and aggregating the individual health state utility gains and losses ingredient in the outcomes. In this view, procedures and programs producing the greatest aggregate net health state utility have, *ipso facto*, the greatest social value.

The point here is not to defend or reject such a welfarist utilitarian position in normative economics or social philosophy. The point is to be aware that other options are available once the distinction between individual utility and social value, as categories of value, is recognized. Social values

do not have to follow welfarist utilitarianism. The paraplegia example is again illustrative. Person trade-off (PTO) questions can reveal equal social value in saving the lives of persons with and persons without paraplegia, even when both rate the individual utility of life with paraplegia as less than 1.0. It would appear, then, that a sophisticated model for discerning the value of health will need to account for the distinction between social value and individual utility. PTO questions can be used to elicit social values, but using them does not rule out the use of visual analog scale (VAS), TTO, or standard gamble (SG) questions to elicit ratings of health state utility.

Answers to the question of whose values to use may thus be different for individual utility and social value ratings. One view on the proper overall structure for health valuation, proposed by Nord *et al.* (1999) divides its answer this way: (1) Patients should be asked questions to obtain quality of life ratings, (2) public representatives should be informed thoroughly of those patient ratings, and then (3) public representatives should be asked PTO questions to obtain social values. Such an architecture for eliciting values illustrates the possibility that the question of whose values needs answers on two different levels: whose judgments to use in discerning health state utility and in discerning social value.

Decision Utility versus Hedonic Experience

Heretofore, the authors have been dealing with 'experience utility' as the quality of life in a state rated by people who have experienced that state themselves. A different, second sense of the term used in economics needs to be acknowledged. Experience utility in this other sense refers to the direct, intrinsic hedonic experience of an outcome. The opposite of experience utility in this sense is not utility estimated by people who have not directly experienced the condition being evaluated, but what economists and psychologists call 'decision utility' – utility measured by people's choices about an outcome. Those choices can be manifested in either actual behavior or expressed preference. For health states, decision utility is measured by TTO or SG preferences about that state, whereas experience utility is measured by direct estimates people make about their hedonic level in a given state. Over the centuries, such direct hedonic experience utility has often been referred to as satisfaction/dissatisfaction, pain/pleasure, and happiness/unhappiness.

If direct hedonic experience judgments are used to rate health states, they will presumably be made by patients in those states. Decision utility ratings, though, can be obtained by asking either actual patients or people imagining themselves as patients. Understandably, then, empirical studies of the difference between patient and nonpatient valuations usually focus on decision utility discerned through TTO or SG. Rarely have the direct hedonic experience utility levels of patients and healthy nonpatients been compared.

With direct hedonic experience ratings of health states thus being made by patients in those states, whereas decision utility ratings can be made by either actual or hypothetical patients, the argument between direct hedonic experience utility and decision utility has implications for the debate about whose values should be used in CEA. If direct hedonic experience utility wins and decision utility loses out, patient values need to be used. However, if decision utility wins, whose values to use remains an open question.

For which of these kinds of utility, decision utility or direct hedonic experience, can the stronger case be made? Decision utility suffers from three significant disadvantages: (1) Because it takes choices as basic in discerning and measuring utility, no independently discerned utility is available with which those choices can be assessed. By contrast, "if we equate welfare with [direct hedonic] experience utility…, it should…be possible to assess whether people's choices actually maximize their own (experience) utility." Such assessment is often seen as necessary because, as noted by Loewenstein and Ubel (2008), "behavioral economists have identified myriad ways in which people take actions…patently contrary to their own interests". (2) Beyond deficiencies in serving people's own interests, moreover, there is abundant evidence that choices and preferences routinely manifest various kinds of inconsistency and irrationality. This poses a challenge: Is it morally defensible to be guided in policy decisions by utilities discerned by preferences that are so frequently flawed? (3) In a number of studies where the SG and/or the TTO have been used to get self ratings from patients, a large share of the subjects have been unwilling to sacrifice any life expectancy in order to become well – even when symptoms and dysfunctions have been quite severe. As noted by Nord *et al.* (2009), this is not necessarily because the health problems are without consequences for well being, but because life itself is so highly valued that it takes quite large health gains to justify any sacrifice of length of life. The 'nontrading' subjects automatically receive utility scores of 1.0, which does not seem helpful in evaluations of programs that clearly have value in terms of improving health and health-related quality of life. To obtain usable (policy relevant) values for health states from patients, one may therefore have to have recourse to measures of happiness, etc., rather than decision utility tools like the SG and the TTO.

But the attempt to discern utility independent of choice and preference may not inspire any greater confidence. To be sure, processes are available through which people can rank their state of well being directly, rather than through expressing some preference like TTO or SG, for example, the 'experience sampling method' (ESM) proposed by Stone *et al.* (1999) and the 'day reconstruction method' (DRM) described by Kahneman *et al.* (2004). In ESM, electronic devices are used to ask people at random times during the day how they rate themselves on certain feelings at the time (happiness, frustration, etc.). In DRM, people are asked to divide the previous day into episodes and rate them for affective elements on a specific scale. However, as noted by Dolan and Kahneman (2008), these methods have their own problems: underestimating losses, misremembering, failing to attend accurately to a given moment because of distraction by other episodes during a day, etc. It is not clear that measurement of direct experience utility is any more accurate and reliable than measurement of decision utility through preference elicitation.

Moreover, playing a role in the larger argument is not just the fact that direct hedonic experience utility has its own difficulties as serious as those of decision utility. Decision utility has its own distinctly positive attractions.

First, Decision utility/preference questions are comparatively clear. Direct hedonic utility, well being, happiness, and satisfaction, arguably, have a befuddling breadth, abstractness, and ambiguity. Asked to rate one's life/day/moment directly in terms of them, a person may wonder exactly what they are. "Am I really happy?" "In relation to what desires am I satisfied?" "What is my well being? – there are a lot of candidates!" Almost any preference question for decision utility is clearer. "How much of your lifetime would you be willing to sacrifice to get a cure for your condition?" may be difficult to answer, but its meaning seems clear. Contrast such TTO or SG questions with the ambiguity of a technique like the VAS that asks for a direct ranking of health, analogous to direct hedonic experience utility more generally. "On this bar that extends from 0 (dead) up to 1.0 (full health), rate your health." How is a person supposed to know what a proportion or an amount of health is? Even if the question is more explicitly focused on value, not health *per se* (as in "On this bar from 0 to 1.0, rate your health-related quality of life"), ambiguity persists: what would living with, say, half the quality of life that living in full health contains mean (e.g., as distinct from two-thirds)? By comparison the meaning of any particular amount of decision utility in health-related quality of life seems clear: one will trade-off a certain portion of remaining life but not more to gain a cure, or one is willing to take a certain risk of death but not more.

Second, within the enterprise of CEA in health care, determining quality of life by eliciting preferences has another major advantage. The core function within CEA of a construct like the QALY is to serve as a common unit of value that incorporates both life extension and quality-of-life enhancement. Otherwise one is comparing apples and oranges and would not know what total value an array of diverse lifesaving and quality enhancing outcomes had. TTO and SG questions are appropriate precisely because they transparently involve a relationship between quality enhancement and longevity. The mystery about how the values of life extension and quality enhancement compare is removed by the very nature of the question(s) used to measure health state utility. Measurement methods for direct experience utility – VAS, DRM, and ESM – leave the relationship murky.

If decision utility has thus not lost in its argument with direct hedonic experience utility, the question in health economics of whose values should be used in CEA remains open and vital.

Normative Considerations in Choosing an Approach

As a normative question, whose values for health states to use in CEA may seem naturally weighted toward patient preferences. Patients presumably know more about living in a given health state. Because the ultimate nature of utility is subjective well being, they would seem to hold a privileged position in health state valuation even if they are no better informed of objective facts about their condition. As shall be seen, however, this initial intuitive case for using patient values faces a number of difficulties. One can begin with what is arguably the standard defense of the relatively common practice of using public, not patient values.

The Standard Defense of Hypothetical Patient Utilities

The standard defense has two main points: the societal role of CEA, particularly its role in a democratic society, and practical feasibility.

In CEA, the health state utilities at stake in a medical practice/policy decision are aggregated to generate a picture of the overall value of expected health changes. The perspective of CEA as an enterprise, then, is necessarily societal, not merely individual. CEA is 'for' society (or some subgroup, such as a private or regional insurance plan). Arguably, then, the perspective on the value of health changes needs to be as encompassing as possible: everyone in the society (or the insurance pool). When CEA is located in a democratic society, this line of thought is reinforced further by a higher level societal value that everyone's perspective should be represented in any process like CEA whose wide scope affects potentially everyone. The best way of ensuring that every person potentially affected by a CEA is represented is to elicit health state values from the public, few of whom will be actual patients with the conditions they are evaluating.

Supplementing this argument for public values is a further consideration about fairness. In its 1996 report, after recommending that CEA employ a 'societal perspective' for reasons similar to those just elaborated, the 1996 US Public Health Service panel on cost-effectiveness in health and medicine (Gold *et al.*, 1996) proceeded to discuss fairness. Fair decisions are best made as choices behind a veil of ignorance, where decision-making parties do not know whether they are advantaged or disadvantaged by the matter at hand. For CEA, then, they claimed that "aggregating the utilities of persons who have no vested interest in particular health states seems most appropriate." If we have already agreed that whatever values are used, they should come from perspectives that are informed, rational and unbiased, a challenge is presented for patient values: patients are likely to be biased by having a vested interest in treating the disease they know they have. One skirts this bias by asking people who express themselves only as hypothetical patients.

Practical considerations also play a role in the standard defense of hypothetical patient utilities. If utility ratings are elicited from hypothetical patients, convenience and efficiency is gained by eliciting ratings for many health states from the same people at the same time. This advantage is difficult to dispute factually. Any procedure for eliciting utilities directly from patients will undoubtedly be more cumbersome and expensive, especially if CEA is used to compare measures and programs across a comprehensive range of health states. The case for using patient values will have to be strong enough to justify additional expense.

Adaptation

Adaptation raises the utility ratings patients express about reduced health-related quality of life. In doing so, it lowers the value achieved from restorative, quality of life improving treatments. In this respect adaptation reduces ill persons' leverage in the competition for health care resources. That alone will make a positive role for adaptation controversial. Critical normative argument about adaptation, however, cuts different ways in the debate between patient and public values.

Epistemic Privilege

A strong first line of argument for using adapted patients' values amplifies the initial, intuitive case for patient values in the first place: because patients presumably know more about living in a given health state and the ultimate nature of utility is subjective well being, they hold a privileged epistemic position for discerning health state utility. To see how attractive this claim can be, suppose, for the moment, that we have adopted the opposite practice and are eliciting health state valuations not from patients but from representatives of the general public. People already agree that the states that public representatives are asked to evaluate need to be described in neutral, factually accurate, and sufficiently complete terms. As part of that, they need to understand what life in fact is really like in the condition they are evaluating. How actual patients typically adapt to a condition is part of that understanding. It is an objectively real aspect of the lives that patients in diminished health states actually lead.

Suppose, in turn, that a hypothetical patient challenges this need to absorb the facts of adaptation and insists that he/she in particular would never evaluate life with paraplegia, for example, at so high an adapted level. The insistence would be suspect. To be sure, a given individual could conceivably be correct in claiming he/she would not adapt (or adapt much), but given the empirical evidence, we ought to be skeptical about what will happen to even such insistent persons when they actually become paraplegic. Very likely most of them, too, would end up adapting considerably. But then, when hypothetical patients refuse to accept these prospective facts, they should be regarded as factually mistaken – they do not understand the state they are evaluating. Thus, we ought to enrich the description of a condition provided to hypothetical patients with information about actual patients' adapted values, and we should insist that hypothetical patients truly absorb those facts. But then why not simply use actual patients' health state values?

Such an argument for the epistemic privilege of actual patients is attractive, but it has not been universally accepted. Brock (1995) argues that because the difference in health-related quality of life ratings stems significantly from an adjustment of substantive goals by adapted patients, they have become 'changed persons'. He concludes that the hypothetical patient's earlier evaluation is not 'mistaken.' If we look back now as chronically ill or disabled, we 'view ourselves as having become very different persons,' not as 'having been mistaken in our earlier aims and values.'

Brock's (1995) move may or may not be a plausible gloss on the meaning of 'mistaken' in relation to 'changed person.' In any case, it is debatable whether it provides a defense for hypothetical patients if they have not absorbed the essential facts about actual adaptation. How can they defend the practice of continuing to evaluate the condition at their own nonadapted level, it will be asked. In most of their prospective years in the chronic condition in question, should they ever experience it, after all, they will likely espouse adapted values. Why should they now be trying to imagine themselves as persons who do not adapt to the condition, people they almost certainly will not be? We find ourselves pushed back to the view that the patient, adapted or not, is in a privileged position in the very enterprise that asking hypothetical patients involves, imagining what it would be like to be someone with the condition.

These considerations establish the initial case for using the values of patients with actual experience of a condition regardless of how much their ratings may be elevated by adaptation. That does not, though, end the moral debate. As noted by Menzel *et al.* (2002), other arguments can be made against the use of adapted values. Among them is the problem of entrenched deprivation.

Entrenched Deprivation

Amartya Sen has focused on this reason to discount adaptation in a well-known critique of utilitarian reasoning generally from 1992. At its very basis, he says, utilitarian ethics is guilty of excessively depending "on what people 'manage to desire' … [that neglects] the claims of those who are too subdued or broken to have the courage to desire much…. A thoroughly deprived person, leading a very reduced life, might not appear to be badly off in terms of the mental metric of desire and its fulfillment, if the hardship is accepted…. In situations of long-standing deprivation, the victims do not go on grieving and lamenting all the time…. The extent of a person's deprivation, then, may not at all show up in the metric of desire fulfillment…." Sen concludes that measuring well being by the fulfillment of people's actual desires is ethically wrongheaded.

Utilitarianism, as a general moral philosophy, must respond to Sen's critique. The issue in the context of whose preferences to use in health state evaluation for CEA, however, is narrower: does the deprivation factor that renders a general utilitarian metric of desire fulfillment ethically questionable also render the ratings of health-related quality of life procured from adapted patients morally dubious? Adaptation in contexts of chronic illness and disability often involves achievement and shrewd and successful control over the trajectory of one's inner life. Here, the adapted person is anything but broken. He/she is hardly subdued. If deprivation is handled by people as challenge and achievement, why is not a metric of actual desire fulfillment appropriate? Sen's argument may serve as an appropriate warning about too readily or generally using adapted patients' utilities, but it is highly problematic as a full rejection.

Social Values and Adaptation

Although the argument from entrenched deprivation against using adaptation influenced values may thus be neutralized, other arguments may be more successful. Suppose that after comprehensive assessment we end up thinking that the basic dilemma posed by adaptation remains unresolved. Here, the distinction between individual utilities and social values detailed previously may provide constructive help. This would be the line of reasoning:

The health state utilities of real life with chronic illness and disability are those expressed by patients. Those are the real utilities of health states, and we should use them in CUA. But people also make moral arguments against the use of

adaptation-influenced values – values that are higher than public values and which therefore reduce the value of the health gained by patients from curative/restorative services. They believe, for reasons of justice or whatever, that those services should be accorded higher value than their real utility value for the disabled and chronically ill alone would indicate. In such beliefs, they are expressing social values. At the level of individual utility itself, there is no 'problem of adaptation' at all; the real utility of moving from illness/disability back to full health is just the utility indicated by patient values. The 'problem of adaptation' occurs only when social value intersects with this utility. Maximizing health state utility is hardly the only philosophical choice for social value, and any influence of adaptation on decision making can be altered at the level of social value. Hypothetical patients need to know that real, adapted patients' health state utility values are higher than their own, and they must absorb those facts in imagining prospective illness. Both hypothetical and actual patients, however, may hold social values that blunt adaptation's effect on social decision making.

The Equal Value of Life

The picture portrayed by the discussion so far is incomplete. We have been wrestling, in part, with the fact that as adaptation increases the value of deficient health states, it decreases the utility gain from restorative, quality of life improving measures. That, however, is only half the effect of adaptation in CEA. In increasing health state utility, adaptation also increases the utility gain from life extension. Insofar as these opposite effects on the utility gain in life-extending versus restorative interventions balance each other out, adaptation may leave the total utility gain claimed for health care relatively unaffected. The matter of whose preferences to use would then be of little consequence.

It would not be correct, however, to conclude that counterbalancing effects rendered the debate about adaptation unimportant. The aggregate effects in the two types of programs – restorative and life extending – may not in fact balance each other out, and specific programs being evaluated will often be largely life extending, others largely quality of life restoring. Most importantly, perhaps, adaptation augments a problem for CEA created by any claim of equal value for different lifesavings. Such equal value for saving the life of a person back to full health and saving the life of a person who will continue in chronic illness or disability was cited as a serious problem for traditional CEA by John Harris already in 1987 and later by Nord (1999). It gives rise to what Ubel *et al.* (2000) called the 'QALY trap.' If the value of life extension for the disabled and chronically ill is equal to the value of life extension for the fully healthy, then the value of curing a chronic illness or disability (restoring such patients to full health) is apparently zero. This implication follows given the very structure of CEA, using as it does a common metric like the QALY to put life extension and quality enhancement on the same value scale. However, if restoration and cure retain value, then the respective life extensions cannot be of equal value. Yet, they do seem to be of equal value. Meanwhile, the first option – no value for cures – also seems contradicted empirically. Virtually everyone, including the disabled and chronically ill, accord considerable value to restorative measures. The traditional QALY model for CEA is then trapped between two propositions – that the different life extensions are of equal value, and that restorative cures have positive value – which the model says cannot both be true.

To save CEA against this challenge, one might take one of two approaches: (1) Use the distinction between social value and individual utility to rescue CEA from the QALY trap. Call this the 'value/utility distinction' approach. Or (2) give up the claim that life extensions for the disabled/chronically ill and the nondisabled/fully healthy have completely equal value but maintain that they have 'almost equal value.' Both approaches affect how we view the issue of whose preferences to use in CEA.

In one particular version of the value/utility distinction approach, Nord *et al.* (1999) suggested that all gained life years should count as one as long as they are deemed preferable to death by those concerned. In this and all other versions of the approach, the claim that the value of life extension for the disabled and chronically ill is equal to the value of life extension for those who can be saved to full health is seen as an expression of a societal value and only societal value. It is not a claim that the individual utilities of the two life extensions are equal. If as a matter of individual utility they are not equal, the claim of positive value for quality restoring measures can be retained. Thus, keeping individual utility and societal value distinct frees CEA from the QALY trap (see Ubel *et al.*, 2000).

A second very different approach backs away from the claim that the two respective life extensions have equal value. If one pays careful attention to the values expressed by patients through, for example, TTO preferences, it is clear that quality of life ratings are not much less than 1.0, but are less than 1.0. In 1993, Dennis Fryback and colleagues reported results on the order of 5–8% for arthritis, severe back pain, migraine, angina, cataracts, ulcers, and other serious conditions and 14–17% for depression, asthma, and chronic bronchitis. This suggests that the tension between 'equal value for lifesaving' and 'cure has value' can be reduced by attending to just how close to a maximum value of 1.0 people with disability or chronic illness rate their quality of life. We can then adjust the claim of 'equal value' to 'almost equal value,' and the claim of 'cure has value' to 'cure has very modest value compared to lifesaving.' Again, we are out of the QALY trap, and now in a way that has not changed the structure of traditional, utility-focused CEA.

In this approach, one still needs to face front and center the question of whose values to use. No easy accommodation of 'use both' (though in respectively different senses of value) is possible, as in the value/utility distinction approach. One must determine whether the higher, adaptation-influenced values of patients that raise the value gained from life-extending measures but lower the value gained from restorative measures ought to be used. The advantage in using the value/utility distinction approach is that one can keep both of the key claims that create the dilemma – equal (social) value for the life extensions, yet significant (individual utility) value for restorative/curative measures (for further discussion, see Nord *et al.*, 2003).

Seeing the claim of equal value for the different life extensions as only a claim of societal value is the key move in the value/utility approach. Arguably, however, the equal value claim may also hold at the level of individual utility. To see this, unpack the reasoning behind claims of value in a QALY framework. In responding to SG or TTO questions, patients are saying that a cure for their paraplegia, for example, has a certain proportion of the value of saving their very own life. In saying that, they have not said that their very life itself has less value than the very life of *another* person in full health. Menzel in 1990 noted that, compared to death they very likely believe that their paraplegic life is as valuable to them as anyone else's allegedly 'better' life is to him or her. This belief comes sharply into focus when people attend to the two comparisons involved: the value of their disabled or chronically ill life relative to death, and its value relative to the same sort of death-comparative value of another's life, even a person in full health. Particularly compelling may be a further step: this very realization of equal individual utility value will likely be shared by healthy and nondisabled persons, too, once they think reflectively about the value of their own very lives compared to death. Who among them would want to claim that the value to them of their life is greater than the value of a paraplegic person's own life to him or her?

The conundrum of the QALY trap thus continues. If lifesavings have equal value even as a judgment about individual utility, we face again the trap's full dilemma: the value of curing paraplegia has become nothing. It is doubtful we would ever accept that. The implication of our resistance is that the QALY trap remains a challenge to the very framework of CEA, an enterprise whose current form requires a common unit of benefit like the QALY in which value for life extension and value for quality improvement are integrated on the same scale. Whether and how the field of CEA will meet this challenge remains unclear.

Evaluating Prevention

Is anything different about 'whose values?' when preventive services, not treatments, are evaluated? Actual patients, arguably, hold legitimate evaluative privilege in rating health state utility. They are the only subjects who experience real life with chronic illness or disability, and hypothetical patients who are imagining themselves to be in such conditions must account for the reality of likely adaptation. The situation is arguably different in prevention. The real recipient of a preventive service's benefit is the unchanged person who, if the prevention is effective, will never need to adapt to the illness or disability in question. It may be suggested, therefore, that while actual patients should be accorded evaluative privilege for purposes of prioritizing curative and restorative services, hypothetical patients retain evaluative privilege for the utility ratings that help determine the value of preventive services.

This suggestion immediately encounters an objection. The person whose health will be damaged if preventive measures are not provided is still a person likely to adapt. The real value of prevention is presumably the difference between people's quality of life before disease and their quality of life with disease or disability – that is, after likely adaptation to the conditions, which have not been prevented. The perceived gains from prevention may be higher than the gains from cure, but the real gains are not.

This is only one interpretation of 'real gains,' however. Most persons receiving a preventive service do not contract the condition the service aims at preventing, a fact that is true regardless of how objectively effective the preventive measure actually is. Recipients of prevention continue to experience its benefits from their perspective as healthy persons. Restorative services, by contrast, are received by persons already experiencing both the burdens of illness and disability and the value raising effects of adaptation. Many lifesaving services also apply to persons already experiencing the ravages of illness. One might argue, therefore, that while the utility value of health gain in the case of treatment services should be determined by patient ratings, the value of the avoidance of illness achieved by prevention should be measured from the perspective of hypothetical patients. Although adaptation lowers the value of restorative services, perhaps it should not be allowed a similar effect on the value of prevention.

Careful analysis of such considerations in comparing how the value of health benefits should be measured for treatment as compared to prevention has received little attention in the literature. An exception is a paper by Nord *et al.* (2009), in which the authors side with using the lower public, nonpatient, ex ante values to measure what then becomes a higher benefit from prevention, while at the same time using the higher health state utility values expressed by patients to calculate what then becomes a lower value to benefit from treatment. They claim there is no inconsistency; with different reference points, the negative value of ill health just is not the same for these two different parties with their different perspectives. Another analysis, by Menzel (2012), of the relative value of prevention also emphasizes reference point differences but leans in the opposite direction, a lower value for prevention.

Summary and Conclusion

Utilities for health states can be measured by values elicited either from people who have experienced those states themselves or by hypothetical patients imagining themselves to have such conditions. Ratings by actual patients are generally higher than ratings by hypothetical patients, rendering the question of whom to ask to measure health state utility for CEA of practical importance.

Factors that help to explain patients' higher ratings include their greater knowledge of the conditions and their adaptation, especially to chronic disease and disability. Adaptation is comprised of numerous different elements, the proportionate influence of which remains unclear.

Two conceptual distinctions affect positions taken on whom to ask. By distinguishing rigorously between individual utility and social value, one proposal argues for asking both patients and the general public but differently: elicit utility ratings from patients, inform public representatives of those ratings, and then elicit social values from the public representatives. Also affecting the debate is a distinction between two different senses of utility, direct hedonic experience and

decision utility expressed through preference or choice. The latter, measured in health utility analysis by choices between preserving life and improving quality of life, has the advantage of yielding a common metric for measuring the value of all changes in health. With decision utility, unlike direct hedonic experience, the debate about whom to ask is kept open; either patient or hypothetical patient values can be sought.

Normatively, the initial intuitive case for patient values sees patients as having epistemic privilege in understanding what real life with disease or disability is actually like. On the opposite side, the standard case for public values cites both the societal perspective that is seemingly natural to CEA – and especially important in a democracy – and practical considerations of convenience and efficiency. The phenomenon of patient adaptation gives rise to numerous and conflicting moral arguments. One attempt to resolve the normative debate would elicit utility ratings from adapted patients but allow societal values elicited from others to discount adaptation's influence on decision making.

The generally higher utility ratings of patients who have adapted to a diminished health state reduce the value gained from curative/restorative services, but those higher ratings have an opposite effect as well: raising the value of life extension for the chronically ill and disabled. Strong arguments are available to defend equal value for different lifesavings – life extension for the disabled/chronically ill and life extension for people returning to full health. Equal value for these lifesavings, while it can be celebrated as a removal of discrimination against the disabled and chronically ill, poses a difficult challenge for the very structure of CEA. A 'QALY trap' emerges: with lifesavings held to be of equal value, restorative care loses its value. It is unclear whether carefully distinguishing between social value and individual utility enables CEA to handle this challenge; if the individual utility, not only social value, of such lifesavings is also equal, the challenge remains unsolved.

Even if the utility values expressed by patients affected by adaptation are the appropriate ones to use for evaluating treatment programs, objections have been made to using them to evaluate preventive programs. Attempts to resolve whether the treatment/prevention difference should affect whose values to use have recently been made; their success is far from clear.

Most of the moral questions about whose values for health utility should be used in CEA are well clarified. Some, perhaps, are even answered, but many are not. Vigorous debate is likely to continue.

See also: Cost–Value Analysis. Quality-Adjusted Life-Years

References

Arnold, D., Girling, A., Stevens, A. and Litford, R. (2009). Comparison of direct and indirect methods of estimating health state utilities for resource allocation: Review and empirical analysis. *British Medical Journal* **339**, b2688, doi:10.1136/bmj.b2688.

Brock, D. W. (1995). Justice and the ADA: Does prioritizing and rationing health care discriminate against the disabled? *Social Philosophy and Policy* **12**, 159–185.

Damschroder, L. J., Zilmund-Fischer, B. J. and Ubel, P. A. (2005). The impact of considering adaptation in health state evaluation. *Social Science and Medicine* **61**, 267–277.

de Wit, G. A., Busschback, J. J. V. and de Charro, F. (2000). Sensitivity and perspective in the valuation of health status: whose values count? *Health Economics* **9**, 109–126.

Dolan, P. and Kahneman, D. (2008). Interpretations of utility and their implications for the valuation of health. *Economic Journal* **118**, 215–234.

Gold, M., Patrick, D., Torrance, G., et al. (1996). Whose preferences should be used in CEA? In Gold, M., Siegel, J., Russell, L. and Weinstein, M. (eds.) *Cost-effectiveness in health and medicine*, pp. 82–133. New York, NY: Oxford University Press.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N. and Stone, A. A. (2004). Toward national well-being accounts. *American Economic Review* **94**(2), 429–434.

Loewenstein, G. and Ubel, P. A. (2008). Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics* **92**, 1795–1810.

Menzel, P. T. (1990). *Strong medicine: The ethical rationing of health care*. New York, NY: Oxford University Press.

Menzel, P. T. (2012). The variable value of life and fairness to the already ill: Two promising but tenuous arguments for treatment's priority. In Faust, H. S. and Menzel, P. T. (eds.) *Prevention vs. treatment: What's the right balance?*, pp. 194–218. New York, NY: Oxford University Press.

Menzel, P. T., Dolan, P., Richardson, J. and Olsen, J. A. (2002). The role of adaptation to disability and disease in health state valuation: A preliminary analysis. *Social Science and Medicine* **55**, 2149–2158.

Nord, E. (1999). *Cost-value analysis in health care: Making sense out of QALYs*. Cambridge, UK: Cambridge University Press.

Nord, E., Daniels, N. and Kamlet, M. (2009). QALYs: Some challenges. *Value in Health* **12**(supplement 1), S10–S15.

Nord, E., Menzel, P. and Richardson, J. (2003). The value of life: Individual preferences and social choice. A comment to Magnus Johannesson. *Health Economics* **12**, 873–877.

Nord, E., Pinto Prades, J. L., Richardson, J., Menzel, P. and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programs. *Health Economics* **8**, 25–39.

Riis, J., Loewenstein, G., Baron, L., et al. (2005). Ignorance of hedonic adaptation to hemodialysis: a study using ecological momentary assessment. *Journal of Experimental Psychology* **134**(1), 3–9.

Schwartz, C. E. and Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal quality of life research. *Social Science and Medicine* **48**, 1531–1548.

Smith, D. M., Sherriff, R. L., Damschroder, L., Loewenstein, G. and Ubel, P. A. (2006). Misremembering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychology* **25**(6), 688–695.

Stone, A. A., Shiffman, S. S. and De Vries, M. W. (1999). Ecological momentary analysis. In Kahneman, D., Diener, E. and Schwarz, N. (eds.) *Well-being: The foundations of hedonic psychology*, pp. 26–39. New York, NY: Russell Sage Foundation.

Ubel, P., Nord, E., Gold, M., et al. (2000). Improving value measurement in cost-effectiveness analysis. *Medical Care* **38**(9), 892–901.

Wallace, C. A. and Bogner, J. (2000). Awareness of deficits: Emotional implications for persons with brain injury and their significant others. *Brain Injury* **14**(6), 549–562.

Vaccine Economics

S McElligott, University of Pennsylvania, Philadelphia, PA, USA
ER Berndt, Massachusetts Institute of Technology, Cambridge, MA, USA

Introduction

Vaccines have been historically hailed as one of the great global public health success stories. Major infectious diseases with significant morbidity and mortality are largely under control in many parts of the world, or have been completely eradicated, for example, smallpox. The World Health Organization (WHO), the United Nations Children's Fund (UNICEF), and the World Bank estimate that 3 million lives are saved worldwide each year through childhood immunization. In the US, routine immunization has been estimated to prevent 10.5 million cases of infection and 33 000 deaths each year. It has been projected that in the US, diphtheria-tetanus-pertussis (DTP); diphtheria toxoids (Td); *Haemophilus influenzae* type b (Hib); inactivated polio (IPV); measles, mumps, and rubella (MMR); and hepatitis B (HepB) vaccines were responsible for saving US$42.5 billion dollars per year from a societal perspective in 2001.

Despite these successes there have been historic signs of market failure. An estimated 20% of the global birth cohort, 24 million children, remains unvaccinated because they live in poor countries that lack the resources to invest in vaccines and infrastructure. More than 2 million vaccine-preventable deaths still occur annually. Even in highly developed countries, there continue to be outbreaks of diseases due to declining immunization rates. For example, cases of whooping cough or pertussis are rising in the US and the UK because of low DTP immunization rates. The vaccine market also experiences episodic supply shortages in both high-income and low-income countries. For example, in the US, many vaccines have only one or two suppliers, and manufacturing disruptions or regulatory actions can lead to insufficient supply. UNICEF has historically experienced periods of potential supply shortages. Several manufacturers have exited the pediatric vaccine market in both developed and developing countries, and concerns have been raised that vaccines may be undervalued by payers, which may lead to underinvestment in vaccine research and development (R&D).

This article examines vaccines from an economic perspective, highlighting the distinct features of the vaccine market that affect the demand, supply, and market outcomes for vaccines. The article is organized as follows. The first section provides an institutional background on vaccines and the vaccine market. The second section discusses the demand for vaccines. The third section examines supply and market outcomes. The fourth section summarizes key points and provides concluding remarks. The analysis and discussion here focuses on preventative vaccines. The emerging area of therapeutic vaccines, although interesting, is outside the scope of this discussion.

Background

The vaccine market can be roughly broken down by consumer age: pediatric, adolescent, adult, and elderly. Pediatric vaccines have historically represented the largest segment. However, new vaccines are spurring growth in both adult and adolescent markets. Globally the recommended vaccines vary by country and are set based on vaccine availability, available funding for immunization programs, and differing perceptions of risk/reward trade-offs. For example, most developed countries use an IPV because of its lower risk profile. However, in low-income countries it is more common to use the oral polio vaccine because it is less expensive and easier to administer, even though there is a risk of vaccine-associated paralytic poliomyelitis.

In the US, the Advisory Committee on Immunization Practices (ACIP) currently recommends that children be vaccinated against approximately 15 different diseases. For adolescents, the ACIP recommends that young adults be vaccinated against five diseases, including yearly influenza. The adult and elderly markets primarily revolve around annual influenza vaccination, booster shots against tetanus, diphtheria, and pertussis (Tdap), and herpes zoster vaccines for the elderly. The WHO recommends vaccinating children for 11 different diseases, and approximately three vaccines for adolescents and adults if they are in a high-risk group or previously not immunized.

Consumers rarely pay directly for vaccines, with the exception of influenza vaccines. In developed countries, physicians and pharmacists are the primary private purchasers of vaccines, thereby assuring proper handling of heat labile vaccines. Physicians, pharmacist, or other healthcare workers typically administer vaccines and then are reimbursed for the cost of the vaccine and its administration by private or public third-party payers. Physicians can theoretically profit from both the vaccine and its administration. In the US, estimates vary on the degree to which physicians profit or take a loss from vaccinations. A significant amount of public purchasing occurs for vaccines in all countries, including the US, where 55% of the volume is purchased by federal or state governments. For less developed countries, UNICEF, the single largest purchaser of vaccines on a volume basis, purchases 40% of all vaccines produced globally. The financing for these vaccines is through the United Nations and the GAVI Alliance (formerly the Global Alliance for Vaccines and Immunization). Required copayments or user fees for vaccines are generally discouraged by the WHO.

In the US there are approximately 43 different types of vaccines protecting against 19 diseases, supplied mainly by six large multinational pharmaceutical companies. These manufacturers are responsible for the supply of a significant amount of global demand to high-income countries. For the low-income markets, there are a total of 30 manufacturers that are WHO prequalified, meaning that they can supply vaccines to UNICEF or any other country that accepts WHO prequalification. However, at the antigen level there are very few to one supplier in most markets (**Table 1**). For example in

Table 1 Number of suppliers by pediatric vaccine in 2012 for UNICEF and the US

Vaccines	*UN*	*USA*
Bacillus Calmette-Guerin (tuberculosis)	5	n/a
Diphtheria-tetanus toxoid	2	n/a
Diphtheria-pertussis-tetanus	3	2
Measles	3	n/a
Meningitis	3	2
Measles-mumps-rubella	3	1
Measles-rubella	1	n/a
Oral polio	4	n/a
Pneumococcal conjugate	2	1
Rotavirus	2	2
Tetanus	3	3
Tetanus Toxoid	4	n/a
Yellow Fever	2	n/a
Hepatitis B and Hepatitis B combinations	4	2
Haemophilus influenzae type b and *Haemophilus influenzae* type b combinations	4	2
Human papillomavirus	n/a	2
Varicella	n/a	1

Abbreviation: UNICEF, United Nations Children's Fund.
Source: For US figures, Centers for Disease Control and Prevention, http://www.cdc.gov/vaccines/programs/vfc/awardees/vaccine-management/price-list/index.html and for UNICEF figures, the UNICEF supply division, http://www.unicef.org/supply/index_57476.html

2012 in the US, most vaccines were supplied by only one or two suppliers. Additionally, UNICEF relies on anywhere from one to five suppliers for any given type of product. There has been some recent entry of new products into the vaccine market, due to the introduction of newer higher priced vaccines such as the human papilloma vaccines (HPV) that protect against cervical cancer, smaller firms focusing on the biodefense market, and entry into low-income markets.

Product Characteristics

The vaccine market differs from the traditional pharmaceutical market for a number of reasons. These differences in product characteristics have consequences for demand, manufacturing, research, and the role of the government.

Vaccines are largely preventative in their aim and are typically consumed only occasionally in a person's lifetime. For example, most vaccines are given in childhood and many confer lifelong immunity or require a booster dose every 10 years. The preventative nature of vaccines means that the consumption decision is based on preventing a future adverse health outcome, whereas many pharmaceuticals and biologics are targeted on treatment, so the consumption decision occurs only when the consumer is ill. Because vaccines are consumed only a few times in a person's lifetime, their demand is concentrated and largely limited by the size of the birth cohort. This contrasts in particular with pharmaceuticals for chronic conditions that may need to be used regularly on a daily basis for a patient's entire remaining lifetime.

Another important distinction is that vaccination may provide benefits to those that are not vaccinated, a positive externality. If a person chooses to be vaccinated, that lowers the probability that an unprotected person will become ill. At a certain point, but below 100% immunization rate, it does not make sense to continue vaccinating because the population has reached the so-called herd immunity rate. The proportion of persons that are immune to the disease from vaccination is high enough that the probability of a disease outbreak is effectively zero. For example, the herd immunity for polio is achieved at an immunization rate of approximately 80%: the remaining 20% of the unvaccinated population is still protected. This herd immunity is important for the protection of those who cannot be immunized, such as the very young or those with a compromised immune system; however, the positive externality from one person's vaccination decision and herd immunity can also lead to a free-rider problem. The existence of the free-rider problem is commonly cited as a justification for government intervention in the vaccine market through mandates and subsidies.

On the supply side, the research, development, and production characteristics of vaccines are distinct from small molecule chemical pharmaceuticals. These differences influence a firm's decision to enter or exit the vaccine market. For example, vaccine clinical trials are typically larger than pharmaceutical trials (Table 2). However, there is some evidence to indicate that vaccine development costs are about the same as that of chemical pharmaceuticals and vaccines have historically had a slightly higher probability of success once they have entered Phase I trials. Additionally, vaccines, as biologics, entail a production process and product characteristics that are intimately linked, necessitating that the final phase of clinical development, Phase III trials, is for both the product and the manufacturing process. This means that investments in full-scale manufacturing need to be made before product approval. Thereafter, the biologic nature and the inability to fully characterize vaccines necessitate greater regulation after product approval. For example, changes in the vaccine production process may require repeating clinical trials to demonstrate that the production changes did not affect product quality. However, changes in chemical pharmaceuticals production process can be demonstrated to not have affected product quality through bioequivalence testing. Finally, defining the regulatory pathway to entry following patent expiry becomes more complex. In the US, for example, generic pharmaceutical companies need to demonstrate 80–125% bioequivalence to receive marketing approval for their drug. Although vaccines are biologics, they may be too complex to currently demonstrate biosimilarity and therefore follow-on entry will likely be more difficult. For example, in their guidance on biosimilars, the European Medicines Agency has stated: "Vaccines are complex biological medicinal products. Consequently, vaccines have to be considered on a case-by-case basis."

Demand

Consumer demand for vaccines is influenced by price and a number of other nonprice factors; these have been studied

Table 2 Clinical trial enrollment by phase for vaccines and nonvaccines (1999–2011)

	Nonvaccines			*Vaccines*		
	N	*Mean*	*Standard deviation*	*N*	*Mean*	*Standard deviation*
Phase I	6 740	62.7	456.7	343	59.4	74.2
Phase I/II	2 544	80.8	358.3	108	108.1	146.6
Phase II	11 024	154.5	3818.9	269	317.8	659.4
Phase II/III	997	336.0	1700.0	7	1 552.0	1 308.5
Phase III	6 449	706.2	2103.9	114	2 830.2	5 010.3
Phase IV	4 201	330.9	2103.9	58	10 535.6	34 762.6

Source: Data from http://ClinicalTrials.gov. To investigate concerns that nonvaccines trials may have disproportionately smaller trials due to orphan drug or specialty drug trials, the mean enrollments for vaccine and nonvaccine trials were also examined for trials with enrollment >100 people and >200 people. Vaccine trials were still larger in these additional analyses.

extensively in the epidemiological and economic literature. Examples of nonprice demand factors are herd immunity thresholds, population heterogeneity around the benefits and risk of vaccines, how members in a population come in contact with one another, the influence of incentives, and the availability of other disease avoidance behaviors or treatments. Finally, policy makers actively try to influence demand for vaccines by heavily subsidizing them or making vaccination mandatory.

In general, demand for vaccines has been modeled in the economic literature using expected utility theory with consumers comparing trade-offs and choosing to remain unvaccinated or becoming vaccinated. Inherent in this stylized model are a number of important factors. First, in the unvaccinated state, the consumer is healthy and has a certain level of utility. The person is susceptible to disease and will become ill and enter a lower utility state with a given probability. The risk of getting the disease increases with the number of people in the population with the disease (prevalence), the ease at which the disease can infect a disease-free person (transmission and infectivity), and the number of exposures a disease-free person has to the disease (mixing). The utility loss between the healthy and sick states increases with increasing disease virulence, with the greatest loss coming from diseases that have high mortality rates or a significant morbidity burden. For example, before eradication, smallpox killed an estimated 400 000 people annually with an estimated mortality rate of 30%, and of those that survived, 30% would suffer blindness.

In this literature the consumer is depicted as comparing the expected utility from remaining unvaccinated with the expected utility derived from vaccination. Choosing to vaccinate will reduce the probability of getting sick and therefore decreases the expected utility loss from the disease, but there are costs to vaccinating. The direct costs are in the form of payments made by the individual for the vaccine and the administration of the vaccine. In actuality, these are typically small in that governments typically subsidize vaccines or require private insurance programs to provide first-dollar coverage for vaccines. In the US, for example, the Vaccines for Children Program provides free vaccinations to all children below a certain income. Internationally, the degree of coverage varies depending on the health system, but typically the direct cost to vaccinate is subsidized to some degree, especially for pediatric vaccines. But the cost of vaccination also includes the actual and perceived side effects of vaccinating. As with any medication, vaccines carry some risk of side effects such as fever and arm soreness, but in the vast majority of cases, these side effects are relatively minor. However, the perceived risks of vaccination are nontrivial in certain groups of consumers who have expressed a strong dislike toward vaccines and may overemphasize the risks associated with vaccination. Understanding consumer reticence for vaccination is important as this will ultimately be useful for designing policy interventions to achieve socially optimal vaccination rates.

Differences in attitudes toward vaccine and vaccine risk suggest that if a subset of the population views vaccines as more risky than the majority of the population, vaccination rates and attitudes toward vaccination can have a wide variance, and demand may oscillate. Vaccination rates will also be below the socially optimal level if the perception of the risk of disease is lower than the actual risk. For instance, consumers have expressed reticence toward getting the varicella vaccine to protect against chickenpox because of the perception that the impact of the disease is minor.

Another potentially important factor influencing vaccine demand is the prevalence elasticity of demand. Put simply, as the prevalence of a disease decreases in a population, the demand for the vaccine will decrease because the probability of becoming ill falls. This has important implications because a stable equilibrium may not be possible, and instead there can be dynamic oscillations in the disease prevalence and vaccination rates. Therefore, static policies that aim at increasing vaccination rates may not have the intended effects. Additionally, during disease outbreaks and product shortages, the prevalence of the disease will rise quickly and may cause people to vaccinate in self-interest. This may not be optimal because there may be some in the population who would derive greater benefits from the limited vaccines. For example during influenza outbreaks, immunization of the elderly and children are typically prioritized because the elderly can have the most severe complications from infection and children are the primary transmission vectors. However, a healthy adult may not account for the higher-than-average marginal benefit of vaccinating these groups and may choose to try to vaccinate based on maximizing their own expected utility. Evidence suggests that during periods of vaccine shortages, providers and consumers do not allocate scarce vaccines optimally.

Free-riding and herd immunity are other demand influencers. A person may decide not to vaccinate if many other

people have chosen to vaccinate because of the benefit from having a lower probability of becoming ill. Additionally, population-level protection can be achieved at the herd immunity level, which varies based on disease characteristics but is approximately 90% for most diseases. Therefore, the marginal net private and social benefit to vaccinating above the herd immunity point are strictly less than zero if there is any cost associated with vaccination.

The demand for vaccine is also influenced by a number of other factors such as how people share information about vaccination and a disease, a person's experience after a health shock, and the social influence of peers. Models of these factors predict that vaccination rates will oscillate over time.

Some authors suggest that moral hazard from vaccination may actually also increase infections. Vaccination provides a type of insurance against getting a disease and therefore a person that is vaccinated may engage in other high-risk behaviors that could increase their chance of getting a disease especially if the vaccine is not perfectly effective. For example, the introduction of a vaccine for Human immunodeficiency virus (HIV) may actually increase the incidence of HIV or other sexually transmitted diseases if the vaccine induces other high-risk behaviors.

Policy Demand Modifiers

Policy makers have a number of tools that have traditionally been utilized to try to influence vaccination rates. These vary by country, but in the developed world the use of mandates and subsidies is common. Mandates raise the cost of not vaccinating, whereas subsidies decrease the cost of vaccination. In the US, each state determines its vaccination policies with regard to mandates and subsidies. The states also vary in the allowable reasons that a person can express to obtain an exemption to the mandatory vaccine policy. Exemptions can be granted for medical, religious, or simply philosophical reasons. In contrast, the UK does not mandate vaccination but provides vaccines at no cost.

Mandates

There is a broad base of evidence examining the impact of mandates on disease incidence and vaccination rates in the US. School-entry mandates are a requirement that a parent must document that their child is fully immunized in order to attend school. A number of studies have found that school mandates significantly reduced the incidence of disease in the US. The estimated impact of mandates on vaccination coverage varies widely with a median increase of 15% (range: 5–54%). Implementation of a mandate also appears to reduce racial differences in immunization rates. Many of these studies were conducted in the US and most did not control for exemptions. It should also be noted that these studies typically do not control for the potential endogeneity of mandate strength and underlying population characteristics despite preliminary evidence that they may be related.

Mandates also appear to have a spillover effect and may be associated with increased immunization rates for nonmandated vaccines. For example, one study found that the implementation of a mandate for the tetanus, diphtheria, pertussis (Tdap) vaccine increased not only the Tdap immunization rate from 29% to 83% but also the tetravalent meningococcal vaccination rate from 10% to 60%.

Exemptions

A study of exemption rates showed that US states that had an easy exemption process, or that allowed exemptions from mandates based on personal beliefs, had an increase in the number of exemptions between 0.99–2.54% and 1.26–2.51%, respectively, from 1991 to 2004. Additionally states with easy exemptions or personal belief exemptions had an increased incidence of pertussis (incidence rate ratio of 1.53 (ease of exemption); incidence rate ratio 1.48 (personal belief exemptions)). Interestingly a survey of parents of children who received a nonmedical exemption reports receiving at least some vaccinations (75.5%), with the varicella vaccine being the most common one not being received (53.1%).

Subsidies

There is limited evidence on the effects of direct subsidies on immunization rates in high-income markets. Providing free influenza vaccine to elderly patients increased the vaccination rate by a modest 0.6% from a base of 40.6% in the US. Additionally, another study estimated that providing free vaccines increased the overall immunization rate for DTP, polio, and MMR vaccines by 7% from 76% to 83% ($p=.03$) based on the implementation of a new insurance program that provided free vaccines to poorer children. In low-income markets, vaccines are highly subsidized and there has been a marked increase in the vaccination rate for older vaccines, through UNICEF, and newer vaccines, through the GAVI Alliance.

Many countries provide free vaccination to all children. In the US there is a mix where some states provide all vaccines to all children, whereas other states provide them for only low-income publicly insured or uninsured populations. All other children receive their vaccine in the private sector. The evidence on the role of US state level insurance policies is mixed. For example, US states that provide Medicaid coverage to a larger proportion of their poor residents are more likely to have residents that are up-to-date with vaccination, but only to a certain point. States that provide Medicaid coverage to more than approximately 50% of their poor residents are actually less likely to have poor and nonpoor residents being up-to-date with vaccination. Also, children residing in states that provided a higher percentage of immunizations through the public sector were less likely to be up-to-date (Mayer *et al.* 1999). Children residing in states that provide all vaccines for their entire pediatric population do not appear to be more up-to-date than those residing in states that provide it only for their low-income populations (Olshen *et al.*, 2007; Mayer *et al.*, 1999; Davis *et al.*, 2003). However, in contrast a different study found that publicly insured children were more likely to be up-to-date on vaccine requirements compared to their privately insured counterparts (Blewett *et al.*, 2008).

Finally, in the US, children who are black, reside in an urban area, are poor, and are the children of young single mothers who do not have college degrees are more likely to be completely unvaccinated. In contrast, white children who live in families with incomes more than US$75 000 and are the

children of married mothers who have college degrees tend to have some vaccination but are not fully up-to-date or undervaccinated. These findings indicate that policies such as subsidies and mandates may need to be more nuanced with regard to groups that are completely unvaccinated and those that are undervaccinated.

Supply

The vaccine industry in the developed world is primarily comprised of large multinational pharmaceutical companies that produce vaccines as part of their overall product offerings. These multinationals typically supply product to most parts of the world. However, developing world manufacturers now supply much of the demand in low- and middle-income countries. In prior decades, governments and universities were major suppliers of vaccines but few still exist globally with most of the vaccines being manufactured in the private sector.

The relatively small number of manufacturers and recurrent shortages has led much of the economic literature on vaccine supply to focus on two interrelated areas, causes of shortages and the impact of sole source suppliers. Shortages are a recurrent feature in the vaccine market as historically there have been only a few suppliers for each vaccine in each market, and because vaccine production typically takes many months. It is difficult for other suppliers to quickly increase production to meet unanticipated demand. The causes of the shortages vary. For example, an increase in demand due to an early and more severe flu season caused shortages of influenza vaccine during the winter of 2004–05. Additionally, regulatory actions have caused shortages of a number of vaccines over time as manufacturers have either decided to cease production or needed time to meet compliance requirements. Finally, the natural uncertainty in biologic manufacturing has also been associated with vaccine shortages. Supply-side research tends to examine why there is such limited capacity and few suppliers at the product level. Causes such as demand and supply shocks, excess regulation, undervaluation of vaccines, greater profitability of pharmaceuticals relative to vaccines, and fixed cost competition along with winner-take-all procurement policies have all been posited as causes for the current market structure. Analysis of the vaccine supply-side features is important to understanding these market outcomes.

Firm perspective

From the firm's perspective a number of considerations are required to determine whether to enter or exit the vaccine market. A 2004 survey of vaccine manufacturers by the Centers for Disease Control and Prevention (CDC) found that manufacturers highlighted the regulatory burden, the high cost of delay between initial investment and sales, and the higher cost of new technologies for new vaccines as barriers to entering the vaccine market. However, the manufactures also noted that they have increased investment in R&D because new technologies have allowed for the development of vaccines previously not thought possible. During the late 1990s and into the early 2000s there was significant exit from the vaccine market. For example, in the US during the 1980s there were 18 vaccine manufacturers and by 2007 there were six. Exit was not limited to the US market; UNICEF received bids from 10 different manufactures to supply their demand for measles vaccine in 1996, and only three in 2001. However, more recently firms have entered the vaccine market. The causes for potential increased entry are multifactorial. For example, more recent vaccines, like the HPV vaccines, have been priced higher than traditional vaccines, leading to greater expected revenues. Additionally, improved technology, such as cell-based influenza vaccines, has also led to new firms entering the market. Internationally, the GAVI Alliance has substantially increased the donor base for funding low-income markets, increasing the overall market size for many vaccines. Finally, manufacturers in emerging markets, including India, China, and South Korea have entered as major suppliers to emerging and middle-income markets.

Vaccine projects must compete for both internal and external capital that could be dedicated to other innovations. Within the pharmaceutical firm these are typically other pharmaceutical or biologic products that may be more profitable. Firms also consider the increased regulatory requirements and the need to make significant investments early in the product life cycle for vaccine projects. Because vaccines are given to healthy recipients and are complex biologics, there is a higher level of regulatory proof required to show product safety, product efficacy, and the reliability of the manufacturing process. Clinical trials for vaccines typically enroll larger number of subjects than pharmaceutical trials because of the need to prove prevention of a disease rather than showing treatment effects (Table 2). The clinical trials must also demonstrate both the safety and the efficacy of the product and the manufacturing process. This is due to the biologic nature of the product, requiring that investment and decisions on full-scale manufacturing be made before product approval. This increases the opportunity costs of vaccines relative to pharmaceuticals. Finally, because of the linkage between the product and the manufacturing process, changes in manufacturing may necessitate new clinical trials. Therefore, firms face high sunk costs when entering the vaccine market relatively early in the clinical trial process and need to wait longer to recoup their investment. A rational firm will invest in a higher fixed cost endeavor only if there is a high enough willingness to pay for the eventual vaccine.

Vaccines also typically tend to have a longer life cycle than pharmaceuticals because of the uncertain regulatory pathway to entry for biosimilar vaccines. Firms wishing to enter an existing vaccine's market will most likely have to do their own full-scale clinical trials for both their product and their manufacturing process. This regulatory barrier to entry means that firms currently in the market will face less competition and have longer product life cycles than a typical pharmaceutical for which entry is typically very rapid after patent expiry. It should be noted that although there has been increased entry into the overall vaccine market, competition at the product level is still limited.

Market outcomes

A distinctive feature of the vaccine market is that for each type of vaccine there is a limited number of suppliers selling in most countries (Table 1). Single (or few) suppliers is not necessarily a bad outcome if the last remaining supplier is the

lowest cost reliable producer and monopsony purchasing can balance monopoly market power. Moreover, this may be the long-run equilibrium for vaccines. However, having few suppliers can cause episodic shortages if there is a regulatory action, a manufacturing problem, or some other unexpected supply shock, especially if storability is limited or costly. For instance, updated regulations may cause firms to exit due to increasing regulatory costs and if firms do not have a reasonable expectation of being able to recoup their investment.

Multiple factors may contribute to exit of individual suppliers and/or temporary shortages of vaccines in the vaccine market. Vaccines have a high fixed cost component. If there are multiple entrants and competition drives prices to marginal cost, none of the competitors can recoup their fixed costs and all but the low-cost supplier will exit the market. Additionally, demand is relatively concentrated because it is limited to the birth cohort for many vaccines, in which case economic theory predicts there will be fewer suppliers. Regulatory actions have also caused firms to exit the market as well as have led directly to shortages. Historically, liability risk for adverse events associated with vaccination was a significant concern as vaccine recipients are healthy infants; these risks were the impetus for the Vaccine Injury Compensation Act in the US. In the US, most vaccine sales have an excise tax, which is used to fund a Vaccine Injury Compensation Fund. Persons harmed by vaccination must go to a special court to seek compensation from the fund. Finally, some have posited that vaccine prices are too low due to undervaluation by payers and the large amount of government purchasing. However, these studies were done before the introduction of newer and higher priced vaccines. Additionally, a US study did not find evidence that government purchasing led to fewer suppliers and therefore may not be a contributing factor to shortages.

A number of solutions to combat shortages have been suggested. In the US, the CDC has established 6-month stockpiles for all pediatric vaccines. This may be the optimal solution if fixed-cost competition along with winner-take-all procurement provisions leads to sole-source supply from the least-cost manufacturer, as has been hypothesized.

Conclusion

The economics of vaccines are complicated and have a number of interesting features that can shed light on various economic phenomena. Vaccines are preventative against contagious diseases, and therefore marginal demand may decline as immunization rates increase because the risk of infection decreases. A rational consumer may prefer not to vaccinate given that the probability of infection becomes very small at high immunization rates. This creates incentives to free ride. For related reasons, immunization rates in private markets may be suboptimal if consumers ignore the positive externalities from immunization. On the supply side, vaccines are distinct from chemical pharmaceuticals due to their biologic complexity. This complexity leads to greater regulatory oversight and more complicated manufacturing processes for both vaccines and biologics compared to traditional chemical pharmaceuticals. Suppliers deciding to enter the vaccine and biologics market need to demonstrate the safety and efficacy of both their product and their manufacturing process because they are intimately linked. This also means that changes in regulations or manufacturing processes may necessitate new clinical trials. Therefore, significant investments must be made early in a vaccine's life cycle and decisions on capacity may be difficult to modify later. However, once in the vaccine market most vaccines have traditionally enjoyed a longer economic life than chemical pharmaceuticals, due to the absence of a regulatory pathway for generic vaccines. Whether this will change in the future remains to be seen.

Finally, the overall vaccine market has historically experienced shortages and few manufacturers per product. The possible causes of the shortages and few suppliers range from prices being below the full social value, changes in regulation causing manufacturer to exit, fixed cost competition, and high levels of centralized purchasing, either by national governments or by groups like UNICEF. Solutions to these problems range from increasing prices to reflect the full social value of vaccines, stockpiling vaccines, and reducing regulatory barriers to increase global competition.

See also: HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Infectious Disease Externalities. Macroeconomic Effect of Infectious Disease Outbreaks. Water Supply and Sanitation

References

Blewett, L. A., Davidson, G., Bramlett, M. D., Rodin, H. and Messonnier, M. L. (2008). The impact of gaps in health insurance coverage on immunization status for young children. *Health Services Research* **43**, 1619–1636.

Davis, M. M., Ndiaye, S. M., Freed, G. L., Kim, C. S. and Clark, S. J. (2003). Influence of insurance status and vaccine cost on physicians' administration of pneumococcal conjugate vaccine. *Pediatrics* **112**, 521–526.

Mayer, M. L., Clark, S. J., Konrad, T. R., Freeman, V. A. and Slifkin, R. T. (1999). The role of state policies and programs in buffering the effects of poverty on children's immunization receipt. *American Journal of Public Health* **89**, 164–170.

Olshen, E., Mahon, B. E., Wang, S. and Woods, E. R. (2007). The impact of state policies on vaccine coverage by age 13 in an insured population. *Journal of Adolescent Health* **40**, 405–411.

Further Reading

Bauch, C. T., Bhattacharyya, S. and Ball, R. F. (2010). Rapid emergence of free-riding behavior in new pediatric immunization programs. *PLoS One* **5**, e12594.

Berndt, E. R., Denoncourt, R. N. and Warner, A. C. (2009). *US markets for vaccines: Characteristics, case studies, and controversies*. Washington, DC: AEI Press.

Briss, P. A., Rodewald, L. E., Hinman, A. R., et al. (2000). Reviews of evidence regarding interventions to improve vaccination coverage in children, adolescents, and adults. *American Journal of Preventive Medicine* **18**, 97–140.

Brito, D. L., Sheshinski, E. and Intriligator, M. D. (1991). Externalities and compulsory vaccinations. *Journal of Public Economics* **45**, 69–90.

Carlson, B. (2011). *The expanding vaccine market 2012*. Rockville, MD, USA: Kalorama Information Inc.

Chaiken, B. P., Williams, N. M., Preblud, S. R., Parkin, W. and Altman, R. (1987). The effect of a school entry law on mumps activity in a school district. *Journal of the American Medical Association* **257**, 2455–2458.

Coleman, M. S., Sangrujee, N., Zhou, F. and Chu, S. (2005). Factors affecting US manufacturers' decisions to produce vaccines. *Health Affairs* **24**, 635–642.

Danzon, P. and Pereira, N. S. (2005). Why sole-supplier vaccine markets may be here to stay. *Health Affairs* **24**, 694–696.
Finkelstein, A. (2004). Static and dynamic effects of health policy: Evidence from the vaccine industry. *Quarterly Journal of Economics* **119**, 527.
Freed, G. L. (2005). Vaccine policies across the pond: Looking at the UK and US systems. *Health Affairs* **24**, 755–757.
Geoffard, P. Y. and Philipson, T. (1997). Disease eradication: Private versus public vaccination. *American Economic Review* **87**, 222–230.
Hinman, A. R. (2005). Financing vaccines in the 21st century: Recommendations from the national vaccine advisory committee. *American Journal of Preventive Medicine* **29**, 71–75.
Institute of Medicine (US). Committee on the Evaluation of Vaccine Purchase Financing in the United States (2003). *Financing vaccines in the 21st century: Assuring access and availability.* Washington, DC: The National Academies Press.
Kremer, M. and Snyder, C. M. (2003). Why are drugs more profitable than vaccines? *NBER Working Paper*, 9833 (http://www.nber.org/papers/w9833)
Offit, P. A. (2005). Why are pharmaceutical companies gradually abandoning vaccines? *Health Affairs* **24**, 622–630.
Omer, S. B., Pan, W. K. Y., Halsey, N. A., et al. (2006). Nonmedical exemptions to school immunization requirements. *Journal of the American Medical Association* **296**, 1757.
Pauly, M. V. (2007). Drug and vaccine pricing and innovation: What is the story? *Managerial and Decision Economics* **28**, 407–413.
Philipson, T. (2000). Economic epidemiology and infectious diseases. *Handbook of Health Economics* **1**, 1761–1799.
Plotkin, S., Orenstein, W. and Offit, P. (2008). *Vaccines*. Philadelphia, PA: Saunders.
WHO, UNICEF, World Bank (2009). *State of the world's vaccines and immunization*, 3rd ed. Geneva: World Health Organization.

Value of Drugs in Practice

A Towse, Office of Health Economics, London, UK

Introduction

To understand the role of establishing value in practice, one needs to briefly recap on the role that value can in theory play in pharmaceutical pricing and use by health systems.

Most industrialized countries have universal coverage with modest patient copayments. By lowering out-of-pocket prices to patients, insurance counteracts the impact on patients of companies using patent protection to charge high prices, protecting consumers from financial risk and, through cross-subsidies, making health services more affordable to low-income consumers. However, because such insurance makes patient demand highly price-inelastic, insurance creates the potential and incentives for manufacturer prices that exceed the level that would result from patents alone. Public and private insurers use various forms of price regulatory strategies to constrain this producer moral hazard. These price regulatory strategies are generally an ad hoc mix of historical policies. It is, however, possible to identify five broad types of measure, often used in combination:

1. Cost-effectiveness requirements: Drugs are assessed for use or for a reimbursement price by looking at incremental health-related effects (often measured and valued using the quality-adjusted life-year (QALY) and incremental costs relative to existing treatments using cost-effectiveness analysis (CEA). Economists regard the use of CEA for drugs (which has the effect of regulating drug prices indirectly through a review of cost-effectiveness) as in theory more consistent with principles of efficient resource allocation than other regulatory methods for drug prices. It means that more effective/safer drugs (delivering more QALYs) can charge higher prices and still be cost-effective relative to less effective/less safe drugs. This provides efficient incentives for research and development (R&D). In addition, by using the CEA approach, the indications in which the use of the drug would be efficient can also be identified and potentially controlled, thereby encouraging a more cost-effective use of the drug in the health care system.
2. Therapeutic added value requirements: These typically involve comparison with other, established drugs in the same class, or with other treatments used in the standard of care (SoC) with higher prices allowed or negotiated for improved health or health-related effects in the form of efficacy, better side effect profile, or convenience (a form of internal reference pricing). If internal reference pricing ignores potential differences between drugs in a therapeutic group and prices them all at the same price, then it is not rewarding innovation efficiently. However, if companies are able to charge higher prices if they can demonstrate superior effect over other products in a therapy class or over the SoC then prices are taking account of the value generated for payers and their patients. This can be done by using an assessment of relative effectiveness (RE) (the term used in Europe) or comparative effectiveness (CE), the term used in the USA.
3. Comparison with the price of the identical product in other countries ('external reference pricing'): This involves setting prices by reference to the prices of the same product in a basket of other countries. It limits the manufacturer's ability to price discriminate across countries. Predicted effects include convergence in the manufacturer's target launch prices across linked markets, with launch delays and nonlaunch becoming an optimal strategy in low-price countries, particularly those with small markets. Parallel trade, which is legal in the EU, has similar effects to external referencing, except that it generally only affects a fraction of a product's sales. The welfare effects of regulatory pressures for price convergence across countries are theoretically ambiguous but likely to be negative. Price discrimination increases static efficiency if volume increases relative to uniform pricing. That differential pricing increases drug use seems plausible, given evidence of new drug launch delays in low price countries when there is external referencing or parallel trade.
4. Cost-based approaches where manufacturers supply production and research cost information: Most countries have now moved away from direct price control based on costs and profit margins. The difficulties of allocating joint costs across global markets and taking account of R&D failures rendered this a particularly inefficient way of regulating pharmaceutical prices. The United Kingdom (UK) uniquely among industrialized countries regulates the rate of return on capital on the whole drug portfolio, leaving manufacturers free to set the initial launch prices of individual drugs (but not to increase them thereafter except in very prescribed circumstances). Indirect price control is, however, operated via the National Institute for Health and Clinical Excellence (NICE) using a cost-per-QALY threshold in making recommendations for the use of drugs in the UK National Health Service (NHS). Following a 2007 review of the Pharmaceutical Price Regulation Scheme (PPRS), the UK Office of Fair Trade recommended that the UK move to a system of 'value-based pricing' regulation, in place of profit regulation. As a result, the renegotiation of the 2009 PPRS may lead to a move away from profit regulation to a more uniform reliance on the use of NICE's cost-effectiveness requirements to constrain pharmaceutical prices.
5. Limits on total spending with various clawback mechanisms to penalize companies when revenues exceed the target set. Controlling expenditure through drug budget caps is a form of 'silo budgeting,' which may create perverse incentives for cost shifting to less efficient inputs or to curb sales of products that are delivering a lot of health gain.

The first two approaches (the use of CEA and a 'therapeutic added value' approach involving a comparison with other

 doi:10.1016/B978-0-12-375678-7.01205-0

drugs or the SoC) link price to value. There has been a substantial increase in the number of third party payers using formal CEA, or pharmacoeconomic approaches for assessing the value of drugs, vaccines, and other heath technologies to inform decisions about pricing, reimbursement, and use within their health care systems. In addition a number of countries use 'therapeutic added value' approaches, in some cases involving the use of CEA.

The main rationale for using CEA is to limit reimbursement to drugs and technologies that meet specified standards of value for money or cost-effectiveness. CEA also has the effect of regulating drug prices indirectly through a review of cost-effectiveness. This is in theory more consistent with the principles of efficient resource allocation than other regulatory methods for drug prices. It means that more effective/safer drugs (delivering more QALYs) can charge higher prices and still be cost-effective relative to less effective/less safe drugs.

Australia was the first jurisdiction to adopt such a policy in 1993 and was quickly followed by New Zealand and several Canadian provinces. The UK established the NICE in 1999 to review the effectiveness and cost of technologies expected to have major health or budgetary impact, using cost per QALY, and to formulate guidance on the use of these technologies in the NHS in England and Wales. In Sweden, the Dental and Pharmaceutical Benefits Board (tandvårds- och läkemedelsförmånsverket (TLV)) undertakes CEA to inform decisions on the reimbursement of drugs. Other European countries requesting economic submissions for some, or all, new medicines, include Belgium, Finland, Ireland, Norway, The Netherlands, Portugal, and Germany. Similar policies have also been recently adopted by some countries in Eastern Europe (e.g., Hungary), Asia (e.g., South Korea), and Latin America (e.g., Brazil).

It would be more appropriate to include the new German pricing system in the second 'therapeutic added value' category, along with the French system. Both place a strong emphasis on the need to demonstrate benefit against an appropriate comparator in order to demonstrate added value, before a higher price can be considered. In Germany, if there is no agreement on the price to be paid for the added value, then a CEA can be required to help resolve the issue. In France a CEA is not required at launch, but is required at the point at which a postlaunch review is conducted.

Both the use of CEA and the therapeutic added value approach therefore link price to value. Price (P) can therefore be thought of as a function of the decision maker's perception of value (V). This can be characterized as

$$P = f(V) \quad [1]$$

For the decision-maker value (V) is additional benefit (B) minus additional cost (C). These costs can be thought of as comprising additional costs associated with using the technology (excluding acquisition cost or 'price') minus cost offsets (including the costs saved by the displacement of other technologies). In addition decision-makers looking at value are also concerned about the opportunity cost of resources (k). In the case of payers using CEA this is explicit (although they may not say what opportunity cost threshold they are using). In the case of payers rewarding price premiums for value it is implicit in their willingness to accept higher prices for additional value. Finally, decision-makers are concerned about the uncertainty (U) associated with their estimation of value. Substantial uncertainty is likely to lead to a lower price, delay in use of the drug pending resolution of the uncertainty with more evidence, or some form of use linked to the collection of evidence designed to resolve the elements of uncertainty (often called coverage with evidence development (CED) or managed entry). The decision-maker value determination can be characterized as being a function of these four elements:

$$V = g(B, C, k, U) \quad [2]$$

To understand how value is determined in practice requires consideration of a number of issues, which are explored in the remainder of this article:

- the elements of benefit and cost payers are willing to have included in their assessment of value and how the most important component, health effect, is assessed;
- the use of opportunity cost in determining price or value;
- how pricing and access decisions are made, given the assessment of value;
- the use of CED to handle uncertainty in decision-making about value;
- the challenges of getting value decisions implemented in practice. Do health systems use recommended drugs;
- can poor value be identified using the same approach, leading to disinvestment from drugs and other treatments that are not good value; and
- trends in regional collaboration in the assessment of and decision-making about value.

The Elements of Value to Be Included

What is theoretically included should, in principle, depend on the perspective of the decision-maker. However, some omissions may be unintended and lead to unintended measurement error. For most decision-makers:

- The health effect is usually the single most important benefit and hence element of any assessment of value.
- Cost-offsets within the healthcare system are a second key benefit.

Other elements of value that are sometimes used by decision-makers fall into three distinct types:

- The 'value' of the health gain to society may be higher or lower depending on who gets it. The severity of the disease is a particular factor. The UK NICE applies a specific value weight when appraising end-of-life medicines. Several health systems treat drugs for orphan diseases differently (where a requirement for designation is that the degree of disease severity is high), allowing higher prices and/or lower evidence standards for evidence of RE or therapeutic added value. In the German arzneimittelmarktneuordnungsgesetz (AMNOG) process, orphan drugs are automatically assumed to be innovative without a consideration of the strength of the evidence, although this is now subject to review. In the UK, orphan drugs were exempt from the NICE review process, but this has now changed.

However, NICE will use a different process to review these drugs as compared with its conventional CEA approach.

- There may be elements of benefit to the patient that are not necessarily captured in the measure of health gain, including:
  - Health-related-quality of life aspects not well captured by a generic measure of health gain such as the QALY and
  - Health-care-process-related aspects such as being treated with dignity, at a convenient time and location, and after only a short wait.
- Information for the patient which, for example, enables life style choices to be made, independent of any health effects.
- Other costs and benefits beyond those to patients and the health care system. Outside of health care a societal perspective is conventionally used by economists, including all costs and consequences related to the initial interventions in a cost–benefit analysis. Applying such an approach would involve expanding the CEA to include unrelated medical costs, costs incurred outside the health care sector, and benefits accruing to all stakeholders in society including those for the patient not captured in the QALY. Several countries, including Norway, Sweden, and the Netherlands, already require that economic evaluations are conducted using a societal perspective.
- Innovative attributes of a technology may be deemed to have value independently of the health gain generated. Japan and Italy use a categorical rating to assess the degree of innovativeness. France uses a categorical rating to estimate the degree of therapeutic added value.

Table 1 is a summary of the ways that value is linked to medicine prices in a range of eight countries whose third party payers use either CEA or a variant of 'therapeutic added value' to assess price. The authors look at:

- Australia, Canada, England, and Sweden which use CEA to determine whether at the price sought by the manufacturer the medicine is deemed cost-effective or not;
- France, Germany, Italy, and Japan – whose approaches include allocation of new medicines into a number of pricing categories defined by assessment of the therapeutic added value of the medicine.

Table 1 sets out the key points to note. For example:

- The Australian approach to medicines pricing is focused around health gain per dollar, i.e., 'clinical effectiveness' and 'cost effectiveness.' Official guidance there notes that cost per QALY is commonly used but does not require it, meaning that a number of approaches may be acceptable. The overall assessment of the value of a new medicine, in the sense of how different aspects of value are weighed up against price, is opaque. For example, there is no specific monetary 'threshold' value applied to QALYs where some of the medicine's benefits are expressed in QALY terms. A national committee representing the payer engages in a deliberative process: there is no formulaic derivation of the price ceiling. Repeat manufacturer submissions are normal as both sides 'negotiate' toward an acceptable price.
- None of the countries has gone so far as to define an explicit method for aggregating qualitatively different nonfinancial elements of a medicine's value, although three of them group medicines into a small number of categories before price determination: five categories in France, three in Italy, and six in Japan.

There is an important division between those markets and payers who use QALYs and those who do not. Typically payers using CEA require or prefer the use of QALYs and those using therapeutic added value do not require it. In the US, the 2010 US Patient Protection and Affordable Care Act (ACA) specifically forbids the use of "a dollars per quality-adjusted life-year (or similar measure that discounts the value of a life year because of an individual's disability) as a threshold to establish what type of health care is cost-effective or recommended…" in public funding decisions (ACA 2010). The wording of the ACA means that the emphasis in the public sector in the US is likely to be around assessments of therapeutic added value (termed CE) that use clinical or disease-specific patient reported outcome measures. Many private health plans and pharmacy benefit managers (PBMs) require elements of a CEA submission for drugs to be provided to them in a standard format agreed by the Academy of Managed Care Pharmacists (AMCP). This format allows for, but does not mandate, use of the QALY.

Estimating the Opportunity Cost of Adopting a Technology

Decision-makers need to know what they are giving up if they adopt a drug. In the case of a health care system with a 'hard' fixed global budget for a specified time period, such as the UK, the opportunity cost in the short terms is usually displacing another health care related activity. This is often referred to as an 'extrawelfarist' approach. In the case of (1) a 'soft' public budget system, (2) as spending budgets are varied over time in 'hard' global budget countries, or (3) in a private sector system, adoption may lead to increases in taxes or premiums and so reductions in private consumption elsewhere. In this context, the relevant opportunity cost is the expected willingness to pay (WTP) for health-related value of the covered population. This is often referred to as the 'welfarist' approach. Which one is relevant depends on the context. In the case of countries using therapeutic added value approaches, a rule of thumb is usually used to estimate WTP for additional value (e.g., by reference to prices sought elsewhere) or a price is negotiated. The approach to the threshold is implicit in these cases and may not reflect either WTP or an estimate of displaced value.

Setting an appropriate opportunity cost threshold for use in decision-making by estimating what is 'displaced' is difficult because (1) it is hard to estimate and (2) it is hard to apply in decision-making as the covered population has usually not accepted explicit rationing of health care on this basis.

The UK has the most explicit policy in respect of using a cost-effectiveness threshold for assessing price in relation to value based on the 'displacement' of other health-providing activities by the NHS within a fixed global budget. In its early days, NICE denied that it was applying a specific threshold. However, as the information on the decisions made by

Table 1 Assessing 'value' and linking to price: current practice in selected countries

Country	*Elements included in 'value'*	*How measured*	*How valued, whose values*	*How aggregated*	*How converted into price*
Australia	• Clinical effectiveness • Cost-effectiveness	Quality-adjusted life-year (QALY) and incremental cost per QALY are commonly used but not obligatory	Not specified	Deliberation – opaque	Negotiation – approximately 30% margin on costs for the most innovative products; effectively therapeutic reference pricing for others
Canada – Federal level	• Cost-effectiveness • Safety • Effectiveness	• Incremental cost per QALY, where possible • ? • QALY where possible	Preferences of general public preferred; patients' preferences may be acceptable	Not specified	Price not linked to value: max price of 'breakthrough drugs' = median of prices in 7 other countries; effectively therapeutic reference pricing for others
France	• Relative Efficacy • Safety • Availability of therapeutic alternatives • Disease severity	• Not specified • Not specified • Not specified • Not specified	Not specified	Categorization by expert clinical committee into one of five categories of incremental health benefit (ASMR)	Negotiation (on price and volume, i.e., total revenue). For drugs with major therapeutic improvements, reference is made to European prices (in Germany, Italy, Spain, and UK)
Germany	• Relative efficacy • Very small or orphan market	• One of three categories for strength of proof • Yes/no	Not specified	Deliberation – opaque	Negotiation for products with additional therapeutic value. If negotiations fail use of a cost-effectiveness analysis and/or pan-European reference pricing. Therapeutic reference pricing for others
Italy	• Clinical effectiveness • Availability of therapeutic alternatives • Disease severity	• Unspecified clinical end-points leading to one of the three categories • One of the three categories • One of the three categories	Not specified	Categorization by expert clinical committee into one of three overall categories	Negotiation
Japan	• Efficacy • Safety • New mode of action • Indicated for children • Small or orphan market	• Not specified • Not specified • Yes/no • Yes/no • Yes/no	Not specified	Categorization by Ministry of Health and Welfare into one of the six usefulness and market size categories	Negotiation
Sweden	• Clinical effectiveness • Cost effectiveness • Cost savings in any sector: health care, nonhealth, public, private, patients, carers, and relatives • Production loss	• QALYs • QALYs • Money • Money (human capital method)	Preference for 'QALY weightings based on appraisals of persons in the health condition in question'	Not specified	Manufacturer selects price and faces coverage decision by TLV

(*Continued*)

Table 1 Continued

Country	*Elements included in 'value'*	*How measured*	*How valued, whose values*	*How aggregated*	*How converted into price*
England	• Health gain • Health service cost savings • Severity/end of life (cancer only)	• QALY • Money • Within 2 years of expected death: yes/no	• General population perspective de facto • Market prices • Appraiser deliberation	As QALYs (weighted if 'end of life')	Manufacturer selects price and faces coverage decision by National Institute for Health and Clinical Excellence

Source: Adapted from Sussex, J., Towse, A. and Devlin, N. (2013). Operationalising value based pricing of medicines: A taxonomy of approaches. *Pharmacoeconomics* **13**(1), 1–10.

NICE accumulated, it was possible to estimate a revealed threshold. NICE then stated that it applied a threshold range: interventions with an incremental cost per QALY ratio below £20 000 have a high probability of funding; and those with a ratio exceeding £30 000 have a low probability of funding although the upper bound of £30 000 can be exceeded, for example, on grounds of equity. This range is set out in the NICE Methods Guide. Research in the UK is beginning to tackle the issue of the value of what might be displaced in the NHS. One study using case studies found it difficult to identify what in practice was displaced at the local level. Another series of studies have attempted to estimate the threshold level implied by a longitudinal and cross-sectional analysis of the current pattern of expenditure by disease area by geographic area within the NHS combined with data on mortality and estimates of morbidity. Their conclusion is that NICE's threshold range may be a little too high (i.e., closer to £18 000 than £20 000–£30 000). Yet other evidence analyzing the revealed preference of NICE decisions suggest the actual threshold used by NICE may be closer to £46 000 per QALY when other factors are taken into account by the NICE Appraisal Committees. NICE has commissioned WTP-based opportunity cost estimates of the social value of a QALY. One study suggested a distinction in WTP between life-saving (£70 000), life-extending (£35 000), and quality-of-life-enhancing (£10 000) considerations. A second study suggested a range of £20 000–£40 000 per QALY. The UK Department of Health currently uses a WTP for a QALY estimate of £60 000 for its own impact assessments, adapted from WTP estimates used elsewhere in the government.

Evidence from other countries suggest the following:

- In Canada, a revealed preference estimate of the ICERs of drugs approved and refused by the CDR between 2003 and 2007 suggesting overlapping ICERs indicating other factors were important (Belgian Health Care Knowledge Center, 2008).
- In Australia, revealed preference estimate of AUS$37 000–69 000 per QALY was derived for decisions in the 1990s (Belgian Health Care Knowledge Center, 2008).
- WTP estimates for the US range between $100 000 and $300 000 per QALY (Eichler *et al.*, 2004).
- The Swedish TLV uses WTP-based estimates ranging from €40 000 to €90 000 per QALY (Persson, 2012).

The main arguments for an explicit threshold are that it may encourage more consistency in decision-making and lead to a more equitable outcome as a result of public debate. Manufacturers will know what payers want to reward and invest in R&D accordingly. There are issues as to whether such a cost-effectiveness threshold encourages the 'right' amount of innovation. Some have argued that the amount of the social benefit going to innovators will be too low; others that allow innovators to price up to the threshold means they can appropriate all of the benefit. The patent system is intended to provide temporary monopoly rights that can enable innovators to exercise some market power. Arguably, the most dynamically efficient outcome is for as much as possible of the social surplus to accrue to the innovator in that period. Payer use of cost-effectiveness thresholds will be (second best) efficient if thresholds reflect societal WTP. For a discussion of this issue, see Danzon *et al.* (2011).

Making Decisions about Value

Use of CEA and/or 'therapeutic added value' requires decision-makers to assess the evidence of value, use judgement, and make a decision. A key part of the process of review is the submission by drug companies of a dossier of evidence to support their claim for a price based on their estimate of value. By way of illustration one can look at the process used by the UK NICE, which is regarded by many as an exemplary process. It involves the appraisal of a single technology:

- The opportunity for early scientific advice (before Phase 3 trials) as to the health and cost outcomes likely to be of importance to NICE. This is nonbinding on both parties and can involve the drug licensing body should the company wish this.
- A scoping exercise before the preparation of the company dossier at which both parties (and other relevant stakeholders) seek to agree the exact scope of the evidence required.
- The submission of the dossier, which should follow the Methods Guidance. Other stakeholders (e.g., patient groups) can also submit evidence.
- Review of the dossier(s) by an independent Evidence Review Group contracted by NICE and preparation of an assessment report by them.
- Appraisal by the NICE Appraisal Committee of the dossier and the Assessment Report and a recommendation by the Committee about the NHS use based on the price offered by the company.
- Opportunity for comment by the company and all other stakeholders before reconsideration by the Committee.
- An Appeal option for the company.

- The option for the company to apply for a Patient Access Scheme should its proposed price proves to be too high. This could involve some sort of price discount or collection of improved evidence to support the proposed price.
- The option of a rereview at a later specified time (usually 2 years).

Note that in practice two types of decision are made by health systems depending on whether (1) the health-technology-assessment (HTA)/pricing and reimbursement body, given an assumption about access, is determining the price or (2) given price, the committee is determining if value is positive for some/ all groups of patients, in order to list or otherwise make the product available for the relevant patient groups. There is inevitably some blurring:

- A company may refuse to accept the price offered and supply more evidence on value.
- If access is denied, the company can resubmit claiming higher value or offering a lower price.

Most payer/HTA bodies have a committee similar to the NICE Appraisal Committee to assess the evidence and make a decision. The mechanism by which the members of a committee combine the various forms of evidence with local context and judgements about interpretation and uncertainty to reach a decision is a deliberative process. This may be particularly valuable in circumstances where there is either uncertainty about technical information (scientific uncertainty), or where issues relating to fairness and social values (value judgements) need to be taken into account. Culyer (2009) defines these as follows:

- Scientific judgment is usually about an effect (positive or negative), its size, the ways in which it can be achieved, for whom, for how long, and how much uncertainty there is about the outcomes;
- Value judgments tend to be in a different territory but they might be about, for example, how worthwhile a technology is, how defensible the tough bits of the decision are, how tolerant of uncertainty the committee ought to be, how important interpersonal comparisons are (who benefits), and whether the outcome measure is a good tracker of the relative health benefits of the interventions that were compared.

Although NICE appears to be more transparent than other HTA bodies, some researchers are critical of its failure to formally codify the impact of decision criteria other than cost-effectiveness, claiming that its statements on these matters have been vague and uninformative. The importance of social value judgements and other factors beyond cost-effectiveness is regularly emphasized, and examples of interventions with high questionable cost-effectiveness being recommended on the basis of such factors are given. These are, however, unusual cases. It is difficult in most cases to understand the extent to which they have contributed to the final recommendation decisions and it is not possible from a review of decisions to find any factors other than the threshold that explain NICE decisions except when they are in the areas of cancer. Thus, though in principle NICE's decision-making fits the description of a sound deliberative process, the lack of explicit reporting of this process means that clarity is not always achieved.

This raises the question as to whether decision support tools can improve the transparency and effectiveness of a deliberative process used by a payer HTA body. Multicriteria decision analysis (MCDA) methods have been advocated for use in health care priority setting. MCDA is a methodology for appraising options on multiple (often conflicting) criteria with the goal of providing a combined appraisal that includes an overall ordering of those options. It provides a framework for explicitly trading off various objectives against each other. It is particularly useful when these objectives do not share a common unit of valuation – for example, health care programs typically involve a mixture of health, monetary, distributional, and political objectives.

Use of MCDA would be attractive if it led to processes becoming more transparent and systematic, so improving both the signals sent to patients and drug developers, and the quality of decision-making. However, it could require a greater time commitment on the part of decision-makers. The burden on decision-makers in using this approach would need to be proportional. To date no HTA body is using formal MCDA techniques.

Using Coverage with Evidence Development Evidence to Handle Uncertainty in Decision-Making

Decisions by payers about the adoption of health technologies are almost always made under uncertainty and on the basis of limited information. Yet it is not at all clear how decision-makers handle uncertainty. Most request evidence on the sensitivity of the evidence to different assumptions or statistical error. NICE asks for probabilistic sensitivity analysis and the presentation of results in a Cost-Effectiveness Acceptability Curve, which estimates the likelihood of a decision to adopt representing good value at different cost-effectiveness thresholds. However, there is no guidance in the Methods Review or in published guidelines as to how judgements about value take uncertainty into account.

Health care payers have three options in respect of uncertainty. One is to adopt the technology and live with the uncertainty. A second is to refuse to adopt the health technology in question until the uncertainty is reduced – either through better evidence or a lower price. The third is to adopt the health technology, but make this decision conditional on the collection of further additional evidence. This is usually termed 'CED or a Managed Entry Agreement (MEA). Adopting a drug may make some forms of additional research more difficult, for example, by reducing the likelihood of enrolling patients in a clinical trial (although data could be collected in another jurisdiction if it is likely to be transferable), and there may be costs of reversing decisions if subsequent evidence suggests that a drug in use is not cost-effective. Yet payers are increasingly using MEAs, many of which are forms of CED. In some cases MEAs are designed to address uncertainty as to how well the drug will perform or as to the overall budget impact. Agreements in France, Australia, and New Zealand are

designed to cap expenditure. In other cases, they are intended to provide an effective price discount, at least for the period of the agreement. The dose-capping agreement that NICE entered into over ranibizumab (Lucentis®) for macular degeneration could be seen as an effective price discount. Cost-effectiveness to NICE was only acceptable if the NHS paid for up to 14 injections per eye of eligible patients. Novartis had to bear the costs of treatment beyond this.

There has been a surge of interest in the use of one particular form of CED – 'performance-based risk sharing' (PBRSA) – that involves an agreement between a payer and a pharmaceutical, device, or diagnostic manufacturer, where the price level and/or revenue received is related to the future performance of the product in either a research or real-world environment. In particular, there is an agreement about a program of data collection to reduce uncertainty about the expected cost-effectiveness of the drug (or device or diagnostic), and the price and/or revenue is linked to the outcome of this program of data collection (Towse and Garrison, 2010). This may be prospective or retrospective. These may be as follows:

- Tackling outcomes uncertainty: The UK multiple sclerosis (MS) drugs scheme addresses outcome uncertainty with a prospective observational study of patient health status with price linked to a cost-per-QALY threshold. In Australia, the agreement for bosentan (Tracleer®) linked price to patient survival using a prospective observational study.
- Tackling subgroup uncertainty, conditional on expected outcomes: The UK bortezomib (Velcade®) example tackles subgroup uncertainty, ensuring identification of responders. There is retrospective payer reimbursement for nonresponders. Responders receive further doses of the product. The Italian Medicines Agency (AIFA) has established several responder-related pay for performance agreements with discounts for trial periods, and rebates for nonresponders. For responding patients, the treatments are reimbursed at full price.
- Tackling subgroup uncertainty via utilization management. In Australia, expenditure caps can also be viewed as risk-sharing agreements that have implicitly tied revenue to outcomes, under the assumption that high volumes mean cost-ineffective care at the prevailing price.

Evidence to date has been mixed. In the UK, for example, the MS risk sharing scheme has attracted much criticism. However, schemes appear to be much more successful in Australia and in Italy, and the UK NICE does operate a form of MEA called Patient Access Schemes, which combine options of effective price discounts and forms of CED including PBRSAs.

Implementation of Technologies Regarded as Good Value

To help facilitate the implementation of decisions and the adoption of cost-effective treatments, national and local health authorities employ a variety of strategies, ranging from provision of financial planning tools; additional funding to cover the adoption of new technologies; and information dissemination. Sweden uses a network of experts to assist decision-makers in understanding guidance and adopting recommended technologies into clinical practice. NICE in the UK takes a similar approach via the use of an in-house Implementation Directorate to ensure that dissemination activities are targeted to the local NHS. It assesses and reports on the level of compliance with guidance across the NHS using a variety of data sources on prescribing and practice patterns, examining utilization trends in relation to the expected level consistent with NICE guidance. Reaching local practitioners may be especially important given their role in the diffusion of technologies – a survey of HTA initiatives in Europe concluded that clinicians frequently fail to change their practice in line with HTA-based recommendations.

Research indicates a range of issues that influence whether recommended treatments are indeed used in health care systems. Such factors include insufficient or misaligned policy aims (i.e., differences in objectives between the HTA process and the needs of decision-makers – the more decentralized a given health system, the more this may be an issue); a lack of a holistic approach to implementation, where not all relevant stakeholders are informed of decisions or there is poor dissemination of guidance; limited use of formal mechanisms to enforce implementation; and rapidly changing political situations. In addition, local authorities often deal with different resource capacities, patient populations, health needs, and available budgets, which can impact their ability to implement national decisions or guidance to make treatments accessible to their populations. In the case of high-cost drugs a problem often arises if local budget holders are reluctant to make monies available to fund the prescribing of a drug approved by the HTA/P&R body.

In the UK, for example, an analysis of public comments on NICE suggests that there is a significant concern regarding the patchy and slow implementation of adoption recommendations, with many stakeholders deeming this a key issue in terms of the Institute's effectiveness, efficiency, and public credibility. A study of NICE guidance implementation found that poor financial planning by local health authorities, in terms of adequately estimating the costs and resource requirements of implementation, was one of the factors contributing to poor implementation. The UK government has published an 'Innovation Health and Wealth Report,' which has reinforced the requirement for local purchasers (commissioners) to provide funds to enable clinicians to prescribe NICE-approved drugs.

Although negative guidance will always be implemented, positive guidance often may not due to the resource consequences and difficulty making disinvestment decisions elsewhere. It may also reflect a local view that the threshold is too high. In Sweden, the local authorities who fund health care are unhappy with recommendations made by the national drug HTA body (the TLV) because they regard the items they will have to displace from their budgets are more valuable than the drugs approved by the TLV.

Although budget holders may have an incentive to resist adoption, financial incentives can be created to reward the use of cost-effective treatments. Several jurisdictions, namely, Denmark, Germany, and England, have introduced regulatory levers to make decisions or guidance legally binding, with the latter also using financial incentives through 'pay for

performance' schemes linked to the uptake of NICE guidance and standards via NHS Quality and Outcomes Frameworks (which make additional payments to GPs for achieving health outcome targets), and a Commissioning for Quality and Innovation payment framework, which holds back a proportion of tariff payments from hospitals who fail to achieve pre-set quality targets and gives the money to those that do achieve them.

Reassessment after a technology has been used in practice is also an important mechanism to facilitate effective implementation and appropriate use of technology. It helps ensure that assessments of value are up-to-date with changes in a technology and the availability of evidence. Several countries, such as France and the UK, have a structured process, conduct re-evaluation at fixed or variable intervals (e.g., every 3–5 years), whereas other jurisdictions initiate subsequent reviews if new characteristics of the product emerge or if new or better clinical and/or economic evidence becomes available.

Disinvestment from Treatments that are not Good Value

HTA processes often focus on new technologies, giving insufficient attention to existing treatments that may be potentially inefficient or used inappropriately. However, given increasingly limited resources and growing emphasis on value for money, several review bodies, such as NICE and the TLV, are implementing 'disinvestment' programs or strategies.

Disinvestment is an explicit process of taking resources from one service in order to use them for other purposes of better value. Rather than a sole focus on allocating new resources, it focuses on eliminating existing 'waste' in the system. These technologies may be effective and hence valued by patients and clinicians. They may, however, be poor value for money.

The notion of disinvestment makes obvious conceptual sense to ensure efficient resource allocation. In practice, however, removing or limiting currently available services, even if cost-ineffective, raises challenges. There is likely to be opposition from clinicians, interest groups, and patients if existing technologies, services, or facilities are no longer made available. Local citizens may give a higher value to services at risk, than to technologies they do not yet have, even if the latter deliver more health care than the former.

A recent HTAi Policy Forum discussion of the issue (Henshall *et al.*, 2012) concluded that the term 'disinvestment' was unhelpful. It was more helpful to think of a process of reassessment, followed by decisions of optimal use, followed by implementation of optimal use decisions. Implementation should be through Managed Exit strategies in the same way as Managed Entry strategies were considered for the introduction of new technologies.

Regional Collaboration in the Assessment of and Decision-Making about Value

As a growing number of jurisdictions request economic data in support of their decision-making procedures for the pricing and/or reimbursement of health technologies, demands on study sponsors and researchers increase, especially as the various national guidelines may insist on the presentation of local data, or the use of specific methods that are not required elsewhere.

There are several reasons (related to benefits, costs, uncertainty, and opportunity cost), why the value of health technologies might vary from place to place. The most important of these are (1) differences in population mix, including the incidence and severity (baseline risk) of the disease in question; (2) clinical practice patterns that may influence relative or CE by impacting on one or more of (a) the relevant comparator or existing SoC, (b) the effectiveness with which the patient is managed with the drug, and (c) the resources associated with current SoC or use of the new technology; (3) relative prices, and (4) the willingness of the health system to pay for the new drug.

However, the requirement that economic evaluations should use local data, or that particular methods should be used, means that analyses increasingly need to be customized for each setting. Transferability is a key issue and there is guidance on good research practices for dealing with aspects of transferability, including analytic strategies and guidance for considering the appropriateness of evidence from other countries.

As discussed above, methods to assess value can vary. There are, for example, differences as to (1) how health effects should be presented, notably in attitudes to the acceptability of the QALY and (2) the acceptability of evidence using indirect comparisons or observational studies. Together with the scientific issues and different WTP these put bounds on the potential for regional collaboration.

The most important attempt to date at increasing regional collaboration is the EUnetHTA project (www.eunethta.net) promoted and funded by the EU designed to produce a 'core' HTA template for an assessment that could potentially be used by decision-makers in several jurisdictions. Progress will depend on the extent to which there is agreement on common methods and requirements and on the extent to which evidence in one jurisdiction is relevant to another. In other words, real benefits from regional collaboration may come not only from methods of convergence but also if some data is transferable with little or no adaption required. This is an empirical issue for which there is currently little evidence. We might expect, however, the potential for common HTA assessments (using agreed methods and data generalizability) to be greater in the systematic reviews of the clinical efficacy data than in the economic evaluation component of assessing benefits and costs. EUnetHTA is exploring the potential for single RE assessments to be shared by EU Member State HTA bodies. The expectation is that (1) the translation of this evidence into an estimate of benefit, (2) prices and resource use (and therefore cost), and (3) the threshold WTP would vary by member state. Thus value decisions would remain local.

The final step in regional collaboration could be to make a common decision, as is currently the case for drug licensing within the EU. Some have argued for a new body making decisions on behalf of all EU Member States. There are scientific, practical, and political issues here. As noted, clinical practice may vary, with implications for the choice of relevant comparator and resource use. RE will depend on patient mix,

baseline risk, and the comparator. Cost-effectiveness is likely to vary between countries. Even where it does not, decision-makers in one country, faced with the same assessment of evidence, may still come to a different decision than those in another jurisdiction about use of the same technology. This is because countries have different levels of resource to devote to health care and different priorities. It would be efficient for prices to differ to increase availability across countries, although the use of reference pricing and parallel trade within the EU make this difficult. EU countries retain different health care systems, with the intent that access decisions will differ. It makes sense, however, to take advantage of economies of scale in information generation, whilst recognizing this does not require that the information is used the same way by different countries.

Regional bodies exist in other parts of the world – notably Asia (HTAsiaLink) and The Americas (HTA Network of the Americas (RedETSA)). However, they are at relatively early stages of collaboration, and there is even greater heterogeneity between the health systems of countries participating in the networks than between the member states of the EU. As in the case of the EU, however, it makes sense to take advantage of economies of scale in information generation, and to share institutional learning, whilst recognizing this does not require that the information is used the same way by different countries.

See also: Cost-Effectiveness Modeling Using Health State Utility Values. Health and Its Value: Overview. Quality-Adjusted Life-Years. Valuing Informal Care for Economic Evaluation. Willingness to Pay for Health

References

Belgian Health Care Knowledge Centre (2008). Threshold values for cost-effectiveness in health care. *KCE Reports 100 C*. Brussels: Belgian Health Care Knowledge Centre.

Culyer, A. J. (2009). *Deliberative processes*. London, UK: Office of Health Economics.

Danzon, P. M., Towse, A. and Mestre-Ferrandiz, J. M. (2011). Value-based differential pricing: Efficient prices for drugs in a global context. *NBER Working Paper w18593*, December 2012.

Eichler, H. G., Kong, S. X., Gerth, W. C., Mavros, P. and Jönsson, B. (2004). Use of cost-effectiveness analysis in health-care resource allocation decision-making: How are cost-effectiveness thresholds expected to emerge? *Value in Health* **7**(5), 518–528.

Henshall, C., Schuller, T. and Mardhani-Bayne, L. (2012). Using health technology assessment to support optimal use of technologies in current practice: The challenge of "disinvestment." *International Journal of Technology Assessment in Health Care* **28**, 203–210, doi:10.1017/S0266462312000372.

Persson, U. (2012). Value based pricing in Sweden: Lessons for design? *OHE Seminar Briefing*. Office of Health Economics. Available at: www.ohe.org (accessed 28.08.13).

Towse, A. and Garrison, L. (2010). Can't get no satisfaction? Will pay for performance help? Toward an economic framework for understanding performance-based risk-sharing agreements for innovative medical products. *PharmacoEconomics* **28**(2), 93–102.

Further Reading

Appleby, J., Devlin, N., Parkin, D., et al. (2009). Searching for cost effectiveness thresholds in the NHS. *Health Policy* **91**(3), 239–245.

Baltussen, R. and Niessen, L. (2006). Priority setting of health interventions: The need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation* **4**, 14.

Devlin, N. and Parkin, D. (2004). Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Economics* **13**, 437–452.

Drummond, M. F., Barbieri, M., Cook, J., et al. (2009). Transferability of economic evaluations across jurisdictions: ISPOR good research practices task force report. *Value in Health* **12**(4), 409–418.

McCabe, M., Claxton, K. and Culyer, A. J. (2008). The NICE cost-effectiveness threshold. What it is and what that means. *Pharmacoeconomics* **26**(9), 733–744.

National Institute for Health and Clinical Excellence (2008). *Guide to the Methods of Technology Appraisal*. London: National Institute for Health and Clinical Excellence.

Neumann, P. J. (2004). *Using cost-effectiveness analysis to improve health care opportunities and barriers*. New York, NY: Oxford University Press.

OECD (2005). *Health technologies and decision-making*. Paris: Organisation for Economic Co-operation and Development.

Rawlins, M., Barnett, D. and Stevens, A. (2010). Pharmacoeconomics: NICEs approach to decision making. *British Journal of Clinical Pharmacology* **70**, 346–349.

Shah, K. K. (2009). Severity of illness and priority setting in healthcare: A review of the literature. *Health Policy* **93**, 77–84.

Sorenson, C., Drummond, M. and Kanavos, P. (2008). *Ensuring value for money in health care: The role of health technology assessment in the European Union*. Copenhagen: WHO. Regional Office for Europe (Observatory Studies Series No. 11).

Sussex, J., Towse, A. and Devlin, N. (2013). Operationalising value based pricing of medicines: A taxonomy of approaches. *Pharmacoeconomics* **13**(1), 1–10.

Value of Information Methods to Prioritize Research

R Conti and D Meltzer, University of Chicago, Chicago, IL, USA

Introduction

Value of Information (VOI) is an outgrowth of advances in Bayesian decision theory and welfare economics that seeks to quantify prospectively the benefits and costs of research and development (R&D) activities under uncertainty. VOI allows for the identification of sources of treatment uncertainty and provides a method to calculate the incremental value of pursuing research to inform clinical practice. This article provides details of the principles behind its estimation, and the different decisions it can inform. Some practical applications of VOI in research prioritization are considered.

Before exploring current applications of the method in detail, it is important for readers to appreciate why the development of economic methods to prospectively assess the value of medical R&D is important. Economists have long noted that economic growth in advanced economies is driven by the creation of innovative ways of producing goods and services, a process that triggers productive investments and allows its benefits to spill over from one country to others. Typically, these economies rely on the profit motives of private enterprises and capital markets to fund innovative efforts. Yet in some economic sectors, such as biomedicine, there is substantial public investment in R&D. According to economic theory, public R&D efforts should act only to complement private investments, where the expected social value is large, but expected profit is small. Economic work since Smith has suggested relief from premature morbidity and mortality offers significant individual and social value. Recent empirical work suggests that the spillovers from public investment in medical R&D to the private sectors' production of novel medical therapies and diagnostics are significant. Schumpeter (1942), Nelson (1959), and Arrow (1962), among others, are the first to articulate the view that R&D and innovative activities are difficult to finance in a freely competitive marketplace. The general argument goes as follows: the primary output of resources devoted to invention is the knowledge of how to make new goods and services, and this knowledge is nonrival so that use by one firm does not preclude its use by another. Moreover, if that knowledge cannot be kept secret, the returns to investment in knowledge cannot be captured by a firm undertaking the research investment, so that firms will be reluctant to invest, leading to under provision of R&D in the economy.

As a consequence, public policies that promote and underwrite innovation are centerpieces of national economic strategy. In the US, National Institute of Health (NIH) funding leads both the public and private sector as the single largest source of support for medical R&D. Within the NIH's purview are direct investments in intramural research, including basic science, preclinical and clinical medical studies, and extramural grants supporting university research efforts. Other advanced economies support medical R&D efforts along similar mechanisms.

And yet, over the past several decades, the unprecedented increase in healthcare expenditures throughout the world, and especially in the US, has prompted increasing concerns that current levels of healthcare expenditures are excessive. It has been suggested that medical care has been provided not only beyond where its marginal benefit exceeds its costs, but often into ranges where there is little or no benefit. The importance of determining when specific medical technologies are worthwhile has intensified with the growing recognition that increases in medical spending have been largely driven by the development and diffusion of new medical technologies. As a consequence, controlling healthcare costs will ultimately require controlling the development and diffusion of medical technology. In accomplishing this goal, reimbursement systems that provide both developers and users of new technology with the appropriate incentives to control costs and produce quality healthcare are essential. Similarly, it is critical to have tools and policies to prioritize investment in public medical R&D efforts.

Retrospective analysis of previous investments may provide useful information to inform assessments of future research endeavors. Recent empirical economic research suggests that improvements in health have been a major component in the overall gain in economic welfare during the plast century for the US, developed, and developing countries. Murphy and Topel (2006) have used cost-benefit analysis to estimate the overall value of medical research and the value of R&D for specific medical conditions. They have found that gains in US longevity due to advances in medical research since 1970 have had an aggregate value of $3.2T, a figure roughly equal to half of gross domestic product. Furthermore, programs aimed to expand public support for specific types of medical innovation, such as the 1970 declared 'War on Cancer' appear to have produced substantial gains in morbidity and mortality. For example, pediatric cancer patients in the US and abroad, and adult patients suffering from breast and prostate cancer, some forms of leukemias, have experienced substantial gains in life expectancy over the past 15–20 years and many believe that such efforts have a great potential to produce more success. Philipson and Jena (2005) have used cost-benefit analysis to estimate the net value of antiretroviral therapies for the treatment of human immunodeficiency virus infection/acquired immunodeficiency syndrome. They have found that patients gained substantial benefits (measured in survival) from the introduction of antiretroviral therapy. In addition, US patients diagnosed with cancer between 1983 and 1999 experienced greater survival gains than their European counterparts; even after considering higher US costs. These findings do not appear to have been driven solely by the earlier application of diagnostic methods.

Whether research funds underwriting these efforts are being allocated to the 'correct' opportunities, and whether the fruits of these investments are valuable at the margin, applied to specific patients, are important and complementary questions. For example, novel approaches to cancer have been a significant focus of public and private sector investment in the past 30 years. These investments are bearing much fruit – in 2012 alone, the Food and Drug Administration has approved 19 anticancer drugs and over 900 anticancer drugs are in various phases of preapproval testing, more than the number for heart disease, stroke, and mental illness combined. Yet, anticancer drugs also rank first in terms of total drug spending by therapeutic area: $23 billion in 2011, up from $18 billion in 2007. Global spending on anticancer therapies is projected to amount to $75–80 billion by 2015, more than any other therapeutic class of pharmaceutical products. Conti *et al.* (2013) have found that commonly used, novel chemotherapies are more often used onlabel than offlabel in contemporary practice. Total national spending on these chemotherapies has amounted to $12 billion (B; $7.3B onlabel, $2B offlabel and supported by additional clinical evidence and expert judgment, and $2.5B offlabel and unsupported by clinical evidence and expert judgment).

Both Congress and the National Institutes of Health have faced increasing pressures from disease specific interest groups in recent years to justify their decisions regarding medical resource allocation, and questions such as these have been sufficient concern to Congress that they have played a role in recent discussions regarding increased funding for research, and have led Congress to request the advice of the Institute of Medicine (IOM) on whether priorities for the allocation of funds at the NIH have been appropriate. Indeed, although the resulting IOM report did not conclude that medical expenditures to date have been allocated inappropriately, it did conclude that NIH should pay greater attention to the burden of illness in assessing research priorities. Building on this, others have suggested that the NIH could better identify the most promising projects if it capitalizes on the formal approaches to assess the burden of illness and opportunities for research to lessen this burden.

VOI is the most well-developed method based in economic theory and may have potential as a practical tool to assist decisionmakers in the process of identifying the value of specific medical technologies and the most promising avenue for future research. In the remaining sections, it is argued that VOI may have the potential to provide important insights into the value of medical research if applied in the right settings with methodological rigor and a thoughtful understanding of its underlying assumptions, strengths, and limitations.

A Review of Value of Information Analyses Applied to Clinical and Policy Questions

VOI has been increasingly used by researchers to inform stakeholders whether additional research would be worthwhile, and to demonstrate the benefits and feasibility of using such analytic methods to inform policy decisions within the timelines demanded by existing procedures. Illustrations of the potential value of the methods include applications to issues of resource allocation in neurology, oncology, and ophthalmology, clinical areas with high burden of disease and/or cost of care. For example, in neurology, VOI decision analytic methods have been utilized to determine the feasibility of magnetic resonance imaging (MRI) as a cost-effective approach to treating multiple sclerosis. The analysis was twofold. It was first determined that the cost of immediate MRI exceeded the cost of the expected value of perfect information. Next, advanced MRI technology had to be shown preferable to the fallback strategy of waiting, given a reasonable estimate of accuracy in MRI. A similar study was completed in orthopedics, utilizing VOI analysis to assess technological advancements in MRI technology to estimate the decision uncertainty that remained after a randomized control trial was completed.

VOI has also been proposed to identify and prioritize medical R&D in a number of clinical areas, where public investment in the later stages of novel therapeutic development has been significant. This interest requires VOI analyses to be calculated for some investment decisions from the public's perspective and with available data and a timely manner consistent with NIH's decision-making process. The idea is to incorporate economic decision analytic tools into trial consideration alongside scientific and trial design criteria to help ensure that public resources are spent efficiently and equitably. For example, in oncology, VOI methods have been analyzed to determine phase III clinical trial research prioritization, feasibility, and areas for greater investment into personalized therapies. Basu and Meltzer's (2007) analysis suggests that identifying cost-effective treatments at the individual level could be greater than 100 times the annual value of identifying the cost-effectiveness treatment on average for the population. In ophthalmology, a VOI analysis was completed to inform governmental health spending and technological priorities. The results of the analysis suggests that the expected value of perfect information (EVPI) could be implemented in a timely fashion to inform the type of research prioritization decisions faced by any healthcare system.

The current reporting standards in the VOI literature is for mean estimates of all stochastic and deterministic model parameters to be described. The uncertainty of the intervention should also be assessed based on the distribution of the incremental costs and incremental quality-adjusted life-years (QALYs) in the cost-effectiveness plane. Additionally, the assumptions underlying the approach should be enumerated. Similar to traditional cost-effectiveness analysis (CEA), typically, main analyses are undertaken with standard assumptions: the discount rate for benefits and costs accrued in the future is 3.5%, and the research findings have a 10 year life span. Sensitivity analysis should be performed to test these assumptions over a range of possible values drawn from the literature. For example, CEAs performed for private insurers tend to use an alternative discount rate that allows for both the timing of costs and revenues and the risk associated with the trial. One commonly used metric is to estimate the adjusted discounted rate based on the capital asset pricing model. The inclusion and exclusion of benefit and costs outcomes and the sources of this information in all analyses should be reported.

A critical practical challenge in the application of VOI methods is that the method has most often been performed by

constructing decision analytic models, which is very time consuming and, therefore, costly. More recently, VOI methods have been developed that use data from existing but often underpowered clinical trials to develop estimates of the value of more definitive trials, or an understanding of the conceptual basis of VOI to bound VOI estimates with even more limited information. Application of practical methods for VOI such as these will continue to be important in developing and validating VOI as a tool to provide timely guidance for decision-making with regard to medical R&D investments.

The promising areas of concern for future methodological advancement in VOI include the following:

1. Individualized care: VOI methods need to be expanded to understand how costs may be better internalized as to capitalize on the value of individualized care, utilizing an expected value of individualized care (EVIC) measure. EVIC is the expected cost of ignorance of patient-level preference heterogeneity and represents the potential value of research that helps to elicit individualized information on heterogeneous parameters, which can be used to make individualized decisions. The heterogeneity parameters of interest are random; hence, rather than larger samples, individualized elicitation will reveal the true values of these parameters. This measure is rather different than the EVPI, in which the parameters of interest have a fixed value in the population. Individualized care offers enormous cost savings, as the value of such may far exceed the value of improved decision-making at the group level; however, such benefits will vary immensely with insurance. EVIC can provide a guide as to when the high value of individualized care may make population-level decision-making especially at risk of providing poor guidance for coverage decisions.
2. Product lifecycle concerns: Despite advancements, uncertainty remains sufficiently high in some potential clinical application areas, hampering VOI calculations, and yet decision-making and research prioritization are required. Analytic methods must evolve further to address significant uncertainty in the potential costs and benefits of novel therapies over the lifecycle of the product. It is believed that questions regarding how risk and uncertainty should be assessed in policy decisions deserves more analytic consideration, because preferences concerning these dimensions are critical to decision-making. Meltzer et al. suggest that it may also be useful to distinguish between uncertainty in insured and uninsured costs in assessing the implications of uncertainty in costs in cost-effectiveness analyses and further characterizations of optimal decision-making when insurance is not complete. Further questions of uncertainty include assessments of the changing value of research due in part to technological or demographic changes.
3. Technological change: Standard CEA and VOI calculations assume a general and uniform rate of technological diffusion across technologies and diseases in clinical practice. However, recent work by Conti, Bernstein and Meltzer (2012) suggests that diffusion patterns of novel molecular-based therapeutics may not follow standard diffusion paths implicit in the standard assumptions of CEA, and may differ substantially from that of new therapies in other clinical areas. Progress on understanding the rate of technological advance across different clinical settings, as well as the product-level, provider-level, and patient-level determinants of this rate, are important inputs for next generation CEA analysis and VOI calculations. *A priori*, replacing standard assumptions with an empirical based model of technological diffusion alters the numerator and denominator of such estimates. How sensitive CEA and VOI calculations are to actual rates of practice that change across a variety of acute, emergent, and chronic disease settings are important subjects for future work.
4. Public versus private investments in research: Finally, future applications of VOI should explore the validity and applicability of the method to help guide decision-making in clinical areas where the public is the main funder of R&D, and also the major source of funding treatment purchases. In the US and other countries, funding for medical R&D, insurance coverage, access to new diagnostic methods and treatment modalities are not shared under the public government budget; rather, the presence of private insurance and private funding for medical R&D challenges the adoption of the social perspective in the widespread use of VOI to guide investment decision-making. VOI can be explored as a tool to guide decision-making in the US, where public monies are a main source of medical R&D, and the main source of insurance coverage and access, once new treatments are developed. The developing world are typically funded by government sources, sometimes in collaboration with experts in public health at the World Health Organization and the Gates Foundation. High profile and sustained gifts from the Gates Foundation in recent years have played an important role in vaccine development successes and in seeding the pipeline for more development in the near future. Recent economic work identifying financing barriers for underwriting R&D in this area have produced novel insights and new approaches to public policy incentives to either 'pull' R&D efforts from the private sector through the credible reward of research activities or to 'push' R&D through direct and indirect underwriting of the perceived costs of R&D and the delivery of vaccines to relevant populations. In this context, the use of VOI methods could be seen as an alternative push mechanism, one that public agencies and public-private partnerships use as a tool to invest funds wisely in the development of new vaccines.

Key Empirical Challenges

A number of empirical challenges are encountered in the practical implementation of VOI, for which practitioners should be aware of, when implementing these methods. First, the most fundamental ambiguity is how to best measure the benefits of a medical intervention. Although disease specific measures such as the number of cancer cases detected or cured may be useful in certain circumstances their effects on mortality (as measured by life years saved) have the advantage of comparability across diseases, they do not capture the important effects of medical care on quality of life. In some

empirical applications, analysts assume for analytical simplicity that quality of life is not a concern so that outcomes may be measured in life-years. This assumption likely provides a lower bound on the total benefits of treatment for some diseases. Yet, there are many clinical examples where including quality of life measurements into a fuller assessment of mortality and morbidity gains could decrease overall benefits of an alternative therapy; for example, if the side effect profile of a treatment that provides mortality gains is quite severe. Recognition of this has led to the development of the concept of QALYs. Using this approach, each year of life is weighted by a factor between 0 and 1, intended to reflect the quality of life in that year, where 0 is equivalent to death and 1 to perfect health. These quality of life weights are most commonly derived by psychometric techniques based on responses to hypothetical choices. Two common approaches to assessment can be fairly readily connected to neoclassical economics; these describe either choices between life with a given illness and a gamble involving life in perfect health and death, with some probability or choices between longer life with illness and a shorter life in full health.

There are also clinical situations where the benefits of alternative therapies that potentially affect morbidity and/or quality of life are not available. For example, quality of life may not be as outcomes in a phase II or phase III trial of novel therapeutic modalities for the treatment of some cancers. In these cases, an extensive literature review of other trials may produce some supportive data. Judgment is required regarding the likely effect of excluding these outcomes on the magnitude and direction of bias introduced into the VOI calculation. Additionally, even when more complete information regarding the impact of treatment on morbidity or quality of life outcomes are available, index QALY weights for these effects may not be available in the published literature for all illnesses and treatment modalities. This hampers the analyst's ability to capture the full range of potential effects of alternative treatments, and also limits the ability of the analyst to compare the full potential benefits of research into one area for research prioritization across alternative uses of supporting funds. When such values are available, it is important to perform a sensitivity analysis over a range of plausible values.

Challenges may also be encountered in the analysis due to the availability of data on the full health costs of alternative treatments. In many settings, treatment costs for standard of care and alternative therapies may not be available from clinical trial data collection efforts, prior studies or estimated using observational data. In addition, innovative therapies that alter the bundle of treatments provided to patients, including the use of diagnostic tests, the length and use of inpatient admissions and physician input, may substantially alter the costs of standard treatment protocols for some illnesses. Validation exercises may need to be performed using actual per person resource use collected on observation data. When the availability of short-term and long-term costs of treatment are lacking, the analyst may choose to ignore costs or perform sensitivity analyses over a range of plausible values. It is important to keep in mind that the dropping of costs from the VOI calculation altogether may be required for analytic convenience, but it limits the comparability of the analysis for research prioritization efforts.

Conclusions

This article provides a review of the rationales behind and the recent practical applications of economic methods to assess priorities in medical research and development. VOI is the most well developed set of tools based in economic theory and advanced cost-effectiveness analysis that could be used by analysts to construct measures of the potential gains from investing in further research. Although these methods have been recently applied to a number of challenging scenarios, the work required to move from what is theoretically possible to the practical application of these principles, to produce valid and reliable estimates of the value of research, involves a series of methodological and empirical challenges. Methodological challenges include the measurement of benefits and costs. Additional issues specific to VOI include developing meaningful priors concerning the parameters of decision models. This may often require extensive review of existing data, primary data collection or even, sometimes, analyses based on a variety of arbitrary priors. It may be difficult to determine priors for the likelihood that the research project will find a meaningful result. Whether it is possible to adequately address these challenges will be resolved through efforts to address these ideas empirically in a number of promising areas.

To apply these approaches to prospectively inform medical research and development decision-making, there are a number of additional and important analytic considerations. These include whether and when typical assumptions of uniform medical technology diffusion rates and discount rates for benefits and costs accrued in the future are justified. Future work in this area needs to empirically grapple with the possibility that the research may be less valuable over time, as other technological or demographic changes can arise that alter the management frequency or natural history of disease and the unpredictability of how the results of research might be useful in areas outside the initial areas of inquiry. These issues imply that the sort of formal analysis suggested here may be more likely used for evaluating clinical research rather than basic preclinical work. Such difficulties suggest that the practical development of VOI for identifying and prioritizing future research is important as one additional tool in the current and evolving armentarium of public research and development decision-making alongside scientific and biostatistical criteria.

Despite these concerns, the importance of making more informed decisions regarding the allocation of resources to medical interventions and medical R&D suggests that work in this area should be an important priority in health economics. It is important to keep in mind – even with evidence that some treatments may have little value at the margin, and with limited evidence of the connection between research and gains in health – health is a domain that people value very highly and at which great strides have been made in recent decades. There is ample reason to believe that such gains may continue in the future. Progress on methods, such as VOI, and the applications for work on the value of medical research and development as a complement to existing methodologies for prospectively evaluating the potential benefits of future investments, have a critical role in ensuring the sustainability of

medical spending and gains in mortality and morbidity that has been conferred by medical science over time.

See also: Information Analysis, Value of

References

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity & individualized care. *Medical Decision Making* **27**(2), 27–112.

Conti, R. M., Bernstein, A. and Meltzer, D. O. (2012). How do initial signals of quality influence the diffusion of new medical products? The case of new cancer treatments. In Grossman, M., Lindgren, B., Kaestner, R. and Bolin, K. (eds.) *Advances in Health Economics and Health Services Research*, vol. 23, pp. 123–148. UK: Emerald Group Publishing Limited.

Conti, R. M., Bernstein, A. C., Villaflor, V. M., et al. (2013). Prevalence of off-label use and spending in 2010 among patent-protected chemotherapies in a population-based cohort of medical oncologists. *Journal of Clinical Oncology* **31**(9), 9–1134.

Murphy, K. M. and Topel, R. H. (2006). The value of health and longevity. *Journal of Political Economy* **14**(No. 5), 871–904.

Philipson, T. and Jena, A. B. (2005). *Who benefits from medical technologies? Estimates of consumer and producer surpluses for HIV/AIDS Drugs*. Article 3, *Forum for Health Economics and Policy*, vol. 0(1), pp. 1005–1005. Berkeley, CA: BE Press.

Further Reading

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5), 941–973.

Bojke, C. K., Sculpher, M. and Palmer, S. (2008). Identifying research priorities: The value of information associated with repeat screening for age-related macular degeneration. *Medical Decision Making* **28**(1), 33–43.

Claxton, K. P. and Posnett, J. (1996). An economic approach to clinical trial design and research priority setting. *Health Economics* **5**, 513–534.

Conti, R., Veenstra, D. L., Armstrong, K., Lesko, L. J. and Grosse, S. D. (2010). Personalized medicine and genomics: Challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities. *Medical Decision Making* **30**(3), 40–328.

Cutler, D. M., McClellan, M. and Newhouse, J. P. (1998). What has increased medical-care spending bought? *The American Economic Review* **88**(2), 132–136.

Groot, K. B., Nikken, J., Oei, E., et al. (2008). Value of information analysis used to determine the necessity of additional research: MR imaging in acute knee trauma as an example. *Radiology* **246**(2), 420–425.

IOM (Institute of Medicine) (1998). *Scientific opportunities and public needs: Improving priority setting and public input at the National Institutes of Health.* Washington DC: National Academy Press.

Meltzer, D., Hoomans, T., Chung, J. W. and Basu, A. (2011). Minimal modeling approaches to value of information analysis for health research. *Medical Decision Making* **31**(6), E1–E22.

Meltzer, D. O. (2001). Addressing uncertainty in medical cost-effectiveness analysis implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *Journal of Health Economics* **20**(1), 109–129.

National Institutes of Health (2005) *Report of the Clinical Trials Working Group of the National Cancer Advisory Board: Restructuring the National Cancer Clinical Trials Enterprise*. Washington, DC: National Cancer Institute. Available at: http://integratedtrials.nci.nih.gov (accessed on July 2012).

Philipson, T., Eber, M., Lakdawalla, D. N., et al. (2012). An analysis of whether higher health care spending in the United States versus Europe is 'worth it' in the case of cancer. *Health Affairs* **31**(4), 75–667.

Value-Based Insurance Design

ME Chernew, Harvard Medical School, Boston, MA, USA
AM Fendrick, Harvard Medical School, Boston, MA, USA, and University of Michigan, Ann Arbor, MI, USA
B Kachniarz, Harvard Medical School, Boston, MA, USA

Introduction

The US healthcare system has widely acknowledged problems with cost, quality, and access. Medical spending is higher than that in any other country, at 17.6% of GDP in 2009, and rising at a rapid rate; such a cost trajectory is unsustainable. Meanwhile, quality is often lacking, and lags behind that of many other nations (Table 1).

Among the most widely used strategies in recent years to address spending has been an increase in patient cost sharing. In addition to shifting the economic burden associated with healthcare from purchasers to patients, economic theory suggests that shifting more financial responsibility onto the patient should reduce wasteful overuse and decrease spending. Unfortunately, patients have been shown to make poor clinical decisions when faced with higher cost sharing by reducing the use of both unnecessary and essential services. Information asymmetry, time-inconsistent preferences, and the impact of marketing and cultural values, all contribute to inefficiencies created by shifting more decision-making power onto the patient. Although such programs inevitably lead to lower spending in the short run, they have been associated with poor adherence and outcomes, increased disparities across socioeconomic groups, and possibly higher long-term spending for some patients.

Value-based insurance design (VBID) was proposed in 2001 as a means to mitigate the negative impact of increased cost sharing and improve the efficiency of our healthcare system. VBID refers to insurance packages that align copays with value, charging patients less for high-value services and more for low value services. By focusing on value, rather than cost alone, VBID aims to improve quality, lower barriers to essential care, and perhaps, if copays are increased for low value services, to save money. VBID programs recognize that different services offer differing amounts of benefit for the money spent. More sophisticated versions can also recognize that value reflects patient traits and can recognize patient heterogeneity.

Most commonly, VBID lowers copayments for high-value services, like diabetes and asthma medications. Cost sharing for low-value services may be increased to help offset program costs and discourage use of low-value services, but this is less common.

Although the concept of VBID is relatively recent, emerging data point to its feasibility and effectiveness. VBID approaches have been successfully adopted by numerous employers with mostly positive clinical results. Both observational data and more systematic controlled analyses support the idea that VBID improves medication adherence and clinical outcomes, and may even lower overall spending. Because its greatest impact is on low-income individuals, it may also help reduce the widespread health disparities seen in the chronically ill.

Rather than a cure-all for our system's problems, VBID is a tool that should be integrated into other innovative approaches, including pay-for-performance (P4P), patient-centered medical homes (PCMH), consumer-driven health plans (CDHPs), and disease management (DM) programs. The ultimate objective of VBID is not to save money, but rather to maximize the health benefit achieved for the money invested; its goal of maximizing value-based limited resources is inherently aligned with that of our entire healthcare system. VBID is growing in popularity among employers, patients, and policymakers. It was included in the Patient Protection and Affordable Care Act (PPACA) and state legislation healthcare reform laws, and has been the subject of plans for new pilot programs within Medicare.

Theory

Consumers seek insurance in order to transfer the unpredictable risk of illness to others and gain access to otherwise unaffordable care. Health insurance helps individuals pool the financial risk of illness. As consumers are risk-averse, they would rather pay a certain fixed premium than risk the possibility of a very high expenditure. In addition, health insurance offers access to treatments that would otherwise not be obtainable. Regular premiums are generally more affordable than the large costs associated with major illness. Thus the benefits of insurance extend beyond risk avoidance to accessibility of care in case of a catastrophic event.

At the same time, insurance introduces moral hazard, which can reduce social welfare. Because the patient's cost of medical treatment is greatly reduced under insurance coverage, patients will utilize more services than they otherwise would. The amount of insurance-induced consumption depends on the price elasticity in demand. The welfare consequences of the extra consumption depend only on the portion of insurance induced over consumption due to the distortion of relative prices. Any extra consumption due to an implicit transfer of income associated with insurance is not a concern. Nevertheless, because insurance distorts prices, it induces greater use of services, necessitates higher premiums, renders health insurance less appealing, and reduces the value of the healthcare system.

Cost sharing has traditionally been used to transfer some risk onto the individual consumer and reduce moral hazard. A price above zero but below the market level allows some risk pooling, while reducing insurance-induced overconsumption. Theory suggests that because of their greater insurance-induced overuse, more elastic services should attract greater cost sharing, whereas inelastic services should be fully reimbursed. Under standard economic theory, because patients utilize services for which their perceived benefit outweighs the

 doi:10.1016/B978-0-12-375678-7.00910-X

Table 1

Setting	*Methods*	*Intervention*	*Conditions*	*Clinical effect*	*Financial effect*
Large firm independently offering DM and VBID options (Gibson *et al.*, 2011a, b)	VBID examined among DM participants and among those who opted out; comparison group consisted of enrollees not offered VBID	Lowered coinsurance from 10–35% tiered structure to 10%	Diabetes	Largest effect seen by combining DM with VBID; drug adherence increased 6.5% over 3 years	Combination of VBID with DM was cost-neutral, with higher drug costs, but lower medical spending
Large employer offering VBID within DM setting (Chernew *et al.*, 2010 Chernew *et al.*, 2008)	Difference-in-differences approach, using a large employer with same DM program as control	Copayments for generic drugs waived, and those for brand-name drugs reduced by 50%	ACEIs, ARBs, beta-blockers, diabetes, statins, and steroids	7–14% reduction in nonadherence for all drug classes except steroids	Possible savings from societal perspective, but less likely for employer. Medical spending would need to drop 9–17% to break even overall, and 29–48% to break even for employer
Pitney Bowes, a large US employer (Choudhry *et al.*, 2010)	Interrupted time series with concurrent control group; commercially insured enrollees with same pharmacy benefit manager were used as controls	Mean copayment for statins decreased by 97%, and that for clopidogrel decreased by 25%	Statins, clopidogrel	Adherence increased by 2.8% for statins; adherence to clopidogrel decreased by 4% in control group, but remained stable in VBID group	
Large pharmaceutical firm (Gibson *et al.*, 2011a, b)	Retrospective, observational with matched comparison group from peer firms	Retail prescription coinsurance lowered from 20% to 10%; Mail-order prescription coinsurance lowered from 10% to 7.5%	Asthma, hypertension, diabetes	Medication adherence rose 5% across all enrollees, and 9.4% for hypertension drugs after 3 years	Entire program was cost-neutral; hypertension patients realized savings of $3700 per enrollee in third year
Over 20 000 enrollees from three CVS Caremark clients who implemented VBID plans (Chang *et al.*, 2010)	Retrospective prepost controlled study	Eliminated insulin and generic drug copayments, and reduced preferred brand copayments from US$30 to US$10–15	Diabetes	Compared to the control group, VBID increased treatment initiation rates by 44% and lowered discontinuation rates by 13–34%. Mean medication possession ratio increased by 4.9% in VBID group, but dropped by 2.3% in control group	
Pitney Bowes, a large US employer (Mahoney 2005, 2008)	Observational with no control	Lowered coinsurance from 25–50% tiered structure to 10%	Asthma, diabetes, and hypertension	Use of medication increased by 35–144%, and ED visits dropped by 22–26%	Pharmacy costs decreased by 7% and overall costs decreased by 6% for diabetes patients; pharmacy and total costs decreased by 19% and 15% respectively for asthma patients

(*Continued*)

Table 1 Continued

Setting	*Methods*	*Intervention*	*Conditions*	*Clinical effect*	*Financial effect*
All adults on cholesterol-lowering therapy (Goldman *et al.*, 2006)	Simulation based on results from retrospective analysis of claims data	Eliminated cost sharing for high- and medium-risk patients, but raised copays from $10 to $22 for low-risk patients	Cholesterol-lowering therapy	Adherence rose by 9–10% in patients with full coverage, but dropped by 8% in low-risk group	No change in total pharmacy spending, with a potential $1 billion in savings from reduced hospitalizations and ED visits nationwide
All Medicare patients (Rosen *et al.*, 2005)	Markov model using literature and Medicare claims data	Full coverage of medications	ACEIs	Assumed adherence increase from 40% to 60%	Program would break even with a 7.2% increase in adherence
Widespread application of VBID to pharmacy benefits and other services (Braithwaite *et al.*, 2010)	Computer simulation	Eliminated cost sharing for high-value and increasing cost sharing for low-value services	All services	VBID would increase benefit of healthcare from 4.7 life-years to 4.73 life-years if applied to drug benefits, and 4.95 life-years if applied to all other services as well	Assumed to be cost-neutral
Medicare beneficiaries (Choudhry *et al.*, 2008)	Markov cost-effectiveness model	Provided full coverage drug therapy to postmyocardial infarction patients	Aspirin, beta-blocker, ACEI or ARB, and statin	Survival after initial event increased from 8.21 QALYs to 8.56 QALYs	Reduced overall spending by $2500 per patient, with cost-savings even if assuming only a 1% increase in adherence. Program was cost effective, but not cost-saving from Medicare perspective.

Abbreviations: ACEI, angiotensin-converting-enzyme inhibitor; ARB, angiotensin receptor blocker; DM, disease management; ED, emergency department; VBID, value-based insurance design.

cost, higher price will reduce consumption disproportionately for low-value services, which will preserve value in the healthcare system.

Across-the-board copayment increases typically do not take into account differences in benefit of different treatments. The expected benefit of a therapy should be inversely proportional to the elasticity of demand, and thus copayment requirements prevail. An essential medical service should have an inelastic demand, and be covered fully by insurance. Most traditional benefit plans only take into account the cost, not benefit, of a service when determining the degree of cost sharing. Such plans also fail to appreciate patient heterogeneity; a single service might have a different benefit and elasticity for different patients depending on the clinical diagnosis. For example, beta-blocker therapy may play a vital role in the management of heart failure patients, or be used more electively in the treatment of anxiety. Differences in risk and outcome preferences among patients further contribute to patient heterogeneity, rendering indiscriminate changes in copayments suboptimal because too much risk is being transferred for inelastic services.

Yet, if patients do not make optimal decisions, placing more risk and decision-making power in their hands may have detrimental effects. As patients are risk-averse, such increased cost sharing reduces the value of the insurance plan. In addition, patients often misjudge the costs and benefits of medical therapies, and make poor clinical decisions. Inherent information asymmetry between patient and provider may lead to underuse of essential services and suboptimal resource allocation. The physician has limited information regarding patient preferences, values, and history. Likewise, patients often fail to fully understand medical information, or are otherwise influenced by external biases like marketing. They lack the clinical training to completely comprehend underlying principles and make objective decisions. Physicians typically have years of experience and better ability to predict disease progression. Similarly, patients' time-inconsistent preferences may bias their clinical decision-making. Individuals tend to undervalue future benefits and prospective cost savings. These phenomena contribute to underutilization of valuable services, increased overall medical spending, and poor outcomes, particularly in chronic disease patients like diabetics or asthmatics.

Higher copayments further tend to have a greater impact on low-income patients, contributing to healthcare disparities. These populations already face significant barriers to essential care, which are only exacerbated by increased cost sharing. Patients with higher education and better understanding of their care are also more likely to make better clinical decisions. The increased responsibility and financial risk associated with increasing copayments places an unnecessary burden on less affluent populations, and tends to preferentially worsen their health outcomes

VBID addresses these problems using a 'clinically sensitive' approach to align financial incentives with value in the healthcare system. It recognizes that if decision-making is flawed, the amount of cost sharing should depend not only on the cost, but also on the evidence-based benefit of a therapy.

Defining value in the context of VBID programs is complex. Conceptually, value relates to the cost effectiveness of a given service for a given patient (health gained per dollar spent). The growing emphasis on cost effectiveness and comparative effectiveness research can support efforts to assess value and implement VBID. But it is unlikely that evidence will be detailed enough to be tailored to specific patients so most VBID programs will be applied on average for groups of patient remain. The crucial assumption is that fully informed consumers, given the economically efficient income transfer associated with insurance, would purchase these services even if they faced the true prices. In cases of high-value (e.g., highly cost effective) services, increased consumption does not represent moral hazard, and thus should not be financially discouraged. This principle is also in line with standard economic theory. Essential high-value care should have price inelastic demand (once implicit income transfers associated with insurance are taken into account), and thus little or no required cost sharing.

VBID ultimately employs evidence-based medicine to reinforce the financial incentives of using high-value care. It addresses the inherent information gap between patient and provider, and may even offer benefits beyond those of patient education for underused services. Lower cost sharing would increase consumption, improve health outcomes, and possibly even reduce long-term healthcare costs. Through lower copays, a value-based benefits design not only encourages optimal utilization of cost-effective services, but also offers a greater degree of risk protection to the consumer.

The overall financial profile of VBID initiatives can be favorable, particularly if cost sharing is increased for low-value services, and depends largely on the disease state being targeted. Services that have elastic demand and reliably prevent expensive complications that are highly likely to develop otherwise, tend to be best candidates for copay reductions. Increased cost sharing for such services might reduce short-term spending the form of lower utilization, but will likely accrue higher long-term costs through increased complications. Conversely, VBID may save money and improve outcomes in such cases.

Although VBID may reduce aggregate healthcare spending by avoiding expensive exacerbations and complications, the financial impact on the employer is less obvious. Employers face increased initial spending due to more generous coverage of high-value services, and greater demand for those services due to improved adherence. Employers take on increased medication costs, and might not be able to reap the savings if they have high employee turn-over rates. As an example, promoting the use of statins will likely increase short-term spending for the employer. A significant part of the savings in the form of avoided complications might go to Medicare once the patient retires. As long as health insurance is largely employer-based, there will be an inherent divide between employer healthcare spending and aggregate spending on a population level.

Nonetheless, some of these employer costs may be offset by savings on other medical spending, such as hospital or emergency department visits. Increased productivity, employee satisfaction, and decreased disability also contribute significant value to the employer. The return on investment will largely depend on the degree and accuracy of patient targeting. Programs that offer copayment reductions for very specific

patient populations and specific medications will tend to have more attractive financial profiles. Although not a major part of VBID, increased cost sharing for other, preferably low-value services, may further help reduce implementation costs. Above all, it is vital for employers to appreciate the often overlooked value of improved productivity and lower disability. VBID will offer greatest financial benefits where the patient population is responsive to changes in cost sharing and expensive complications may be reliably prevented using cheap medications. This is true of many chronic diseases, such as diabetes or asthma, which have been the first targets of VBID programs. VBID principles may be extended to other high-value therapies as well. Nonetheless, it is important to recognize that increasing the use of cost-effective services will not in itself be cost-saving. The chief benefit will be improved efficiency and value of the healthcare system, with reductions in spending possibly requiring increased cost sharing for low-value services.

By lowering copayments for high-value services, VBID may support a number of other health system reforms, including pay-for-performance, patient-centered medical homes, and CDHPs. Both P4P and PCMHs are supply-side interventions that encourage evidence-based medicine and improve access to high-value care. These programs allow clinicians to claim a portion of savings from reduced medical spending and offer financial rewards for improving outcomes. Similarly, VBID offers patients a financial incentive to pursue lower-cost higher-value therapies in the form of reduced cost sharing. VBID naturally complements P4P and PCMHs by aligning patient and provider incentives. DM programs would likewise benefit from value-based benefit designs. DM utilizes a variety of strategies, including patient education and coaching, to encourage high-value care. Patients are often given easier access to doctor visits and relevant medications. Reducing copayments for these drugs naturally complements DM initiatives by reducing the financial barriers to care. Like VBID, CDHPs emphasize consumer incentives to improve value and curtail costs. However, CDHPs significantly increase patient cost sharing for all services below the deductible with likely reductions in the use of both low-value and essential care. Implemented together, these programs would promote cost-conscious decision-making while encouraging use of high-value care. For example, the use of 'VBID waivers' for certain services would mitigate the negative impact of higher cost sharing in CDHPs. Applying VBID principles to subsidize high-value services would increase their use and improve efficiency within the healthcare system.

Although drug benefits are a very natural application of VBID, the concept may be extended to other health services. As an example, it has been proposed that the field of oncology would be a natural candidate for VBID implementation for several reasons. Different therapies will offer varying degrees of benefit for patients; whereas some treatments add years of life, some might only extend survival by a few weeks. The benefit of one drug may also depend on the diagnosis. Although the same chemotherapy or radiation might be used for many different cancer types, some will be more responsive to the therapy than others. Finally, the expected value of a treatment often depends on the particular patient. Biomarkers may be used, as in the case of breast cancer, to identify patients likely to respond to certain therapies. Oncology would naturally benefit from evidence-based targeting to encourage use of high-value services. In parallel, gastroenterology has also been proposed as a possible target for VBID. For example, a colonoscopy will rather have different value for a high-risk or elder patient, than for a young patient seeking the same procedure. The basic principles of targeting and adjusting copayments according to value of a service may be applied in a variety of clinical situations, ranging from drug benefit design to oncology and gastroenterology.

Practitioners

Over the past decade, VBID has grown steadily in popularity, and by one estimate is currently utilized in some form by 20–30% of large employers. It has garnered significant support for its adaptability, depending on employer goals and the patient population. The basic principles of VBID may be applied for any balance between improved employee health and reduced spending. In practice, it is impossible to achieve perfect targeting and evaluate the value of each service for every patient. A balance must be struck between program effectiveness and feasibility, often limited by availability of evidence-based data, accurate assessment of patient's clinical condition, and health information technology. To address these issues, several approaches for implementing VBID have been used that target patients based on service, condition, condition severity, participation in other health programs, or a combination of them.

One approach is to simply reduce cost sharing for certain drugs and services that are deemed to be of high value. All employees would face the same copayments, irrespective of clinical diagnosis or use of the therapy. Pitney Bowes and Marriot have adopted such a solution for diabetes, hypertension, and asthma medications. Pitney Bowes was among the most widely celebrated employers of VBID; although there was no external control, it reported $1 million in savings after introduction of the program in 2002. Most importantly, the program has received widespread attention and has demonstrated that VBID is feasible and may be effective.

Another possibility is to target a specific patient population and offer reduced copayments for high-value evidence-based treatment. Patients with a specific condition would be eligible to participate and receive free or subsidized care. Such programs typically target chronic diseases with known evidence-based therapies, including cancer, cardiovascular disease, obesity, respiratory conditions, and diabetes. The University of Michigan, MI, USA, is among the first to utilize this approach. All employees with diabetes are eligible to enroll and receive subsidized insulin, beta-blockers, diuretics, and other high-value medications. Started in 2009, the University of Michigan Focus on Diabetes Program is the first prospective controlled trial of VBID, and will likely shed light on its effects on outcomes and spending. The city of Asheville in North Carolina and United HealthCare have used a similar approach to target diabetes.

Less commonly used approaches include targeting high-risk patients either eligible for, or actively enrolled in, a DM program. These patients would likewise receive reduced or waived copays for certain medication classes. This design is

offered by WellPoint, although it has not been widely adopted by its clients; Gulfstream offers subsidies for utilizing providers that meet certain evidence-based care criteria. Other major providers using VBID include Caterpillar, Service Employees International Union, Mid-America Coalition on Health Care, and Health Alliance Medical Plans (HAMP). Each company targets different combinations of chronic diseases, depending on the employee population and claims data. Employers may fine-tune their VBID implementation to reach a desired level of medical costs and employee health. Many providers had incorporated VBID into more comprehensive novel healthcare delivery systems. Hannaford Brothers, for instance, combined VBID targeting certain diseases and minimally invasive surgical procedures, with promoting healthy lifestyle habits and better information technology. All available results point to improved drug adherence and outcomes, especially for diabetes and hyperlipidemia.

Finally, there are increasing calls on the government to promote the adoption of VBID, and include it in any healthcare reform laws. The American Academy of Actuaries recognized the importance of VBID and recommended that any new legislation do not discourage its implementation. The government should also continue investing in comparative effectiveness research (CER) and health information technology (HIT) as means to improve value and clinical outcomes in the healthcare system. Unlike many other interventions, VBID has garnered bipartisan support, with 73% of healthcare leaders generally in favor of its adoption. VBID has the benefit of offering important financial incentives without limiting patient choice. By targeting primarily high-value services, patients are encouraged to pursue valuable care, but are given the freedom to access other services as well. In addition, it avoids placing physicians into the role of healthcare gatekeepers, and maintains low administrative barriers to care. VBID has gained support from patients, providers, and payers, by aligning their incentives.

Accordingly, VBID has gained much attention among policymakers. The PPACA includes language permitting the use of VBID for high-value preventive services, such as immunizations and screenings (2010). As required by the new law, the Department of Health and Human Services has devised a National Strategy for Quality Improvement in Health Care, which promotes the use of VBID models at the federal level (2011). Similarly, there has been interest in using value-based principles to improve the financial profile of Medicare. There have been proposals to introduce a VBID pilot program for Medicare to evaluate its effectiveness. More recently, the Medicare Payment Advisory Commission (MedPAC) report to Congress has underscored the importance of using VBID to steer patients toward higher-value services. President Obama has likewise included VBID in his Deficit Reduction Plan, and has called for vesting the Independent Payment Advisory Board with power to promote value-based benefit designs (2011).

Empirical

Patient cost sharing has been steadily on the rise over recent years, in an attempt to curtail growing healthcare spending. Between 2000 and 2009, the average generic, preferred, and nonpreferred prescription drug cost sharing increased by 25%, 80%, and 59%, respectively. Paradoxically, copayments have also risen both within DM programs and for services used as quality indicators. As DM programs implement innovative approaches to encourage use of essential services and improve adherence, rising copays discourage the consumption of those same therapies. Likewise, services accepted as high-value and used as quality measures for hospitals often lack demand-side financial incentives. Indicators contained within the Health Plan Effectiveness Data and Information Set (HEDIS) have suffered increases in copayments similar to that of other services. HEDIS is widely used to evaluate health plan performance and includes measures such as receipt of beta-blockers after a heart attack and treatment of asthma. Higher cost sharing for HEDIS services may lower health plan performance.

Further, there are significant data demonstrating the effects of greater cost sharing on adherence and outcomes. Many studies have shown that patients tend to indiscriminately cut use of both essential and low-value services when faced with greater copayments. Even vital medications, like those used for hyperlipidemia, rheumatoid arthritis, diabetes, and asthma, suffer the effects of increased cost sharing. Doubling of copayments for diabetes and hypertension drugs has decreased medication use by 23% and 10%, respectively. This is particularly alarming, considering that many crucial services are widely underutilized by patients. The resulting decreased adherence often leads to poorer outcomes and higher rates of complications. This has been particularly evident in the case of asthma, diabetes, and cardiovascular disease. Under some circumstances, increased cost sharing may actually raise long-term costs by increasing the incidence of expensive and preventable complications. Importantly, the effects of increased cost sharing have the greatest impact on low-income patients. These patients are more likely to delay treatment due to cost concerns; for example, higher out-of-pocket expenses are also associated with more frequent asthma exacerbations in children of low-income families. The higher incidence of chronic illnesses like diabetes and asthma, combined with the effects of increased copayments, contributes to worse outcomes in these populations with resulting greater health disparities.

There is less evidence regarding the effects of lowering copayments, which is the principle instrument of VBID. The impact of lower cost sharing might be different than that of higher cost sharing because of psychological phenomena, but in practice, it is generally similar in magnitude. Some indirect data had come from the introduction of Medicare Part D, which lowered out-of-pocket drug spending for seniors. There was a 3–13% increase in medication use, with the opposite effect seen in the coverage gap. This corresponded to improved outcomes and a 4.1% reduction in hospitalizations relating to diabetes and several cardiovascular and pulmonary conditions. The effects of Medicare Part D were largest for patients with previously high copays or no coverage. In this subgroup, savings on medical expenditures generally offset increases in drug costs. In other settings, it has been observed that lower cost sharing for diabetes patients is associated with better adherence and better glycemic control, as measured by the degree of hemoglobin glycation. Fixed-effects modeling further suggests that lower medication copayments may lead to

higher pharmacy benefit costs, but significant overall savings in congestive heart failure, hypertension, diabetes, and dyslipidemia patients.

Direct evidence on the impact of VBID programs is relatively recent; the concept of VBID is fairly new, and it takes years to see long-term outcome and spending effects. The data are also more heterogeneous because the programs' impact largely depends on the particular implementation. Results that are not peer-reviewed and lack a control group suggest that the experience has been generally positive. Pitney Bowes had boasted of one of the first widely celebrated programs. Caterpillar, Hannaford Brothers Company, United Healthcare, and others, reported similarly improved outcomes with no change in, or reduced, spending. Controlled studies and more systematic analyses are fewer, but offer important, often less positive insights into the consequences of VBID programs.

Earliest data on the effects of VBID come from several mathematical models. A very broad implementation targeting various high-value services throughout the healthcare system would confer an additional 5–9% health benefit, as measured in life-years, without increasing overall or out-of-pocket spending. Better targeting of high-value therapies, such as angiotensin-converting enzyme (ACE) inhibitors or cholesterol-lowering drugs, offers even more advantages. Simulation analysis suggests that eliminating cost sharing for ACE inhibitors for Medicare patients with diabetes would both improve outcomes and lower costs by up to US$1600 per patient. Adjusting copayments for cholesterol-lowering therapy based on the patient's risk level would offer similar benefits. Thousands of hospitalizations and emergency department visits would be avoided, with over US$1 billion in annual aggregate savings. These simulations are particularly sensitive to estimates of the impact of lower cost sharing on adherence. Nonetheless, even with conservative assumptions, VBID is expected to confer clinical benefit with little change in spending.

More recently, there have been emerging data from employers implementing VBID principles. Analyses of two large firms with VBID options had demonstrated improved adherence and outcomes, with potentially neutral effects on aggregate spending. The financial impact on the employers was somewhat less favorable, but some of the cost could be offset by improved employee satisfaction and productivity. Both the State of Maine and the City of Springfield in Oregon had initiated pilot programs that targeted diabetes patients. In addition to waiving copayments for drugs and physician visits, the latter program also provided free individualized pharmacist consultations. Compared to randomly chosen controls, patients in the intervention group in each case had improved medication adherence and had better glycemic control. Sick leave had declined and productivity had improved for both programs. Although Maine reported significant savings, Springfield's healthcare costs had actually increased. Nonetheless, it is possible that savings in the form of employee productivity and reduced disability have helped offset any program costs, and long-term savings are likely to accrue beyond the timeframe of the initial study. Both pilot programs were considered a success, with VBID options becoming more widely available soon thereafter. Overall, these data demonstrate the varied consequences of VBID implementation. The financial profile will largely depend on the level of targeting, patient population, and other employer-specific parameters. Nonetheless, there is consistent evidence that VBID improves clinical outcomes and value of the healthcare system; by promoting the use of high-value services, it offers improved employee health and productivity. These indirect benefits are often overlooked in cost-effectiveness analyses.

VBID has also been applied in the context of DM programs and patient-centered medical homes, with very positive results. Although both DM programs and VBID improve medication adherence, a combination of the two strategies offers further benefits. Within a single DM program, VBID had increased medication adherence by 7–14% for statins, ACE inhibitors, beta-blockers, and diabetes medications. In a separate study, a combination of DM and VBID to target diabetes patients had proved to be cost-saving and improved drug use by almost 7%; these effects were significantly better than controls in either program alone. Similarly, several employers have combined VBID with PCMHs. Among others, the City of Battle Creek in Michigan and the State of Minnesota have reported positive results using this approach. Various performance measures have greatly increased by at least 20–35% following introduction of the programs. Patients have received more preventive care, and have avoided both expensive hospitalizations and emergency department visits. Blood pressure, glycemic control, and cholesterol levels have improved by 5.7–22%. Other examples further reinforce the benefits of integrating VBID into innovative payment reform approaches. Although effective on its own, VBID may be easily and effectively combined with other strategies.

Conclusion

As the persistently growing healthcare spending is addressed, it is important to maintain a focus on value and not cost alone. The purpose of healthcare is not to save money, but to provide the greatest health benefit given limited resources. Limiting access to essential care might save money, at least in the short term, but is not socially desirable. Further, focusing on shortsighted interventions like indiscriminate increases in copays may actually have opposite effects on spending in the long term. Curtailing spending should not be at the expense of reducing essential high-value care. VBID is an important approach that aims to improve the value of the healthcare system, as well as reduce barriers to essential care and health disparities.

Nonetheless, there are some challenges that lay ahead of a more widespread acceptance of VBID. Patients might have concerns of privacy and fairness. Different patients might be charged different fees for the same service. Some of the patients' clinical data are also used for benefit design. Nonetheless, most of these issues may be addressed through patient education and careful program design. Another major challenge to VBID implementation is a lack of CER and HIT infrastructure. VBID relies on CER to identify high-value services; there are currently few studies that compare the effectiveness of competing therapies. Likewise, HIT is necessary to incorporate the data from CER into benefit design. Nonetheless, there are known high-value therapies for the treatment

of many chronic diseases, and HIT is adequate for basic targeting. In these areas, VBID could be implemented successfully. As there iscontinuous expansion in targeting capabilities, opportunities for VBID will expand.

Employers have also been cautious in adopting VBID because of its somewhat uncertain financial impact. The return on investment profile of VBID largely depends on the particular implementation. Employers can improve the financial profile of their program by finer targeting. Reducing copays for a smaller group of high-risk patients is more likely to reduce program costs. Employers may also choose to raise copays for all other services, or preferentially target low-value services, though it may be difficult to identify low-value services based on easily identifiable patient characteristics. Few treatments are low-value for entire patient groups, requiring more sophisticated targeting and incorporation of clinical judgment. Another concern of VBID is that such plans might preferentially attract sickest patients who would receive lower copays for high-value services targeted at chronic disease.

The success of VBID will require a new mindset of simply embracing value over costs. After allocating the most efficient amount of resources to healthcare, the health benefit to patients will need to be maximized. This will also call for more comprehensive ways of assessing costs, benefits, and value than the often shortsighted methods being used today. Such a new approach to insurance design will require an integration of clinical medicine, economics, and actuarial analysis. The feasibility of VBID will grow with continued investment in CER and HIT. As new data become available on the relative value of services, it will be crucial to align financial incentives with the highest-value options. The recent growth of HIT will also make better benefit design feasible. As communication between patients, providers, and insurance benefit managers improves, better plans that align incentives between the parties will become a reality.

Although not a sole remedy for the healthcare system's issues of cost, quality, and access, VBID is a powerful tool that aligns financial incentives for patients with evidence-based medicine. By adjusting cost sharing based on the value of a service in the context of a particular clinical situation, VBID can facilitate more efficient resource allocation and better health outcomes. Integrated with other approaches, including DM, CDHPs, and patient-centered medical homes, VBID may prove to be a powerful tool for tackling major health care reform in the coming years.

See also: Access and Health Insurance. Adoption of New Technologies, Using Economic Evaluation. Cost-Effectiveness Modeling Using Health State Utility Values. Cost–Value Analysis. Demand Cross Elasticities and 'Offset Effects'. Demand for and Welfare Implications of Health Insurance, Theory of. Disability-Adjusted Life Years. Efficiency in Health Care, Concepts of. Health and Health Care, Need for. Health and Its Value: Overview. Health Insurance and Health. Health Insurance in Historical Perspective, I: Foundations of Historical Analysis. Health Insurance in Historical Perspective, II: The Rise of Market-Oriented Health Policy and Healthcare. Incorporation of Concerns for Fairness in Economic Evaluation of Health Programs: Overview. Managed Care. Measurement Properties of Valuation Techniques. Medical Decision Making and Demand. Moral Hazard. Multiattribute Utility Instruments and Their Use. Multiattribute Utility Instruments: Condition-Specific Versions. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. Prescription Drug Cost Sharing, Effects of. Price Elasticity of Demand for Medical Care: The Evidence since the RAND Health Insurance Experiment. Private Insurance System Concerns. Quality-Adjusted Life-Years. Rationing of Demand. Resource Allocation Funding Formulae, Efficiency of. Supplementary Private Insurance in National Systems and the USA. Time Preference and Discounting. Utilities for Health States: Whom to Ask. Value of Drugs in Practice. Welfarism and Extra-Welfarism. Willingness to Pay for Health

References

Braithwaite, R. S., Omokaro, C., Justice, A. C., Nucifora, K. and Roberts, M. S. (2010). Can broader diffusion of value-based insurance design increase benefits from US health care without increasing costs? Evidence from a computer simulation model. *PLoS Medicine* **7**, e1000234.

Chang, A., Liberman, J. N., Coulen, C., Berger, J. E. and Brennan, T. A. (2010). Value-based insurance design and antidiabetic medication adherence. *American Journal of Pharmacy Benefits* **2**, 39–44.

Chernew, M. E., Juster, I. A., Shah, M., et al. (2010). Evidence that value-based insurance can be effective. *Health Affairs* **29**, 530–536.

Chernew, M. E., Shah, M. R., Wegh, A., et al. (2008). Impact of decreasing copayments on medication adherence within a disease management environment. *Health Affairs* **27**, 103–112.

Choudhry, N. K., Fischer, M. A., Avorn, J., et al. (2010). At Pitney Bowes, value-based insurance design cut copayments and increased drug adherence. *Health Affairs* **29**, 1995–2001.

Choudhry, N. K., Patrick, A. R., Antman, E. M., Avorn, J. and Shrank, W. H. (2008). Cost-effectiveness of providing full drug coverage to increase medication adherence in post-myocardial infarction Medicare Beneficiaries. *Circulation* **117**, 1261–1268.

Gibson, T. B., Mahoney, J., Ranghell, K., Cherney, B. J. and McElwee, N. (2011a). Value-based insurance plus disease management increased medication use and produced savings. *Health Affairs* **30**, 100–108.

Gibson, T. B., Wang, S., Kelly, E., et al. (2011b). A value-based insurance design program at a large company boosted medication adherence for employees with chronic illnesses. *Health Affairs* **30**, 109–117.

Goldman, D. P., Joyce, G. F. and Karaca-Mandic, P. (2006). Varying pharmacy benefits with clinical status: The case of cholesterol-lowering therapy. *American Journal of Managed Care* **12**, 21–28.

Mahoney, J. J. (2005). Reducing patient drug acquisition costs can lower diabetes health claims. *American Journal of Managed Care* **11**, S170–S176.

Mahoney, J. J. (2008). Value-based benefit design: Using a predictive modeling approach to improve compliance. *Journal of Managed Care Pharmacy* **14**, S3–S8.

Rosen, A. B., Hamel, M. B., Weinstein, M. C., et al. (2005). Cost-effectiveness of full medicare coverage of angiotensin-converting enzyme inhibitors for beneficiaries with diabetes. *Annals of Internal Medicine* **143**, 89–99.

Valuing Health States, Techniques for

JA Salomon, Harvard School of Public Health, Boston, MA, USA

Introduction

Cost-effectiveness analyses of health interventions and policies are often conducted using quality-adjusted life-years (QALYs) as the metric for quantifying health outcomes. A related metric called disability-adjusted life years (DALYs) has been used to assess the burden of disease attributable to different causes as well as in cost-effectiveness analyses, especially those in low- and middle-income settings. Both QALYs and DALYs provide summary measures of health outcomes that (1) combine information on survivorship and the health experience among the living; and (2) accommodate comparisons across diverse types of health problems by expressing outcomes in a common 'currency'. A critical feature of DALYs and QALYs is that they attach weights to time spent in different states of health to reflect the relative severity of these outcomes. These weights have been called, among other names, 'health state valuations,' 'health-related quality of life weights,' and 'health utilities,' with some associated variation in the interpretation of the meaning of the weights; in this article, these will be referred to as 'health state valuations.' Health state valuations are given on a scale that ranges from 0 to 1.0. For QALYs, 1.0 implies a state of optimal health and 0 implies a state equivalent to being dead, whereas for DALYs, the scale is reversed: 0 implies no health loss, whereas 1.0 implies severity equivalent to being dead. For use in QALY and DALY calculations, health state valuations must have interval scale properties, that is, differences between two values on the scale must be meaningful, with a given distance between two scale values having the same significance, no matter where the points are located on the scale. For instance, the difference between 0.4 and 0.6 must be understood as equal to the difference between 0.7 and 0.9.

Various methodological and empirical issues relating to health state valuations have inspired a rich and growing literature. The focus of this article is on techniques for eliciting valuations. Where other relevant topics are mentioned, cross-references are provided to those articles in which these topics are treated in greater detail.

Overview of Techniques

There are six types of techniques that have been used prominently in eliciting health state valuations.

Standard Gamble

The standard gamble is a method that has its theoretical basis in the von Neumann–Morgenstern axioms of expected utility theory. It aims at measuring the 'disutility' of a health state by observing the willingness to accept a certain risk of death in order to avoid the state. In a typical framing of the standard gamble, a respondent is asked to consider a choice between two alternatives. In alternative A, the person would live with a particular health problem (the one for which the valuation is needed) with certainty, for the remainder of his or her life. Alternative B is usually characterized as a risky treatment, with two possible outcomes: life in a state of optimal health, with probability p, or immediate death, with probability $(1-p)$. The measurement objective for the standard gamble is to identify the probability of optimal health, p, at which the respondent is 'indifferent' between alternatives A and B, in other words, the point at which the two alternatives seem equally attractive. Once this indifference point is identified, a health state valuation for the particular health problem of interest is equal to p. The logic of this inference derives from setting the utility of optimal health to 1.0 and that of death to 0 and assuming that at the point of indifference, the respondent considers the expected utility of alternatives A and B to be the same. In mathematical terms, the equality is stated as $p \times \mathrm{U(optimal)} + (1-p) \times \mathrm{U(death)} = \mathrm{U(health\ \ outcome)}$, or $p \times 1.0 + (1-p) \times 0 = \mathrm{U(health\ outcome)}$, which simplifies finally to $\mathrm{U(health\ outcome)} = p$.

Time Trade-Off

The time trade-off is another of the most widely used methods for eliciting health state valuations. Like the standard gamble, it invokes the notion of willingness to sacrifice something that is valued in order to avoid an inferior health state. In the standard gamble, what is sacrificed is the certainty of survival, whereas in the time trade-off, what is sacrificed is the length of life. The time trade-off asks respondents to consider a choice between two alternatives. The first is to survive for a specified amount of time, t_1, with a particular health problem, followed by death. Different time trade-off studies have taken different approaches to defining t_1, including using an arbitrary duration such as 10 years or using the respondent's estimated life expectancy (or some rough approximation to this). The second alternative in the time trade-off is to survive a (presumably) shorter amount of time, t_2, but in optimal health. The measurement approach in the time trade-off is usually to hold t_1 constant and vary the amount of time t_2 until the indifference point is identified. A health state valuation may then be computed as the ratio t_2/t_1. The logic of this inference, similarly to the standard gamble, is to equate the overall value between the two alternatives at the respondent's indifference point. In case of the time trade-off, the value of an alternative is taken to be the product of its health valuation and its duration. Thus, the indifference point implies the equality $\mathrm{U(health\ outcome)} \times t_1 = \mathrm{U(optimal\ health)} \times t_2$. Again taking the valuation of optimal health to be 1.0, this simplifies to $\mathrm{U(health\ outcome)} = t_2/t_1$.

 doi:10.1016/B978-0-12-375678-7.00502-2

Rating Scale

The rating scale approach comes from psychometrics. In contrast to both the standard gamble and time trade-off, it consists in eliciting a numerical valuation for a health outcome directly, without invoking the notion of sacrifice. A series of health outcomes are often simultaneously located on a numerical scale such that a respondent evaluating outcomes A, B, and C must consider whether A is preferable to B, B preferable to C, and A preferable to C, and also to decide the strength of these preferences, in other words the distances between them on the numerical scale. A number of different ways of operationalizing a rating scale are possible but most feature a straight line with marked intervals (like a meter stick), with the endpoints marked with numbers (e.g., 1.0 and 0 or 100 and 0) and labels referring to the best outcome (e.g., as 'perfect health' or 'best imaginable health state') and the worst outcome (e.g., as 'worst imaginable health state' or 'dead'). Sometimes the marked intervals are accompanied by numerical labels. Rating scales can also be constructed without marked intervals, in which case they are called 'visual analogue scales.' In practice, the latter term is sometimes used in a generic way that includes both marked and unmarked scales. There has also been variation in practice concerning the range of the scale. If researchers want to accommodate states that are regarded as worse than being dead, then the scale spans from the best to the worst imaginable outcomes, and respondents are asked to locate 'dead' on the scale amidst one or more nonfatal outcomes. Issues around states regarded as worse than being dead are mentioned in the Section on Key Conceptual and Methodological Issues. Typically, health state valuations are derived from rating scales by taking the ratio of the distance between a particular state and the point on the scale assigned to 'dead', divided by the distance between the upper endpoint of the scale (the best outcome) and the point assigned to 'dead.'

Magnitude Estimation

Another technique that arose, like the rating scale, from the direct measurement tradition of psychometrics is magnitude estimation (sometimes called 'ratio scaling'). In this approach, a respondent is given one health state as a reference benchmark, and then asked to indicate how many times better or worse some other states are compared with the reference state. Sometimes, the reference state has been defined as an endpoint of the scale (e.g., the most desirable outcome), although other studies have chosen an intermediate state as the reference. For example, a seminal magnitude estimation study by Patrick *et al.* (1973) anchored comparisons to a reference item describing 'a day in the life of a person who was as healthy as possible on that day,' which was assigned an arbitrary score of 1000. Other days were to be scored in relation to this reference, for instance, the instructions noted that a day that was regarded as 'half as desirable as the standard' should be scored at 500. Based on this scheme, results may be rescaled to the unit interval simply by dividing the scores by 1000. In fact, by translating the ratios into scores in this way, the operationalization of the task comes to bear a strong resemblance to the rating scale. Another influential magnitude estimation study by Rosser and Kind (1978), anchored the comparison task with the second best outcome (no disability and mild distress) as the reference, and asked respondents to indicate how many times worse other states were compared with this reference. In this case, rescaling of the results depended on normalizing the scale, so that the best state had a value of 1.0 and death a value of 0. Thus, if death were regarded as 200 times worse than the reference state, this would result in the reference state having a value of $(1-(1/200))=0.995$; a state that was considered 10 times as bad as the reference state would then have a value of $(1-10\times(1-0.995))=0.95$.

Person Trade-Off

The person trade-off is a technique that has been used less commonly than many of the other techniques mentioned so far. Unlike these other techniques, the person trade-off asks respondents to answer from the perspective of a social decision maker considering alternative policy choices rather than as an individual making choices for himself or herself. The person trade-off has been framed in various ways, but a typical presentation asks respondents to consider two options, one that will result in longer survivorship for a group of people and the other that will result in prevention of a nonfatal, usually chronic, condition. For example, a respondent might be asked to weigh an option that would prevent $x_1=1000$ deaths in a healthy population versus an alternative that would prevent $x_2=5000$ cases of some particular chronic disease outcome. The measurement approach in the person trade-off would usually be to hold constant the number of averted deaths (x_1) and vary the number of nonfatal outcomes averted (x_2) to find the indifference point between the alternatives. A health state valuation for the nonfatal outcome being considered would then be computed as $(1-(x_1/x_2))$. For instance, $x_2=10\,000$ would yield a value of 0.9.

Ordinal Response Methods

Finally, there has been renewed interest recently in ordinal response methods. Over much of the history of measuring health state valuations, ordinal methods such as ranking have been deployed primarily as a 'warm-up' exercise, for example, as a preliminary step to eliciting rating scale values for a range of health states. However, there have been a number of examples of analyzing ordinal response data in order to infer latent cardinal values that are consistent with these responses, and these examples have grown numerous over the past several years. Methods for collecting ordinal responses fall into two main categories: (1) rank ordering of health states and (2) paired comparisons of health states, residing within the broader methodological tradition of discrete choice analysis. Analysis of ordinal response information has been based largely on a random utility framework operationalized using regression models for discrete outcomes. These models are based on the presumption that ordinal responses may be related to differences between values on an unobserved cardinal scale. Specifically, regression-based approaches formalize the intuitive notion that two states that are distant from each other on some underlying measurement scale are more likely

to produce agreement in the pairwise ordering of the outcomes than will two states that are very near to each other. If distributions of values on the latent scale are assumed to be normal, then the ordinal responses can be modeled using probit regression; analogously, the assumption that the values follow an extreme value distribution leads to logit regression.

Historical Development of Health State Valuation Techniques

The development and adaptation of techniques for measuring health state valuations have occurred mostly since the early 1970s, but historical antecedents for this work may be found decades earlier. The history of the standard gamble is perhaps easiest to trace, as the technique debuted alongside the introduction of the expected utility theorem of von Neumann and Morgenstern (1944). Following this introduction, various approaches were proposed to assess von Neumann–Morgenstern utilities through specific types of standard gamble comparisons. The type of comparison that has been commonly adopted for use in health state valuations, in which the respondent chooses between a certain prospect of an intermediate outcome on the one hand and a gamble with the best and worst extreme outcomes on the other, was featured originally by von Neumann and Morgenstern and used subsequently in formulations for general utility assessment (i.e., not specific to health) by Frederick Mosteller, Duncan Luce, Howard Raiffa, and others. In 1968, Arnold Packer explicitly noted the applicability of the standard gamble to evaluation of health programs. Torrance *et al.* (1972) presented what may be the earliest published example of a comprehensive approach to measuring effectiveness of health programs with a utility assessment strategy based on the standard gamble. George Torrance had previously (in an unpublished dissertation, in 1971) undertaken a pilot test of the standard gamble technique, among others, in the context of a health care program evaluation.

The time trade-off approach, as currently implemented in valuing health states, appears to have been devised and named in the same study comprising Torrance's dissertation work, although the basic approach was discussed around the same time by Fanshel and James (1970), who referred to the notion as 'weighting through equivalence in time.' Torrance himself described the time trade-off as evolving from the so-called 'direct measurement technique' attributed to the psychologist Stanley Smith Stevens (1959), although the final format of the time trade-off bears little resemblance to this earlier proposal to directly elicit ratio assessments for two quantities. The time trade-off was originally developed to assess values for states considered better than being dead; another important milestone in development of the time trade-off was the elaboration by Torrance in 1982 of the method to accommodate states regarded as worse than dead.

Rating scales in health state valuation draw on a long history of related scaling approaches used in psychology and attitude measurement, including work by Louis Leon Thurstone in the 1920s. Patrick *et al.* (1973) applied a 'category scaling' approach to health measurement based on the method of equal-appearing intervals attributed to work published by Warren Torgerson in 1958. Patrick *et al.* operationalized this approach by having respondents place cards labeled with various health outcomes into equally spaced slots in a desk file sorter, numbered between 0 and 16. In another study published in the same year, the same authors used a linear rating scale, which has become the conventional rating scale approach in health state valuation. Subsequently, George Torrance adapted the approach with a 'desirability line' representing 101 equal interval categories spanning the range between 'Death, Least Desirable' and 'Healthy, Most Desirable.' Following these early precedents, numerous applications of rating scales in health state valuation have introduced a number of variations on this basic theme.

Magnitude estimation was proposed by Stevens (1951), in part as a response to the chief limitation he saw in the use of rating scales, which is that responses on rating scales appear to be nonlinearly related to the actual underlying scale that is being measured. Patrick *et al.* (1973), citing earlier work from the field of criminology, presented what appears to be the first application of magnitude estimation in valuation of health states. Another prominent use of the technique was in the Rosser and Kind index in 1978.

The person trade-off approach was named by Erik Nord in 1992, but the technique itself was applied already by Patrick *et al.* (1973) under the name of the 'equivalence' method. A proposal for 'weighting by equivalence in population' had appeared in the work of Fanshel and Bush (1970), but that earlier study presented the concept without applying it in empirical study. The person trade-off gained prominence through the publication of a review and empirical study by Nord (1995), which summarized prior applied work using the person trade-off and related techniques and presented the first comprehensive assessment of the reliability and possible biases in the technique. The profile of the method was also raised by its adaptation in the measurement of disability weights for DALYs in the Global Burden of Disease Study, as described by Murray (1996). The DALY study used two variants of the person trade-off in a deliberative group exercise. One of these variants – which compared life extension among disabled and nondisabled groups and thus differed from the typical person trade-off format described above – inspired criticism from Trude Arnesen and Erik Nord in 1999 for its potential ethical implications.

One of the most recent trends in measuring health state valuations actually relates to one of the oldest methodological traditions, which concerns estimation of cardinal measures based on ordinal responses. In the 1920s, Louis Leon Thurstone developed the 'law of comparative judgment' that provides the conceptual foundation for most approaches to deriving cardinal values from ordinal assessments. Following Thurstone, Ralph Bradley and Milton Terry, Duncan Luce, and Daniel McFadden further developed the axiomatic basis for choice models and refined analytic approaches based on a random utility model. Kind (1982) presented the first application of the Bradley–Terry–Luce approach to health state valuation, and there has been a recent revival of interest in these methods due to the relative simplicity of eliciting ordinal responses and a widening range of analytic tools to accommodate these responses.

Key Conceptual and Methodological Issues

There have been various conceptual interpretations of health state valuations that have produced some amount of ambiguity in defining the basis for measuring and understanding these valuations. When valuations are measured with the standard gamble, some people refer to these valuations as 'health utilities.' In fact, some have suggested that the standard gamble is the only method that produces 'utilities,' according to the von Neumann–Morgenstern framework. Others are less restrictive in the use of this term. Richardson (1994) and others have questioned the primacy of the standard gamble and challenged the prevailing argument that the standard gamble is preferred because its inclusion of risk aligns the technique with the inherently uncertain nature of medical decision making.

There has also been variation in the use of terms like 'quality of life' or 'health-related quality of life' in reference to health state valuations. The term 'quality of life' has been used widely in various social science contexts to refer to the overall subjective appraisals of happiness or satisfaction experienced by individuals. In health, the term 'quality of life' has sometimes been used in a more particular way to refer to a multidimensional construct relating to symptoms, impairments, emotional states, and domains of functioning. Because this use of 'quality of life' diverges from more general uses of the term, health researchers often refer to the distinct construct of 'health-related quality of life'. To the extent that an individual's health-related quality of life is understood in terms of a vector of levels on 'health-related' dimensions of life, it is similar to the conceptual notion underlying health state valuation, which can be used to attach an overall scalar value to such a multidimensional profile. Where health-related quality of life is viewed in terms of the contribution of an individual's health to his/her overall well-being, conceptual problems emerge from the fact that well-being is not clearly separable into independent health and nonhealth components (as, for instance, philosopher John Broome has argued).

In considering empirical differences between the different techniques for eliciting health state valuations, it is useful to recognize how the different constructs embodied in the techniques, for example, the 'utility' notion reflected in the standard gamble, may combine judgments about health with other values such as risk aversion. There has been a general consistency in the ordering of values (for the same state) produced by responses to the different valuation techniques, with rating scale values tending to be lowest (on a scale in which higher numbers imply better outcomes); standard gamble and person trade-off values highest; and time trade-off values tending to fall between these extremes. One interpretation of this typical finding is that the systematic variation across valuation techniques relates to the specific types of other values that are invoked by the particular framing of each technique. For example, a highly risk averse person will answer standard gamble questions in a way that produces values near 1.0, as the person will be unwilling to entertain even small probabilities of mortality. Several commentators have suggested that person trade-off responses are susceptible to an analogous set of values at the population level, which may be understood in terms of the 'rule of rescue,' by which respondents tend to choose a program that averts a relatively small number of deaths over a program that averts a very large number of nonfatal outcomes. Time trade-off responses may be influenced by a range of factors, such as discounting of future events, but the net effect of these factors may be relatively modest compared with the impact of risk aversion. Finally, various possible biases in rating scale responses have been considered, including a propensity to avoid values near the extreme ends of the scale, which is consistent with an overall downward shift in rating scale values.

Some health states are considered to be worse than being dead. Assignment of values to these has presented some challenges, especially in the use of the time trade-off. A typical approach to the time trade-off is first to ask whether a state is regarded as better than dead or worse than dead, as in the protocol developed by the Measurement and Valuation of Health (MVH) Group in 1994. For a worse-than-dead state, respondents in the MVH study were asked how many years spent in the health state (t) followed by a period of perfect health, summing to 10 years, would be equivalent to immediate death. By assigning values of 0 and 1.0 to dead and optimal health, respectively, valuations for a worse-than-dead outcome may be derived from the following equality: $\mathrm{U}(\text{health outcome}) \times t + \mathrm{U}(\text{optimal health}) \times (10 - t) = 0$, which simplifies to $\mathrm{U}(\text{health outcome}) = 1 - (10/t)$. In principle, this implies that the weight for a worse-than-dead state falls in the interval $(-\infty, 0)$. In practice, the lowest possible valuation using the MVH protocol is -39 (due to reporting of responses in quarter-year increments). Several studies have observed that treating worse-than-dead responses as originally intended – although faithful to the conceptual development of the time trade-off question – can lead to a large number of health states having negative average valuations, challenging face validity. In response, George Torrance, Paul Dolan, Leida Lamers, and others have considered various transformations of the worse-than-dead responses, which have prompted some controversy and a range of alternative proposals.

Conclusions

A large and growing literature on health state valuation has been directed toward a range of key issues including: the choice of technique for eliciting valuations; whose values to elicit; related issues around changing valuations over time (e.g., due to adaptation to decreased function); how to describe states for valuation; and the relevance of other values that may influence responses to health state valuation questions. This article has introduced six prominent techniques for eliciting valuations, discussed certain milestones in the historical development and evolution of these techniques, and mentioned some of the most salient conceptual and methodological issues relating to measurement of health state valuations.

See also: Cost–Value Analysis. Disability-Adjusted Life Years. Measurement Properties of Valuation Techniques. Multiattribute Utility

Instruments: Condition-Specific Versions. Quality-Adjusted Life-Years. Utilities for Health States: Whom to Ask

References

Fanshel, S. and Bush, J. W. (1970). A health-status index and its application to health services outcomes. *Operations Research* **18**, 1021–1066.

Kind, P. (1982). A comparison of two models for scaling health indicators. *International Journal of Epidemiology* **11**, 271–275.

Murray, C. J. L. (1996). Rethinking DALYs. In Murray, C. J. L. and Lopez, A. D. (eds.) *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, pp. 1–98. Boston: Harvard School of Public Health.

von Neumann, J. V. and Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making* **15**, 201–208.

Patrick, D. L., Bush, J. W. and Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research* **8**, 228–245.

Richardson, J. (1994). Cost utility analysis: What should be measured? *Social Science & Medicine* **39**, 7–21.

Rosser, R. and Kind, P. (1978). A scale of valuations of states of illness: Is there a social consensus? *International Journal of Epidemiology* **7**, 347–358.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S. (ed.) *Handbook of experimental psychology*, pp. 1–49. New York: Wiley.

Torrance, G. W., Thomas, W. H. and Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research* **7**, 118–133.

Further Reading

Lamers, L. M. (2007). The transformation of utilities for health states worse than death: Consequences for the estimation of EQ-5D value sets. *Medical Care* **45**, 238–244.

McDowell, I. (2006). *Measuring health.* New York: Oxford University Press.

Nord, E. (1992). Methods for quality adjustment of life years. *Social Science & Medicine* **34**, 559–569.

Salomon, J. A. (2003). Reconsidering the use of rankings in the valuation of health states: A model for estimating cardinal values from ordinal data. *Population Health Metrics* **1**, 12.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review* **34**, 273–286.

Torrance, G. W. (1976). Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences* **10**, 129–136.

Valuing Informal Care for Economic Evaluation

H Weatherly, R Faria, and B Van den Berg, University of York, York, UK

Glossary

Genomics The science of the function and structure of genomes, i.e., the DNA within a single cell of an organism.
Longitudinal study Any study using time series data.
Meta-analysis Using statistical techniques to synthesis the results from separate but related studies in order to obtain an overall estimate of treatment effect.
Multi-criteria decision analysis (MCDA) A framework for decision makers who need to consider multiple and sometimes conflicting factors when assessing the advantages and disadvantages of comparator interventions. MCDA has been applied broadly to inform health-care decisions. It has been used for shared decision making across different health-care stakeholders in the analysis and selection of interventions and for setting health-care priorities.
Regression discontinuity A cutoff or threshold above or below which a discrete intervention is performed or not performed.
Selection bias A distortion created when using data from a sample that differs systematically in its characteristics from the general population due to a feature of the selection process.
Shadow price The marginal cost or marginal value of a service as revealed in experiments or estimated by adjusting market prices.

Introduction

Informal care is the mainstay of support for many people living in the community, particularly those with long-term care needs. It refers to the care provided to individuals who would have difficulties managing without this help, by family or friends who are unpaid, although they may receive some nominal payment or state benefits. Some definitions also add that informal carers spend a significant proportion of their life providing care and support, however, if the amount of time spent caring is an important consideration, this requires specification. Informal care tasks include providing a range of care and assistance with activities of daily living, such as support with mobility, social support, personal care, and domestic assistance. In low-income settings, carers might even provide support with a broader range of medical-related tasks which could involve health-related care, particularly where state funding of health care is highly limited.

The size of the informal care economy, while substantial, is challenging to quantify and value. On an average one in nine people aged 50 years or more reported providing care for a dependent relative across all Organization for Economic Co-operation and Development (OECD) countries in 2007 (see http://www.oecd.org/health/health-systems/long-termcare.htm). The World Health Organization predicts that during the next 40 years the dependency ratio in China and India, in particular, will increase greatly. There is also a concern that young carers do not get adequate support. Substantial differences in informal care can be seen across OECD countries and, in part, this reflects the input of informal care versus formal, paid long-term care. For example, the level of informal care is twofold larger in Italy and Spain as compared to Sweden where care provision is more formalized. In economic terms, informal care can be described as a quasi-market commodity as there is no explicit market where informal care is bought and sold, and it does not have a directly observable value, such as a price, to reflect either the resources required to provide care-related tasks, or the benefits of doing so.

Informal care does not have an explicit value; however, this does not imply that it is a free resource. From an economic perspective, there are two key considerations relating to the carer when considering the value of informal care: (1) people offering informal care face an opportunity cost in that time spent on caring replaces other activities, such as education, employment, and leisure, and (2) there is an opportunity cost of using informal care as, while generating some potential benefits, such as the utility gained from caring for a loved one, being an informal carer is also associated with burden and disutility, such as reduced quality of life and negative health impacts. It is worth noting too that carer and care recipient's values are interdependent but this article focuses solely on the value of informal care in relation to the carer.

Informal care is rarely valued for inclusion in economic evaluation studies. Economic evaluation is increasingly used to inform national policy decisions about the efficient allocation of public funds across the economy. It offers a transparent framework to assess the relative costs and benefits of comparator interventions and has been applied in a range of sectors including the environment, health, and care. Many countries have developed detailed methods guidance for the economic evaluation of health-care interventions and now more attention is being given to the development of methods for the economic evaluation of care interventions. For example, from 1 Apr 2013 the remit for the National Institute for Health and Clinical Excellence in the UK is extending beyond health care to provide evidence-based guidance on social care interventions. Such guidance requires a systematic approach to measuring and valuing costs and benefits, consistent with the perspective chosen and considerably more research is required to identify the best approach to dealing with informal care in economic evaluations.

Valuing informal care for economic evaluation might be important in the following three respects: (1) for evaluating the cost-effectiveness of interventions or technologies to directly support informal carers (and care recipients), such as carers breaks or use of technologies such as robotic devices

for house cleaning, (2) to assess interventions in which there is an indirect impact on informal carers, such as health-related packages of care in which informal carers contribute, for example, use of medication for Alzheimer's disease, in which a carer is involved in administering the medication, or (3) for use in testing and redesigning services to evaluate the cost-effectiveness of different levels of access to more formal input, for example, from universal access to means-tested access. Once informal carer input is valued, their contribution is accounted for in an economic evaluation for use in informing the decision maker. In a scenario in which a new health-care intervention reduces health-care costs and increases use of informal care, relative to the comparator intervention, an analysis omitting this might result in a potentially undesirable shift of resource use from the health-care sector to the informal economy.

A broad range of methods have been developed to value informal care for economic evaluation but to date no formal guidance is available. This article reviews the methods for valuing informal care, including the methods used to measure time spent on informal care and the monetary valuation of this time, as well as nonmonetary methods and offers a brief review of the advantages and issues associated with applying currently available methods. The entry concludes by considering the implications of valuing informal care for economic evaluation. The same approach is used to review monetary and nonmonetary methods (see Figure 1). Four key steps comprise: (1) conceptualization of informal care, (2) identification of the caring activities or description of the effects of informal care, (3) measurement of the time spent on informal care (monetary) or the effects of informal care (nonmonetary), and (4) valuation of the time spent caring or the effects of informal care. This article focuses on the measurement and valuation, steps (3) and (4).

Monetary Valuation of Informal Care

Informal care can be incorporated within an economic evaluation on the cost side of the analysis. Figure 2 summarizes the methods involved in monetarily valuing informal care based on the four above-mentioned steps. Obtaining a clear and consistent definition of what constitutes informal care is far from straightforward. The majority of informal care is provided by one or more individuals known to the care recipient, such as their family, however informal care might be offered by individuals who are friends or who befriend them. As van den Berg *et al.* (2004) have discussed, this raises issues about the nature of the interaction between the care recipient and the carer. Activities such as cooking and cleaning ordinarily might be provided by a spouse when neither partner requires care. Typically, only those activities that take place because the care recipient requires care, owing to, for example, deteriorating health, should be included as informal care. These can be less than straightforward to identify, particularly when caring has taken place over a long time horizon and, additionally, life styles vary across different healthy couples. There may be shared benefits or some joint production involved in the carer undertaking tasks which complicates the ability to distinguish the component that is categorized as informal care. For example, the carer may provide informal care while participating in a leisure pursuit with the care recipient. Another consideration is the multiple tasks that carers might undertake. Some activities might be health related, such as supporting the care recipient with mobility and may substitute the need for health-care assistant input, whereas other aspects of care such as personal care and support with administrative tasks or keeping the care recipient company are more likely to impact beyond narrowly defined health. The impact of the caring task on the informal carer might be affected by the intensity and duration of the care provided. All these issues need to be specified upfront so that assumptions underlying the analysis are transparent. To aid comparability across economic evaluations and across settings, a systematic approach is required to conceptualizing, identifying, measuring, and valuing informal care.

Measuring Informal Care

Once tasks are identified as informal care, a method is required to measure time spent on informal care and four methods are reviewed next.

Figure 1 Steps involved in the measurement and valuation of informal care.

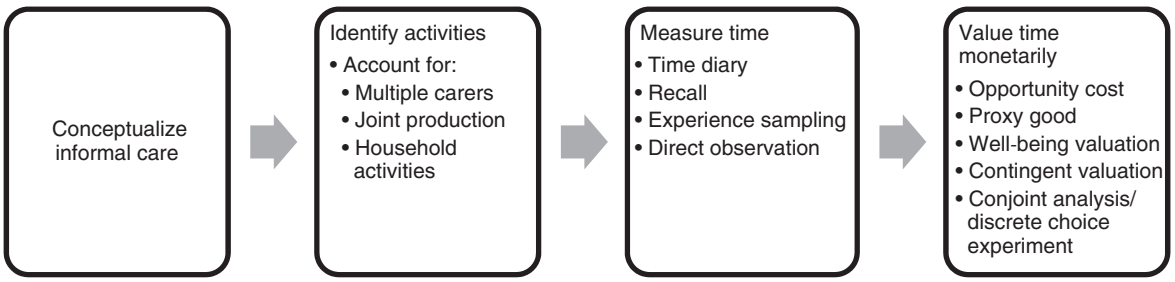


Figure 2 Monetary valuation of informal care.

Time diary method

Respondents are asked to report the sequence of activities undertaken as defined by the analyst and the amount of time spent on each activity over a set period of time, such as every quarter of an hour. An example of use of the diary method developed to measure informal care time is found in van den Berg and Spauwen (2006). Advantages include systematic collection of data within a brief space of time of the activities occurring, thus minimizing the recall period. In addition, respondents can indicate whether they are undertaking more than one care-related activity at the same time and, in principle, this allows the analyst to correct for joint production. There is, however, no generally agreed method to do this and in practice there is often no correction for joint production. Disadvantages of the method include that it can be very time-consuming to complete the diary and this can put carers off participating in such studies and completing the diary can impact on the time spent caring. In principle, this method might be thought to be the most accurate, but it has problems with compliance.

Recall method (also known as the stylized or questionnaire method)

This is probably the most commonly used method of time measurement. It involves asking respondents to report the amount of time spent on a particular activity and the frequency of the activity over a set period of time in the past. Examples of its use include the Carer Activity Survey by Davis *et al.* (1997), and the recall questionnaire used by van den Berg and Spauwen (2006). On other occasions, the recall method only asks about the amount of time spent on providing informal care even without referring to any care task. Compared to the time diary method, the recall method is less time-consuming to complete. Disadvantages, however, include that it is more sensitive to recall bias including systematic error because of differences in the accuracy of reporting past events by respondents. Although this method might be less accurate than the diary method, this consideration needs to be weighed against potentially greater compliance as carers seem able and more willing to use this method.

Experience sampling method (also known as beeper or buzzer method)

This has been used in time use and in wellbeing research, but to date, to the best of knowledge available, this approach has not been applied to measure informal care time. The method prompts respondents to register their activity at random instants over a prespecified time using a signal emitted from an electronic device. As the device beeps, respondents record the activities they are undertaking at that point in time. Gershuny (2011) found that on average this method provides highly accurate weighted sample estimates of time use across the sample. Experience sampling is less prone to recall bias than the time diary method because of immediate recording of the activity. If the respondent records their own responses rather than being given a closed-ended questionnaire, this method is likely to record all activity taking place. Disadvantages include respondent burden. A beeper system can intrude in daily life and might result in failure to respond. The method does not record the duration of the activities, nor are activities recorded sequentially, so there is a lack of context to the activity.

Direct observation method (also known as the continuous observation or outsider method)

This method involves observers recording activities and as such might be considered objective, rather than relying on self-report, and therefore it is considered to be highly accurate. However, it is very time- and resource-consuming and can also be very intrusive.

Valuing Informal Care in Monetary Terms

Once the time spent on informal care is measured, a method is required to attach a monetary value to informal care time. Examples of studies valuing informal care in monetary terms are provided in **Table 1** and this shows that the monetary values derived vary considerably by method and by caring task. There are two broad economic approaches for valuing time of using revealed preferences or stated preferences, as described next.

Revealed preference

These preferences are obtained by analyzing individuals' behavior or indirectly via preferences revealed in other markets such as from datasets recording individuals' decisions about services which are close substitutes for activities undertaken by informal carers. Typically, revealed preference methods use wages or income data to derive monetary values. Note that both the opportunity cost and proxy good methods do not incorporate the full impact of informal care on carers. However, whether the full impact should be included in the analysis will depend on the decision problem and the perspective taken.

Opportunity cost method

This method values informal care as the income forgone by the carer when spending time on informal care. There are many examples of how the opportunity cost method has been used to value informal care time, for example, in van den Berg *et al.* (2006). Income forgone is the carer's current wage rate (if employed) or can be estimated based on the previous wage rate if the carer worked in the past. The average (or median) net wage of people employed in the labor market who have the same sociodemographics might be used for those who have never worked. Use of this method is less straightforward for children and younger people as time spent on informal care may reduce the time available for education, which can have consequences later in life. Therefore, this method can result in different values for informal care, for example, a person with the potential of higher wages (e.g., a skilled professional) will have higher income forgone and the value attached to an hour of informal care will be higher for this person than for a person with the potential to earn a lower wage. Another issue is how to account for informal care activities that replace leisure time or unpaid work. Applying the wage rate of paid work to leisure time assumes that the wage rate reflects the marginal value of the time across the different uses of time. Another issue relates to the implications

Table 1 Examples of studies valuing informal care in monetary terms[a]

Method	*Reference*	*Application*	*Unit cost per hour (price year if stated/country)*	*Cost converted to GB sterling 2013*[a]
Opportunity cost	Smith and Frick (2008)	Average hourly income for all employed county residents	$17.34 (2004/US)	£12.56
	Wilson *et al.* (2009)	Average gross hourly wage rate for both genders	£13.11 (2004/UK)	£17.44
Proxy good	Dewey *et al.* (2002)	Hourly rate per type of activity	Community and domestic services, A$11.20	£8.53
			Personal care, A$13.45	£10.13
			All care by secondary carers, A$11.20 (1997/Australia)	£8.44
	Gaugler *et al.* (2003)	Hourly rate for home care services	$2.93 (1993/US)	£3.38
Wellbeing	van Den Berg and Ferrer-i-Carbonell (2007)	Extra compensation to maintain same level of wellbeing after providing additional hour of care	€8–9 (2001/the Netherlands)	£7.13–8.02
Contingent valuation	Gustavsson *et al.* (2010)	Carers' monthly willingness to pay for 1 h per day of reduction in informal care	UK, £105, Spain, £121, Sweden, £59, US, £144 (date not stated)	As stated
	van den Berg *et al.* (2005)	Care recipients' WTP for an additional hour of informal care per week and their and willingness to accept (WTA) for a reduction in 1 h of the informal care received	Rheumatoid arthritis:	
			Care recipients: WTP €7.84	£6.99
			WTA €8.22	£7.33
			Carers: WTP €7.80	£6.96
			WTA €9.52	£8.49
		Carers' WTA to provide an additional hour and WTP to provide one less hour of care	Heterogeneous sample:	
			Care recipients:	
			WTP €6.72	£5.99
			WTA €8.62	£7.68
			Carers: WTP €8.61	£7.68
			WTA €10.52	£9.39
			(2001/the Netherlands)	
Conjoint Analysis/ Discrete Choice Experiment	van den Berg *et al.* (2008)	Extra compensation per hour required to provide 21 h instead of 7 h of informal care per week	€12.36 (2001/the Netherlands)	£11.02
	Mentzakis *et al.* (2010)	Carers' willingness to accept to provide an additional hour of care of a number of tasks	Personal care: £0.12–2.29, Supervision: £0.07–0.81, Household tasks: £0.25–1.04 (date not stated/UK)	As stated

[a]The two websites used to convert the monetary values to GB sterling for the financial year 2012–13 are given in the section Relevant Websites.

of using net or gross wage rates. The wage rate net of tax reflects the opportunity cost to the carer, whereas the gross rate reflects the opportunity cost to society. The choice of wage rate needs to be consistent with the perspective chosen for the analysis.

Proxy good method

Also known as replacement cost method, proxy good method values informal care time use at the price of a close substitute. The relevant substitute depends on the activities undertaken, for example, a health-care assistant wage could be used to value informal care time spent on help with feeding and the wage of a housekeeper for help with cleaning the house or doing the laundry. To use this approach the analyst requires data on caring activities undertaken, time spent on these activities and proxy values for each activity. This method assumes exact substitutability between formal and informal care, including assuming that the informal carer and care recipient are indifferent or have the same preferences for informal care as compared to formal care. The method implicitly assumes that any prices used are appropriate reflections of value.

Wellbeing valuation method

This method estimates the monetary value of providing informal care. It does so by estimating the carer's wellbeing as a function of income and time spent caring, among other things. This allows the analyst to estimate the income required to compensate the carer for the loss in wellbeing because of providing informal care. The wellbeing valuation method uses data directly obtained from carers. The first application of this approach to valuing informal care time was undertaken by van Den Berg and Ferrer-i-Carbonell (2007). An assumption underlying the wellbeing valuation method is that wellbeing can be measured empirically and some analysts even assume that this measurable wellbeing is a proxy for utility, more specifically for experienced utility. An empirical finding is that providing informal care is associated with wellbeing losses as

well as a positive association between income and wellbeing. In cases in which informal care would not be negatively correlated with wellbeing, it would not be possible to calculate the related monetary compensation and in case of a positive empirical association between informal care and wellbeing one could argue that the informal carer would be willing to pay to provide the care themselves: In other words, they may have a strong preference to provide the care and may benefit from caring. Although the authors are not aware of any published study that has found this positive association, the information obtained using this method might still be included in an economic evaluation.

Stated preference

Stated preferences are obtained directly from respondents by asking them to consider hypothetical situations, typically using survey methods. The method relies on statements of preference, and not on actual choices and, hence, may not reflect the respondent's actual behavior and are, therefore, criticized by mainstream economists.

Contingent valuation

Contingent valuation values informal care either in terms of the maximum monetary amount informal carers would be willing to pay for reducing caring activities, or the minimum monetary amount that they would be willing to accept for supplying extra informal care. Contingent valuation has been used to estimate the value of informal care in a few studies, for example, Gustavsson *et al.* (2010) estimated the willingness to pay for reductions in informal care need in Alzheimer's disease. Although contingent valuation questions might be relatively simple to ask, they might not be straightforward to answer as carers might not be used to thinking about monetary valuation of informal care time and, therefore, some respondents may be unwilling to value carer time in monetary terms. This might be especially true if the question is framed as if the care recipient might pay the carer. In an attempt to solving this problem van den Berg *et al.* (2005) suggested framing the question as if the government was going to compensate the carers, as sometimes happens via carer allowances for example. There is a substantial literature on biases involved in applying this method especially as willingness to accept values tend to be larger than willingness to pay ones and this difference cannot be explained based on economic theory. When van den Berg *et al.* (2005) explored this issue in relation to informal care in the Netherlands, the differences were quite small suggesting this bias is less persistent when applying the method to value informal care. It might be necessary to explore this further and to consider whether responses might be influenced by culture and the health and care system of the respondents participating in the study.

Discrete choice experiment (although strictly speaking based on different theories also known as conjoint measurement or conjoint analysis)

Discrete choice experiment uses survey methods to obtain respondents' estimation of the relative value of different attributes of a service which might include health, nonhealth, and process attributes. In addition, if a cost or a price is included as an attribute, a monetary value of the other attributes can be derived. Mentzakis *et al.* (2011), for example, included an attribute of the amount of money (£0, £4, £10, and £17) a respondent would receive as compensation per hour for the informal care they provided and in answering this, respondents were asked to consider the effect of the caring role on their own health and wellbeing. The methodology assumes that a service can be described by its constituent characteristics and that the total utility, satisfaction, or preference that a respondent derives from a service is determined by the utility they gain from each of the constituent parts. Examples using this method applied to informal care, have, to date, mainly included care tasks and care time and although this method in theory can also take into account the full impact or value of undertaking informal care on the carer themselves, including for instance involved health losses, this has not been contested in empirical applications so far.

Nonmonetary Valuation of Informal Care

Informal care can also be incorporated in the effect side of an economic evaluation, rather than on the cost side. As for monetary valuation of informal care, there are four key steps (see **Figure 3**): (1) conceptualize informal care, in a similar way as described in Section Monetary Valuation of Informal Care, (2) describe the effect on the informal carer, (3) measure the effect on informal care, and (4) attach a nonmonetary valuation to the effect. This section focuses solely on the impact on carers, however, as noted earlier, there is likely to be considerable interdependence between the outcomes for the carer and the care recipient.

Methods to incorporate informal care as an effect comprise three key measures: Burden, health, and health-related quality of life and informal care-related quality of life and some examples of these are provided in **Table 2**. In practice there may be some overlap between these measures.

Table 3 compares the dimensions included in three measures of care-related quality of life, the Carer Quality of Life Instrument (CQLI), the care-related quality of life (CarerQoL), and the Carer Experience Scale. As can be seen from the table, not all measures include the same dimensions, for example, CQLI and CarerQoL include physical health and energy, whereas the Carer Experience Scale does not.

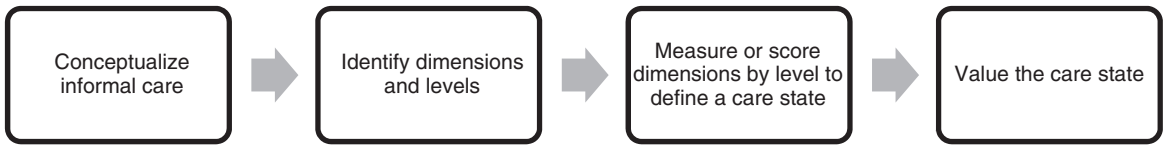


Figure 3 Nonmonetary valuation of informal care.

Table 2 Examples of nonmonetary methods for measuring informal care

Constructs of informal care	*Measures*	*Examples of nonmonetary instruments*
Burden of care	• Objective burden of informal care in terms of caring activities undertaken • Subjective burden in terms of impact of informal caring on carer wellbeing	Objective burden: • Carer Activities Time Survey (Clipp *et al.*, 1996) • Carer Activity Survey (Davis *et al.*, 1997) Subjective burden: • Caregiver Reaction Assessment (Given *et al.*, 1992) • Carer Strain Index (Robinson, 1983) Objective and subjective burden: • Carer Subjective and Objective Burden Scales (Montgomery *et al.*, 1985) • Cost of Care Index (Kosberg and Cairl, 1986)
Health and health-related quality of life	• Health • Health-related quality of life	Health: • SF-36 (Hughes *et al.*, 1999) Health-related quality of life: • EQ-5D (Dixon *et al.*, 2006)
Informal-care related quality of life	• Carer quality of life • Carer wellbeing	• Carer Quality of Life Instrument (CQLI) (Mohide *et al.*, 1988) • The care-related quality of life (CarerQoL) instrument (Brouwer *et al.*, 2006) • The Carer Experience Scale (Al-Janabi *et al.*, 2008)

Table 3 Comparison of the dimensions included in three measures of care-related quality of life

Dimensions	*CQLI*	*CarerQoL*	*Carer experience scale*
Physical health and energy	☑	☑	
Emotional and mental health	☑	☑	
Social relationships and other activities	☑	☑	☑
Sleep	☑		
Relationship carer/care recipient	☑	☑	☑
Fulfillment with carer situation		☑	
Financial consequences of carer		☑	
Support (formal and/or informal)		☑	☑
Carer's control over care-giving activities			☑
Carer's perception of fulfilling a duty			☑

As different dimensions are incorporated across the three instruments, this limits their direct comparability and this complicates synthesis of effects across the instruments. An additional consideration is to what extent each measure incorporates the full effect on informal carers.

Measuring the Effects of Informal Care

Burden of informal care

Carer burden is one of the most commonly used indicators of informal care. It is an attempt to quantify the physical, psychological, social, and financial impacts of caring. Sometimes a distinction is made between objective and subjective burden with objective burden referring to observable aspects of informal care, such as the events and activities, with subjective burden referring to the perception of the caring experience by the carer, including feelings, attitudes, and emotions. Both concepts of burden are complementary as the time spent on caring and the activities involved (objective burden) also involve the perception of the burden of care by carers (subjective burden). Several instruments are available to measure burden as, for example, reviewed by Deeken *et al.* (2003). This heterogeneity adds to the complexity of cross-study comparisons and makes synthesis of results across studies difficult. Although the concept of burden traditionally focuses on the negative aspects of caring and places less obvious emphasis on the benefits or utility obtained by the care from caring, there have been attempts within the burden literature to incorporate benefits and the Caregiver Reaction Assessment is an example of such a measure.

Informal carer's health-related quality of life

Measures of health-related quality of life have been used to account for the consequences of informal care on carers.

Caring may impact positively and negatively on carers' health-related quality of life. Instruments for use in measuring health-related quality of life include the SF-36 and the EQ-5D. Using health-related quality of life instruments to incorporate the impact of informal care on carers has been criticized because it may not capture the full effects on quality of life and well-being. However, the use of generic health-related quality of life instruments is helpful for comparing results across economic evaluations, including across disease areas.

Informal caregiver quality of life

As discussed by Stull *et al.* (1994), this concept is broader than health extending to incorporate all the factors that may impact on the carer's life. Three examples of such instruments are reviewed briefly here including the caregiver quality of life instrument, the care-related quality of life instrument and the Carer Experience Scale.

The caregiver quality of life instrument

This was the first carer-specific instrument developed to value carers' wellbeing states. Mohide *et al.* (1988) report the CQLI as comprising five dimensions which were identified from a review of the literature and clinical opinion. The dimensions comprise two social (amount of time to socialize with family and friends and quality of the interpersonal friendship between the carer and the care recipient), two physical (adequacy of amount of sleep and degree of physical wellness and energy), and one emotional (degree of happiness and freedom from anxiety and frustration). Each dimension is described using one of four levels, comprising almost always, most of the time, half of the time, rarely, or almost never. Four standardized hypothetical health states are included, that is, the ideal caregiver quality of life reference state where the caregiver was almost always well on each of the dimensions, mild, moderate, and severe caregiver wellbeing states. Twenty-nine family caregivers and ten relatives of well older people describe their state of wellbeing, relative to the standardized states, over the preceding fortnight.

The care-related quality of life instrument

CarerQoL was developed to incorporate the impact of informal care on carer's quality of life for use in an economic evaluation. Brouwer *et al.* (2006) presented the conceptualization and first test of the CarerQoL instrument. The CarerQoL-7D questionnaire measures how satisfied the carer is with their care-giving situation. The CarerQoL-7D contains seven dimensions comprising fulfillment, relational, mental, social, financial, support, and physical and each dimension is judged on one of three levels; some, a lot of. The dimensions were identified by selecting those most frequently assessed in several carer burden scales. To the authors' knowledge, the CarerQoL instrument has not been applied in an economic evaluation in practice.

The carer experience scale

This scale also values the carer experience for possible use in an economic evaluation. Developed by Al-Janabi *et al.* (2008), it has six dimensions comprising the carer-recipient relationship, institutional support, informal support, activities outside caring, control, and duty. Each dimension is described on one of three levels, that is most, some, or sometimes, few, little, or rarely. The dimensions were identified through a metaethnography of qualitative research in informal care together with interviews with carers.

Valuing the Effects of Informal Care

Informal carer's health-related quality of life

Published algorithms are available to convert health state descriptors from generic health-related quality of life instruments, such as the SF-12, SF-36, and the EQ-5D to HRQoL weights. Valuation tools include the visual analog scale (VAS), which strictly speaking does not generate a utility value, and the time trade-off (TTO) and standard gamble techniques which both incorporate choice, and standard gamble also includes uncertainty and for this reason tends to be the most favored valuation tool. However, VAS and TTO are considered easier to use and less cognitively burdensome.

It is worth noting that in undertaking informal care, caring may negatively impact on health-related quality of life but carers may also have health problems themselves: The causality of undertaking care and having poor health-related quality of life is unclear and this presents a challenge to analysts who aim to separate these issues. If providing care causes health-related quality of life loss and if carers are happy to take the loss, that is, the utility function of the carer might be interdependent with that of the care recipient's: A key challenge is how to value this preferred health loss. Further consideration is required on the role of interdependent utility functions in economic evaluations, particularly if the local context has an influence.

Informal caregiver quality of life

The carer quality of life instrument

CQLI uses the TTO to value states by indicating the number of years of future life in the 'burdened' test state they would exchange for a year in the 'ideal' reference state. The 'ideal' reference state refers to the best wellbeing state, in which carers almost always feel physically well and energetic, almost always feel happy and free from worry or frustration, almost always have sufficient to socialize with family and friends, almost every night get an adequate amount of undisturbed sleep, and almost always gets along well with the person being cared for. Drummond *et al.* (1991) used the CQLI in an economic evaluation comparing a support program for carers of elderly people with dementia with usual care. Quality-adjusted life-years (QALYs) with and without the support program were calculated using the CQLI values to inform the wellbeing experienced by the carers.

It should be noted that this is not the same as standard QALYs which are solely health-related.

The care-related quality of life

CarerQoL includes a VAS to quantify how happy the carer feels currently on a scale of 0 (completely unhappy) to 10 (completely happy).

The carer experience scale

This scale was valued using best-worst scaling in an orthogonal main effects design. In best-worst scaling, respondents compare statements within a profile and select the most and

the least desirable statement. The best state, i.e., that with the lowest burden, has the value of 0 (zero), whereas the worst state, i.e., that with the worst burden and which every dimension has the worst level, was set to 100. Each dimension has different values and together these descriptors define the health state. For example, the health state with dimensions at the middle level has the value of 64.58. This provides a cardinal scale, however, there is no means of combining it with length of life to generate QALYs.

Discussion

This section offers an overview of methods to value informal care in monetary or nonmonetary terms. The choice of method depends on the research question, the data available, and the type of research being undertaken. This section has focused on the application of these methods describing their use in practice and has not discussed the normative foundations underpinning different approaches. However, in undertaking an economic evaluation, assuming a decision is taken to value informal care, a key issue is how to combine the method chosen with other methods used in economic evaluation and to ensure no double-counting. If monetary valuation is chosen, as a consequence of which the values are included on the cost side of an economic evaluation, then it could be combined with the cost of other inputs. A challenge for monetary valuation methods is that the values obtained are dependent on the individual's income, as greater ability to pay will drive up their willingness to pay for a service and possibly willingness to accept to provide informal care. Valuing informal care in terms of the effects of health-related quality of life raises the question on how to aggregate the values with care recipient's health-related quality of life, and it is not clear how to incorporate this formally into economic evaluation. It has been suggested that multicriteria decision analysis could be used to get around this issue, although this approach is not without its detractors.

Stronger guidance for analysts on whether and how to value informal care would result in greater uniformity in valuing informal care within studies and this would enhance the transferability and comparability of economic evaluations. Before this, more conceptual work needs to be undertaken to examine how to identify, measure, and value informal care. Informal carers provide a vital service within the community and this review illustrates that there are many methods available to quantify the economic value of this essential service.

See also: Cost-Effectiveness Modeling Using Health State Utility Values. Quality-Adjusted Life-Years. Valuing Health States, Techniques for. Willingness to Pay for Health

References

Al-Janabi, H., Coast, J. and Flynn, T. (2008). What do people value when they provide unpaid care? A meta-ethnography with interview follow-up. *Social Science and Medicine* **67**, 111–121.

van den Berg, B., Al, M., van Exel, J., Koopmanschap, M. and Brouwer, W. (2008). Economic valuation of informal care: Conjoint analysis applied in a heterogeneous population of informal caregivers. *Value in Health* **11**(7), 1041–1050.

van den Berg, B., Bleichrodt, H. and Eeckhoudt, L. (2005). The economic value of informal care: A study of informal carers' and patients' willingness to pay and willingness to accept for informal care. *Health Economics* **14**(4), 363–376.

van den Berg, B., Brouwer, W., Exel, J. and Koopmanschap, M. (2005). Economic valuation of informal care: The contingent valuation method applied to informal caregiving. *Health Economics* **14**(2), 169–183.

van den Berg, B., Brouwer, W., van Exel, J., et al. (2006). Economic valuation of informal care: Lessons from the application of the opportunity costs and proxy good methods. *Social Science and Medicine* **62**(4), 835–845.

van den Berg, B., Brouwer, W. and Koopmanschap, M. (2004). Economic valuation of informal care. *European Journal of Health Economics* **5**(1), 36–45.

van den Berg, B. and Ferrer-i-Carbonell, A. (2007). Monetary valuation of informal care: The well-being valuation method. *Health Economics* **16**(11), 1227–1244.

van den Berg, B. and Spauwen, P. (2006). Measurement of informal care: An empirical study into the valid measurement of time spent on informal caregiving. *Health Economics* **15**(5), 447–460.

Brouwer, W., van Excel, N., van Gorp, B. and Redekop, W. (2006). The CarerQoL instrument: A new instrument to measure care-related quality of life of informal caregivers for use in economic evaluations. *Quality of Life Research* **15**(6), 1005–1021.

Clipp, E., Moore, M. and George, L. (1996). The content and properties of the Caregiver Activities Time Survey (CATS): An outcome measure for use in clinical trial research on Alzheimer's disease. *American Journal of Alzheimer's Disease and Other Dementias* **11**(6), 3–9.

Davis, K., Marin, D., Kane, R., et al. (1997). The Caregiver Activity Survey (CAS): Development and validation of a new measure for caregivers of persons with Alzheimer's disease. *International Journal of Geriatric Psychiatry* **12**(10), 978–988.

Deeken, J., Taylor, L., Mangan, P., Yabroff, R. and Ingham, J. (2003). Care for the caregivers: A review of self-report instruments developed to measure the burden, needs, and quality of life of informal caregivers. *Journal of Pain and Symptom Management* **26**, 53–922.

Dewey, H. M., Thrift, A. G., Mihalopoulos, C., et al. (2002). Informal care for stroke survivors. *Stroke* **33**(4), 1028–1033.

Dixon, S., Walker, M. and Salek, S. (2006). Incorporating carer effects into economic evaluation. *PharmacoEconomics* **24**(1), 43–53.

Drummond, M., Mohide, E., Tew, M., et al. (1991). Economic evaluation of a support program for caregivers of demented elderly. *International Journal of Technology Assessment in Health Care* **7**, 19–209.

Gaugler, J., Zarit, S., Townsend, A., Parris Stephens, M.-A and Greene, R. (2003). Evaluating community-based programs for dementia caregivers: The cost implications of adult day services. *Journal of Applied Gerontology* **22**(1), 118–133.

Gershuny, J. (2011). *Time-use surveys and the measurement of national wellbeing*. Oxford: University of Oxford. Centre for Time-use research, Department of Sociology.

Given, G., Given, B., Stommel, M., et al. (1992). The caregiver reaction assessment (CRA) for caregivers to persons with chronic physical and mental impairments. *Research in Nursing and Health* **15**(4), 271–283.

Gustavsson, A., Jönsson, L., McShane, R., et al. (2010). Willingness-to-pay for reductions in care need: Estimating the value of informal care in Alzheimer's disease. *International Journal of Geriatric Psychiatry* **25**(6), 622–632.

Hughes, S., Giobbie-Hurder, A., Weaver, F., Kubal, J. and Henderson, W. (1999). Relationship between caregiver burden and health-related quality of life. *The Gerontologist* **39**, 534–545.

Kosberg, J. and Cairl, R. (1986). The cost of care index: A case management tool for screening informal care providers. *Gerontologist* **26**, 273–278.

Mentzakis, E., Ryan, M. and McNamee, P. (2010). Using discrete choice experiments to value informal care tasks: Exploring preference heterogeneity. *Health Economics* **20**(8), 930–944.

Mohide, E., Torrance, G., Streiner, D., Pringle, D. and Gilbert, R. (1988). Measuring the wellbeing of family caregivers using the time trade-off technique. *Journal of Clinical Epidemiology* **41**, 475–482.

Montgomery, J., Gonyea, J. and Hooyman, N. (1985). Caregiving and the experience of subjective and objective load. *Family Relations* **34**, 19–26.

Robinson, B. (1983). Validation of a caregiver strain index. *Journal of Gerontology* **38**, 344–348.
Smith, C. A. and Frick, K. D. (2008). Cost-utility analysis of high- vs. low-intensity home- and community-based service interventions. *Social Work in Public Health* **23**(6), 75–98.
Stull, D., Kosloski, K. and Kercher, K. (1994). Caregiver burden and generic wellbeing: Opposite sides of the same coin. *Gerontologist* **34**(1), 88–94.
Wilson, E., Thalanany, M., Shepstone, L., et al. (2009). Befriending carers of people with dementia: A cost utility analysis. *International Journal of Geriatric Psychiatry* **24**, 610–623.

Relevant Websites

http://www.gocurrency.com/v2/historic-exchange-rates.phpccode2=AUD&ccode=GBP&frMonth=3&frDay=7&frYear=1997
GoCurrency.

http://www.greenbook.treasury.gov.uk
HM Treasury.

http://www.oecd.org/health/health-systems/healthataglance2011.htm
Organization for Economic Co-operation and Development.

http://www.thisismoney.co.uk/money/bills/article-1633409/Historic-inflation-calculator-value-money-changed-1900.html
This is MONEY.

http://www.who.int/chp/knowledge/publications/ltc_needs.pdf
World Health Organization.

Waiting Times

L Siciliani, University of York, York, UK

Glossary

Inpatient waiting time Time between the addition to the list, following specialist assessment, and the date of admission for treatment.

Maximum waiting-time guarantee Establishes that no patient should wait more than a predetermined maximum.

Outpatient waiting time Time between family doctor visit and specialist visit.

Introduction

Publicly funded systems are often characterized by limited budgets and free-of-charge (or highly subsidized) access to healthcare. These two features often translate into an excess demand which generates a waiting list. Patients may have to wait for a significant time before accessing health care. Waiting times generate dissatisfaction for patients as they postpone benefits from treatment, may induce a deterioration of the health status of the patient, prolong suffering, and generate uncertainty. How to deal with or reduce waiting times is often the subject of debate in political campaigns: it is not surprising that waiting times have become a key health policy concern in many Organization for Economic Cooperation and Development (OECD) countries.

This article is devoted to presenting some key ideas on the role of waiting times in the market of health services. It draws selectively on the existing health economics literature. The article discusses: (1) different types of waiting-time measures; (2) how waiting times can be thought of as a rationing mechanism which brings together the demand for and the supply of health services; (3) the potential role of patients' choice and competition among public healthcare providers to reduce waiting times; (4) the role of waiting times in allocating patients between the public and the private sector; (5) the scope for policies based on the maximum waiting-time guarantees; (6) the equity implications of using waiting times in the health sector. The Section 'Technical Appendix' provides formal frameworks on waiting-time measurements, waiting-time dynamics and hospital competition, which are covered more intuitively, respectively, in the first three sections. The final section provides references for further reading.

Waiting-Time Measures

Patients in need of health care can experience different types of waiting. In many countries, the first contact point of the patient is the family doctor (also known as general practitioner). The family doctor will then refer the patient to a hospital specialist. Between the visit of the family doctor and of the specialist the patient will have to wait. This is often referred to as the outpatient waiting time. Once the specialist visits the patient and thinks that the patient needs (a medical or surgical) treatment, then the patient is typically added to the waiting list. The time between the addition to the list and the date of admission for treatment is often referred to as the inpatient waiting time.

For many nonemergency (also known as 'elective') treatments, like hip and knee replacement, and cataract surgery, inpatient waiting times can be substantially long with an average of 3–6 months. Waiting times are typically shorter for more severe conditions (e.g. patients in need of cancer treatment). This article does not focus on emergency treatment, where waiting times are short (and a matter of minutes or hours).

Waiting times for non-emergency treatments are routinely collected in many OECD countries (e.g., Australia, Canada, Norway, Portugal, and the UK) through administrative databases. Each country collects several measures. Definitions, however, tend to vary across countries. They can refer to specific procedures (like hip replacement, cataract surgery) or broader categories (like specialties or all nonemergency patients). Waiting times are often reported according to basic descriptive statistics: the mean or the median waiting, the number or proportion of patients waiting more than a given time (say 6, 9, or 12 months) or the waiting times at the 80th or 90th percentile of distribution. Given that the waiting-time distribution is skewed and characterized by a small number of patients with long wait, the mean is typically longer than the median (up to 20–30% difference). Measures based on the number (or proportion) of patients with a long wait capture only the upper tail of the distribution. For example if 5% of patients wait more than 12 months, it could be that the remaining 95% wait for 10 months or 5 months. The reason for reporting such figures is that the number of 'long waiters' are seen as the most problematic ones.

Regardless of which specific figure is employed to report waiting times, waiting-time information can serve different purposes. For example, waiting-time information can be used to set targets for health care providers, with hospitals reporting a longer wait being subjected to penalties or a closer monitoring. Waiting-time information can be used to enhance patients' ability to choose providers: in such cases the information is provided either to the family doctors or is publicly available on the internet. Waiting-time information may also be necessary to establish and enforce maximum waiting-time guarantees, under which governments state that no patient should wait more than a predetermined waiting time.

For a given procedure, condition or specialty, two common measures of waiting times can be recorded: (1) the waiting

 doi:10.1016/B978-0-12-375678-7.01304-3

time of the patients on the list at a point in time (or a census date) and (2) the waiting time of patients who have received a treatment (during a predetermined period). The first measure refers to an 'incomplete' measure of waiting time as the patient is still on the list (all patients are still waiting): there will be some patients who have just entered the list and some patients who have been on the list for a long time. The second measure refers to the full duration of the waiting-time experience from the time the patient is added to the list to the time the patient is admitted to the hospital for treatment: the measure is computed retrospectively once the wait is terminated.

Figure 1 provides an example which illustrates the difference between the waiting times of the two distributions. In each period two patients enter the waiting list, the first waiting for 1 month and the second waiting for 5 months. In period t only two patients have their wait completed. One has waited for 1 month and one has waited for 5 months. The average wait of patients treated (which is the correct expected wait for the patient) in period t is therefore three periods. At time t there are 6 patients waiting on the list: two patients have waited for one period, one for two periods, one for three periods, one for four periods, and one for five periods. The average wait is therefore 2.7 periods. The key difference between the two distributions is that the distribution of patients on the list tends to ''oversample'' long-wait patients. Indeed, in the example five out of six are long-wait patients at the time of observation while in fact there is an equal proportion of long and short-wait patients. The reason why in the example the average wait of patients on the list is shorter than the average wait of patients treated is that most of the patients in the list at the time of observation have not completed their wait yet. Hence the data on patients on the list suffers from ''interruption'' bias. In the present example the interruption bias dominates the long-wait oversampling bias. In Section 'Waiting Time of Patients on the List Versus Waiting Time of Patients Treated: An Example' differences between the two distributions are illustrated more formally. An example where the average wait of the patients on the list is higher is also provided.

Waiting Times as a Mechanism to Bring the Healthcare Market to an Equilibrium

A Simple Theoretical Framework

Publicly funded healthcare sectors in many countries are often characterized by zero or little copayments and simultaneously by capacity constraints. If capacity is less than potential demand, this generates an excess demand. If patients in 'excess' are added to a waiting list, then patients will have to wait a certain time before receiving treatment.

Economists have argued that waiting times can then be thought of as a price (a non-monetary one), which patients have to pay to access healthcare. In other sectors of the economy, monetary prices bring demand and supply in equilibrium; higher prices reduce demand by discouraging some consumers to buy a certain good and increase supply by encouraging the provider to expand production. Equilibrium prices are determined where the demand curve crosses the supply one. The standard economic demand–supply framework can readily be adapted to the health sector. Like monetary prices are determined such that demand is equal to supply, in the health sector waiting times (the non-monetary price) are determined to bring demand and supply in equilibrium. Note that if demand would systematically be larger than supply, and both demand and supply do not respond to waiting-time variations then the waiting list and the waiting time would keep growing over time.

There are different mechanisms through which waiting times affect the demand for and supply of healthcare (as measured, e.g., by the number of discharged patients). On the demand side, patients may opt for care in the private sector if they are not willing to wait and if they can afford to pay the (monetary) price charged by the private sector. Some patients may also simply give up the (public or private) treatment or opt for a pharmaceutical one. On the supply side, higher waiting times may induce providers to increase activity either because waiting times are used as key performance target (associated with penalties or monitoring) or because providers (doctors) feel bad about patients waiting for a long time.

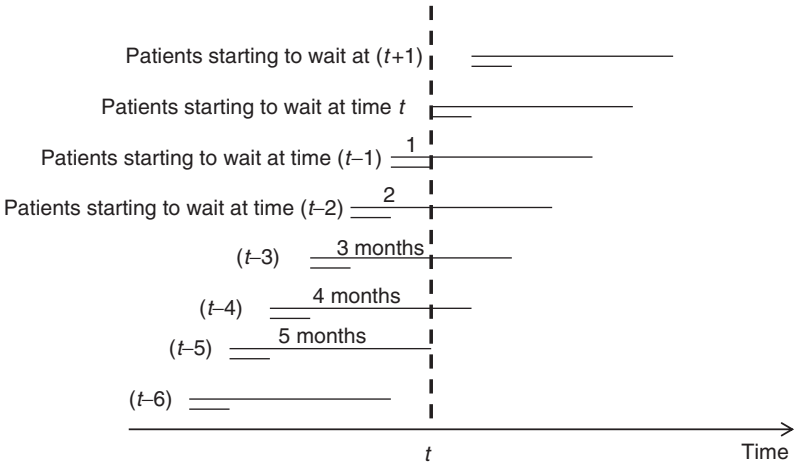


Figure 1 Distribution of waiting time of patients treated and on the list.

Perhaps, most importantly when waiting times are higher, policymakers may be willing to allocate more resources to increase the supply of healthcare (although this may also generate perverse incentives if providers artificially inflate the list to attract more resources).

A simple formal model is provided. Define $D(w)$ as the demand for healthcare which depends on the waiting time w. If assumed that higher waiting times reduce demand, then $D'(w)<0$. Similarly, the supply curve is defined by $S(w)$ with $S'(w)>0$. In such a market the equilibrium waiting time is determined such that $D(w^*)=S(w^*)$. The equilibrium waiting time is described in Figure 2 as point A.

Using this framework, the effect of exogenous shifts on the demand and the supply curve can also be explored. If, for any given waiting time, demand increases (an upward shift in the demand curve) then its equilibrium waiting time and activity will also increase. A higher demand implies that more patients are added to the waiting list, which ultimately leads to an increase in waiting time. Providers will respond to such an increase in waiting time by making more effort to increase activity (point B in Figure 2).

Similarly, if, for any given waiting time, supply increases (a shift to the right of the supply curve) then in equilibrium waiting time will reduce and activity will increase (point C in Figure 2). This shift may be due to an increase in capacity, as measured by an expansion in the number of hospital beds and doctors. A higher supply implies that more patients are taken out from the waiting list, which reduces waiting times; such a reduction is offset to some extent by an increase in demand.

Figure 2 also shows that the responsiveness of demand and supply to waiting times play a crucial role. If demand is very inelastic (nearly vertical, as drawn in Figure 3) an increase in supply will lead to a large reduction in waiting times. In contrast if the demand is elastic (nearly horizontal, as drawn in Figure 4), an increase in supply will lead only to a small reduction in waiting times.

Figures 3 and 4 illustrate a point which is critical in policy discussions on how to reduce waiting times. Some policymakers argue that increasing supply to the health sector will not reduce waiting times because the increase in supply will be offset by an equal increase in demand. As clearly depicted in Figure 4, this type of reasoning implicitly assumes that the demand is elastic. Others argue that a higher supply will bring down waiting times. As depicted in Figure 3, it implies that the demand is inelastic. Empirical estimates of how demand responds to waiting times are therefore paramount to decide whether increasing supply can or not significantly reduce patient waiting times.

Most of the existing empirical estimates of the demand curve from England put the elasticity of demand to waiting time at -0.1 (a 10% increase in waiting times leads to a reduction in demand by 1%). This suggests an inelastic demand curve: higher supply will lead to significant reductions in waiting times (as in Figure 3). It is worth emphasizing however that the existing empirical studies on demand elasticity have been derived mainly for the English healthcare system and should therefore not necessarily be transposed to other countries. One study from Australia, for example, has observed that demand was elastic to waiting times (with elasticity greater than one). Estimates of the supply elasticity even for England vary depending on the study (see discussion below).

Figures 2–4 also show the critical role of demand and supply shifts over time in determining the evolution of waiting times. In the health sector demand tends to increase over time. This is both due to the aging of the population that increases the healthcare needs and also due to technologIcal development, which makes new treatments available (some patients can be treated now who could not be treated in the past). This generates increasing pressures on waiting times. Supply may also increase over time due to the technological advancement that allows treating patients safely with less invasive treatments and a shorter length of stay. Whether waiting times

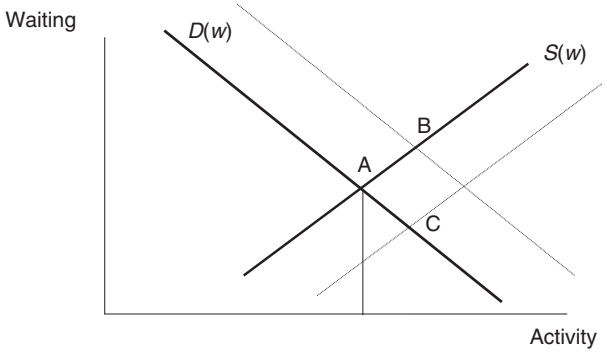


Figure 2 Demand and supply. A positive shock on demand implies a higher waiting time. A positive shock on supply implies a lower waiting time.

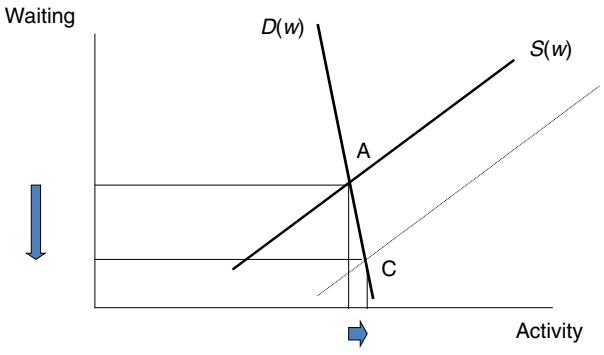


Figure 3 Inelastic demand.

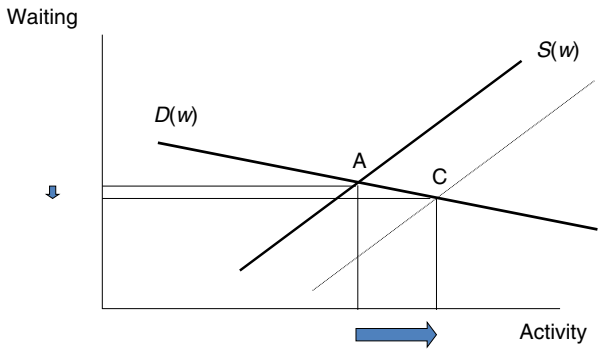


Figure 4 Elastic demand.

increase or reduce over time depends on the difference between the increase in demand and increase in supply. The gap between demand and supply may also depend on the type of system, with National Health Services being generally associated with tighter capacity constraints, and longer waiting times, than public (or private) insurance systems.

Policymakers have also tools available to influence the demand and supply of healthcare. For example, on the demand side countries with a gatekeeping system (where the patient has to see a family doctor before seeing a specialist) may have lower levels of demand. Moreover, eligibility to public treatment may be made conditional on certain criteria (like severity thresholds, benefits, appropriateness, and cost-effectiveness). More stringent criteria will lead to lower demand. Policymakers may also control the type of technology that is available to hospitals (and e.g. decide not to provide a certain treatment).

On the supply side, policymakers can influence supply by deciding the level of capacity in the public sector (the number of hospitals, doctors, nurses). The payment system for doctors and hospitals will also influence supply with fee-for-service rules and activity-based funding possibly being associated with a higher supply compared with a salary system and fixed budget rules.

Empirical Evidence

Estimates on demand and supply responsiveness to waiting times have been the subject of several empirical studies, mainly with data from England. They have used either a cross-sectional (i.e., a sample of hospitals or small areas at a point in time) or a panel-data (a repeated cross section over time) approach. The cross-sectional approach aims at identifying whether regions with higher waiting times have lower demand, for a given supply, or have higher supply, for a given demand. The analysis can be carried out either at a hospital level (variations across hospitals) or at a small-area level.

One problem with the estimation of demand and supply responsiveness is that separate measurements of demand for healthcare and supply are generally not available. What the researcher can typically observe, say at the hospital level, is different combinations of waiting times and activity levels. The researcher cannot tell whether the different activity levels reflect demand or supply variations.

To disentangle the differential effect of waiting times on the demand and supply of healthcare more sophisticated approaches than Ordinary Least Squares (OLS) need to be implemented. Estimating the responsiveness of the demand function to waiting times by regressing activity on waiting times and demand shifters (e.g., health needs or population) will lead to biased (distorted) results. Similarly, estimating the responsiveness of the supply function regressing activity on waiting times and supply shifters (e.g., the number of doctors and the number of beds) will lead to biased results. The endogeneity of waiting times in determining both demand and supply is typically addressed through the use of one or more 'instrumental' variables (a variable that is correlated with waiting times but not the dependent variable). In the demand equation this entails finding a variable that affects waiting times but does not affect the demand: a natural choice is a supply-shifter; for example, more doctors reduce waiting time but have no effect on demand. Similarly, in the supply equation the researcher needs to find a variable that affects waiting times but not the supply. In this case a suitable instrument is one or more demand shifters; for example, a higher proportion of elderly population will be correlated with waiting times but not with the supply.

Using a cross-sectional approach, Martin and Smith (2003) suggest that the elasticity of demand in the English NHS is between 0 and -0.2 depending on the year (-0.2 in year 1991–92, 0 in years 1993–94 and 1996–97, and between -0.12 and -0.15 in the remaining 4 years). The demand function is therefore inelastic: a 1% increase in waiting time reduces demand by at most 0.2%. In contrast, the supply function is rather elastic: between 2.1 and 5.9 depending of the year considered.

An alternative approach to cross-sectional analysis is the panel-data analysis, which involves data collection on activity and waiting times over several years (say approximately 3–7 years). In this case, the analysis tries to exploit how variations over time for a specific provider affect the demand and supply of healthcare. The results are very much in line with the cross-sectional one with an overall demand elasticity of -0.09 and a supply elasticity of 5.3. The elasticities also vary for different specialties; for example, the demand elasticity is higher (in absolute terms) for general surgery and oral surgery (equal to -0.24 and -0.21, respectively) and smaller for orthopaedics (-0.07).

Waiting Times versus Waiting List

The above discussion refers mainly to the duration of waiting times and not to the size of the waiting lists. This is for two reasons. First, from the patients' perspective what matters is how long they wait, and not necessarily how many patients are on the waiting list; if each patient waits for 2 weeks the disutility from waiting is the same regardless of whether there are 100 patients or 1000 patients on the list.

Second, waiting times and waiting lists do not necessarily move in the same direction. To illustrate this point, recall that the number of patients treated in equilibrium is $\gamma^* = D(w^*) = S(w^*)$. Define L as the waiting list. Note that at steady state (when the waiting time and the waiting list do not vary over time), $w^* = L^*/\gamma^*$. This expression is intuitive; it simply says that the waiting time for a typical patient is equal to the number of periods necessary to treat all the patients on the list. Therefore, if the waiting list has 200 patients and only 10 patients per week are treated, then the waiting time is 20 weeks.

Perhaps less intuitively, the waiting list can be rewritten as the product of activity and waiting time: $L^* = \gamma^* w^*$. Figure 2 can then again be used to investigate the effect of shocks on the demand and supply side on the waiting list. Figure 2 shows that an exogenous increase in demand increases both equilibrium waiting time and equilibrium activity (compare point A with point B). As the waiting list is the product of the two, then the waiting list will also increase. An exogenous increase in supply instead reduces waiting times and increases activity.

The waiting list which in the long run is given by the product of activity and waiting time can then either increase or reduce in response to an exogenous increase in supply (e.g., due to an increase in the number of doctors and beds).

Historically, waiting lists have been initially collected by policymakers as they were more easily available but in recent years a shift in focus on waiting times has been observed.

This section has used a static framework. The Section 'Waiting Times Dynamics over Long Periods of Time' expands such framework and allows demand and supply to increase over time driven, for example, by the technological development.

Competition and Choice

Many OECD countries have introduced in the last 20 years hospital payment systems of the Diagnosis Related Groups (DRGs) type where a public insurer (or the government) pays each hospital a tariff for every patient treated with a given diagnosis. This type of system is also known as activity-based financing. Patients typically receive treatment free of charge or are subject to small copayments.

As tariffs are the same across providers (they are regulated), hospitals have to compete on quality and waiting times to attract patients. Policymakers often argue that such competition among hospitals can also be beneficial to reduce waiting times. The intuitive idea is that under a competitive system, the provider will have an incentive to reduce waiting times to attract more patients and ultimately increase hospital revenues.

A prerequisite for implementing such competition policies is patients' choice; it is only if patients (or family doctors on their behalf) actively compare waiting times and quality across providers that providers will have an incentive to reduce waiting times to attract patients.

The higher the tariff paid to the hospital, the stronger will be the incentive to attract patients. DRG prices are normally based on average-cost rules. However, in some countries, like Norway, the price can be significantly less than the average cost (between 40% and 50%) in which case the incentive is reduced. Moreover, in some cases even countries that make use of DRG pricing have caps on the number of patients treated; the price can be significantly reduced once a certain volume of patients has been reached. In such cases the incentive to compete on waiting times can be significantly reduced.

Whether the hospital has a financial incentive to attract more patients by reducing waiting times depends on the difference between the tariff and the marginal cost. It is only if the profit margin is positive that the financial incentive is present. However, even if the profit margin is positive the provider may still lack the incentive to compete on waiting times. If the number of doctors is fixed and if the demand for treatment is high and there are strict targets on waiting times, doctors may be working at a point where the marginal utility from treating an extra patient is negative. In such cases, the introduction of competition may have an adverse effect on waiting times; intuitively by increasing waiting times, the provider can shift the cost of an unprofitable patient to another provider, which may generate a spiral towards high waiting times. Such results can be derived more formally through the stylized model provided in the Section 'A stylized Model of Hospital Competition'.

Empirical Evidence

There are two main empirical approaches to test the model outlined above. First, one prerequisite for competition to 'work' is that patients are willing to move when waiting times are higher. Empirical studies that model patients' choice suggest that at least in two countries (England and Norway) patients do react to variations in waiting time but the elasticity with respect to waiting time is low. The results are derived by regressing the hospital's choice of the patient (measured through a dummy variable equal to one for choosing a given hospital) as a function of waiting times in all hospitals in the catchment area of the patient (e.g., where the patient resides).

A second approach is to focus directly on the relationship between waiting time and competition, regressing waiting time of a given hospital on the number of hospital within the catchment area of the hospital. Using a cross-sectional framework one empirical study finds that in England an increase in competition by one hospital (from five to six hospitals) has a modest effect on waiting times reducing them by half week, from 17 to 16.5 weeks. A different study also tested the effect of competition on waiting times using a natural experiment framework in England: the control group includes providers who are 'local monopolists' and have no potential competitors within their catchment area. During 1997–99 competition (the policy of interest) was highly encouraged whereas during 1992–96 market boundaries were assigned geographically. The study shows that the introduction of competition leads to a reduction in waiting times by 0.8 months (see final section for detailed references).

Waiting Times and the Private Sector

Waiting times play a key role in the interaction between the public and the private sectors. Patients who are not willing to wait for a free public treatment may opt for the private sector paying a fee. Higher waiting times will lead to more patients being treated in the private sector. Anticipating that they will want care in the private sector, higher waiting times may also induce more individuals to buy private health insurance to cover the price charged by the private sector.

The incentive to set waiting times in the public sector may differ when a private sector is present or banned. Without a private sector, a public sector incentive to increase waiting times is such that the probability of idle capacity is optimally minimized. With a private sector a marginal increase in waiting times has the additional benefit (from the public provider perspective) of shifting patients to the private sector, which reduces costs in the public sector. Waiting times may therefore be higher in the presence of a private sector.

The incentive of the public sector to increase waiting times may be reinforced in the presence of 'dual practice', that is, when doctors are allowed to work in both the public and the private sectors. An increase in waiting times increases revenues in the private sector and generates a conflict of interest for doctors working in the public sector.

When the public and the private sectors interact, and patients differ in their severity, doctors may have an incentive to cream-skim the patients in the private sector to gain higher profits. Which patients are cream-skimmed depends on the rationing rule used by the public provider and the threshold severity level over which patients are entitled to public treatment. Intuitively, it would be expected that patients with the lowest severity to be cream-skimmed in the private sector. This may not always be the case. Patients with lower severity may not be willing to pay the price charged by the private sector and may also not be eligible for public treatment. Incentives to cream-skim will be highest for patients with middle severity when the severity threshold required for public treatment is high; for those patients severity is high enough to be willing to pay the price in the private sector but not high enough to be eligible for public treatment.

Maximum Waiting-Time Guarantees

In a number of countries, in response to rising waiting times, policymakers have introduced maximum waiting-time guarantees. In its simplest form, a maximum waiting time guarantee says that no patient should wait for more than a predetermined number of months or weeks. This has been the case, for example, in England. Such guarantees are 'unconditional;' they hold for every patient regardless of the treatment.

One problem with such unconditional guarantees is that they may conflict with clinical prioritization. As they refer to a maximum, the patients who benefit from this guarantee are likely to be those with a lower priority, who indeed wait the longest. There is therefore a risk that patients with lower severity may be given priority over patients with higher severity, who tend to wait less. Empirical evidence from England suggests that this is indeed the case; the probability of being admitted for treatment increases as it approaches the target but it decreases after the target.

To address this issue, unconditional guarantees have been replaced in some countries (like Sweden in the 1990s) with 'conditional' ones where the guarantee is given only conditional on having a certain severity. The maximum waiting time can differ between different severity groups (normally three or four). The most sophisticated form of conditional guarantee is one that specifies a 'personalized' waiting time (like in Norway), which depends on several criteria including urgency, benefit and cost-effectiveness from treatment. In contrast to unconditional ones, conditional guarantees conflict less with prioritization (and can actually be seen themselves as prioritization rules). They however suffer from their own limitations: as the provider has discretion over assigning a guarantee to a certain group and a provider can use such discretion to make sure it complies with the guarantee. One possibility to reduce the scope of such discretion is to develop explicit guidelines (though developing a guideline is quite a costly procedure as it involves extensive consultations).

One common problem to maximum waiting-time guarantees is how to enforce them. Although it is quite simple to state that no one or a subset of patients should not wait for more than a predetermined amount of time, it is not clear why the provider should have the incentive to respect such guarantees. The empirical evidence suggests that maximum waiting times work only when they are in the form of targets and clear penalties are attached to them. In England, one study made use of a 'natural experiment' to test whether or not a waiting-time target policy combined with sanctions for hospital managers resulted in a reduction of hospital waiting times. Scotland was used as the control group. The study finds that 'targets and terror' policy significantly reduced the number of patients waiting more than 6 and 12 months by 20% and 60%, respectively.

Equity Issues

This article has argued in the Section 'Waiting Times as a Mechanism to Bring the Healthcare Market to an Equilibrium' that waiting times act as a rationing device that brings about an equilibrium to demand and supply. It can also be argued that, despite waiting times generating disutility to patients, one advantage of rationing by waiting is that such form of rationing is equitable because the ability of affording such nonmonetary price does not depend on the ability to pay. This is in contrast to price rationing, for example, in the form of a copayment, which can be more easily afforded by richer individuals.

A recent empirical literature suggests, however, that waiting times may not be as equitable as they appear because individuals with higher socioeconomic status (usually measured by income or educational attainment) tend to wait less, and therefore pay a lower nonmonetary price, in publicly funded hospitals, than patients with lower socioeconomic status. Such gradient is found in separate studies in Australia, England, Norway, and Sweden and may be interpreted as evidence of inequity, which favors the rich and more educated patients over poorer and less educated ones.

One advantage of copayments over waiting times to contain demand, as opposed to waiting times, is that the cost to patients generated by long waiting times is not necessarily recovered by anyone else (except from the reduction in idle capacity that are exhausted quite rapidly). In contrast, copayments raise resources for the provider, which can be recovered by the government. Moreover, copayments could be income-tested to address equity concerns though there may be administrative costs associated with it.

A final argument in favor of waiting times is in terms of redistribution. Waiting times induce some better-off patients to opt for the private sector. Patients who opt for the private sector pay twice; they pay the price to receive the treatment in the private sector and also pay taxes, which contribute to the funding of public health systems. If governments have limitations to the extent to which they can redistribute between different income groups, the presence of waiting times may then help to redistribute resources form the rich to the poor (although indirectly).

Conclusions

Waiting times are a pervasive feature of many public-funded healthcare systems and increase with the gap between demand

and supply. In future, demand is likely to grow driven by the aging population and by the technological advancements. In contrast, given the current economic climate, there are limits to the extent to which governments can allocate additional resources to increase supply or identify significant efficiency gains. Waiting times may therefore be on the rise. As waiting times generate a significant dissatisfaction among patients and the general public, optimal demand management and patients' prioritization will play a key role in future policy developments.

Technical Appendix

Waiting Time of Patients on the List versus Waiting Time of Patients Treated: An Example

To illustrate the difference in the distribution of waiting times for the patients on the list versus the distribution in waiting times of patients treated, a simple illustrative example (based on Dixon and Siciliani, 2009) is provided. For expositional clarity, one period is referred to as 1 'month.'

Suppose that in each period there is a fixed number of patients who enter the waiting list, which is normalized to one, and an equal number who are treated (therefore a steady state is assumed). Suppose also that the proportion of patients entering the waiting list at any point in time and waiting more than 1 month is p_1, so that $(1-p_1)$ get treated in the first month. In the second month, conditional of having waited for 1 month, a proportion equal to $(1-p_2)$ get treated and p_2 keep waiting. By the third month, everyone is treated so that $p_3=0$ or $(1-p_3)=1$: no patient waits for more than three periods.

The distribution of the patients treated at any point in time (say in any given month) is the following. There are $(1-p_1)$ patients waiting for 1 month, $p_1(1-p_2)$ waiting for 2 months, and the remaining (p_1p_2) patients waiting for 3 months. The average waiting time of the patients treated (AWT), measured in months, is therefore

$$\text{AWT} = 1 \times (1-p_1) + 2^{*}p_1(1-p_2) + 3^{*}p_1p_2 = 1 + p_1 + p_1p_2$$

The length of the waiting list at any point in time, denoted by L, is equal to $L=1+p_1+p_1p_2$. At any point in time there is a whole cohort of patients (equal to one) who waits for 1 month (i.e. the minimum wait), plus those who started to wait the previous period (equal to p_1 as $1-p_1$ were treated in the previous period) and those who started to wait 2 months before (equal to p_1p_2). The average waiting time of the patients on the list (AWL) is therefore equal to

$$\text{AWL} = (1 + 2^{*}p_1 + 3^{*}p_1p_2)/L$$

It is straightforward to show that the AWT>AWL if $(p_1(1+p_2)^2 - p_2)p_1 > 0$. The comparison is, in general, indeterminate and depends on p_1 and p_2.

As a numerical example suppose that $p_1=0.2$ and $p_2=0.8$, which implies that only 20% of the patients keep waiting to the second month, and conditional on having waiting for 1 month, 80% keep waiting for 3 months. Then, AWT=1.32 months and AWL=1.38 months, and the average waiting time of the patients on the list is higher than that of patients treated. The opposite holds if $p_1=0.4$ and $p_2=0.8$.

As the first measure refers to an 'incomplete' measure it would intuitively be expected to be shorter than the second measure. This is however not necessarily the case. When looking empirically at the distribution of waiting time it may well be the case that the proportion of patients waiting more than 6 months or the average waiting time for the patients 'on the list' is actually larger than those treated. This arises because 'long waiters' are overrepresented (or oversampled) under the first measure, which causes the waiting time to go up.

Waiting Times Dynamics over Long Periods of Time

This section provides a simple stylized model of waiting-time dynamics, which illustrates the long-run determinants of waiting times. The analysis provided above (and in Figures 2–4) is static as demand and supply depend on waiting times and not directly on time. Changes over time have been analyzed only through exogenous shocks on demand and supply. As argued above, both demand and supply can vary over time possibly at different rates due to the technological development and a range of other factors. A simple way to model time-varying demand and supply function is to assume that they are linear and, respectively, equal to $D(w,t)=a+bt-cw$ and $S(w,t)=d+et+fw$, where t is the time and a, b, c, d, e, and f are positive parameters. Other parameters a and d can be interpreted, respectively, as demand and supply at time zero when the waiting times are zero; c and f as the responsiveness of demand and supply to waiting times; b and e as the degree at which demand and supply increase over time due, for example, to technology developments. The waiting time dynamics can be formally represented by the following differential equation:

$$\frac{\partial w}{\partial t} = D(w,t) - S(w,t) = (a-d) + (b-e)t - (c+f)w$$

which suggests that waiting times increase over time when demand is higher than the supply and reduce when waiting times are lower than supply. The closed-form solution of the above equation (assuming $a=d$ to keep the exposition simple) is as follows:

$$w(t) = \left[w_0 + \frac{b-e}{(c+f)^2} e^{-t} - \frac{b-e}{(c+f)^2} \right] + \frac{b-e}{(c+f)^2} t$$

The first term in the square bracket goes to zero as time passes. The long-run dynamics of waiting times is then driven by the difference between b and e, that is, the difference in the speed at which demand and supply grow over time. Intuitively, if demand grows faster than the supply, then waiting times will increase over time. *Vice versa*, if supply grows faster than the demand, then waiting times will reduce and eventually disappear. The smaller the growth in waiting times, the higher is the response of demand and supply to waiting times.

Given the current economic climate, public budgets allocated to healthcare may reduce or stagnate, and therefore slow down the growth in the supply of healthcare. As a result, waiting times may be on the rise. Policymakers will have to

either seek efficiency gains on the supply side or intervene on the demand side to prevent rapid growth of waiting times.

A Stylized Model of Hospital Competition

Some of the arguments provided in Section 'Competition and Choice' can be illustrated through a simple stylized model (adapted from Brekke *et al.*, 2008). Suppose that there are two hospitals i and j. It is assumed that the total demand is inelastic and equal to 1 (which can be thought of as 100% of the patients being treated) and every patient receives treatment in one of the two hospitals. A Hotelling set up is used, which assumes that the two hospitals are located at the extremes of a unit line is used. The demand function of hospital i is given by $D_i(w_i,w_j) = \frac{1}{2} - (w_i - w_j)/2t$, where w_i and w_j are, respectively, the waiting times in hospital i and j. If waiting times are identical, then each hospital has half of the market. If waiting times are higher in hospital i, then the demand in hospital i is less than half of the market. The parameter t is a transportation (or other) cost parameter, which can broadly be interpreted as the extent to which patients are willing to switch from one provider to the other. Lower transportation costs imply that patients are more willing to switch from one provider to the other when waiting times are lower. This parameter can be influenced by policymakers: making, for example, waiting-time information easily available to patients is equivalent to a reduction in transportation costs.

Assume that the payoff function of the hospital, defined by U, is given by the difference between revenues and the sum of monetary and non-monetary costs, which is given by the following expression:

$$U_i(w_i,w_j) = T + pD_i(w_i,w_j) - C(D_i(w_i,w_j),w_i) - \frac{k}{2}w_i^2$$

where p is the price for each patient treated, T is a fixed-budget component, $C(\cdot)$ is the cost function of treating patients (discussed in more detail below), k is a parameter that is proportional to the penalties from having long waiting times (e.g., deviations from waiting-time targets): higher waiting times imply a higher disutility in the form of more monitoring and a higher threat of dismissal for senior managers. To keep the presentation simple assume that the cost function $C(\cdot)$ is quadratic:

$$C(D_i(w_i,w_j),w_i) = F + \frac{c}{2}D_i(\cdot)^2 - zw_i + \frac{d}{2}w_i^2$$

where F is a fixed cost, and c, d, and z are positive parameters. Assume also that a higher activity increases costs at an increasing rate $(C_{D_i}(\cdot) = cD_i)$. This cost function captures both monetary and non-monetary costs. As in public hospitals the number of doctors is fixed, it is plausible that the marginal disutility (i.e. the nonmonetary marginal cost) from treating an extra patient is increasing. This specification also allows for the possibility that higher waiting times reduce costs by lowering the probability of idle capacity $(C_{w_i}(\cdot) = -z + dw_i)$. When waiting times are zero, a marginal increase in waiting reduces costs. However, this happens at a decreasing rate and there may be a point where higher waiting times increases costs due to the higher costs of managing the waiting list.

Differentiating the utility function, the following condition for optimal waiting times is obtained:

$$(p - cD_i(w_i,w_j))\left(-\frac{1}{2}\right) - kw_i + (z - dw_i) = 0$$

A marginal reduction in waiting times generates benefits and costs. On one hand, it increases activity which generates higher revenues and also makes waiting-times target figures look better. On the other hand, it increases monetary and nonmonetary costs either directly through higher activity or indirectly through higher idle capacity. In the symmetric equilibrium,

$$w^* = \frac{z + (p - (c/2))(-1/2t)}{k + d} > 0$$

which is always positive whenever the price p is not too high (or the cost c not too low). As expected, stricter targets (higher k) or higher prices p reduce waiting times, and higher costs (c) increase waiting times.

Critically, in equilibrium the price-cost margin $(p - (c/2))$ can be positive or negative. This depends on the price and the steepness of the marginal cost. Figure 5 illustrates the two possible cases with one price p and two marginal cost curves being characterized by $c_1 > c_2$. When the marginal cost is higher (lower), the profit margin is negative (positive).

How does more competition affect waiting times? Policies that encourage competition can be thought of as policies which reduce the transportation costs or the cost of switching between providers. Differentiating the equilibrium waiting time with respect to transportation cost parameter yields the following:

$$\frac{\partial w^*}{\partial t} = \frac{(p - (c/2))}{k + d}\left(\frac{1}{2t}\right)^2$$

Whether more competition increases or reduces waiting times depends on the price-cost margin $(p - (c/2))$. More competition (lower transportation costs) reduces waiting times only if the margin is positive (i.e., the price is sufficiently high). In contrast, more competition increases waiting time if the margin is negative (i.e., the price is sufficiently low). The intuition for the latter result is that if providers are working at a negative margin, then under more competition a marginal increase in waiting times becomes more effective in shifting

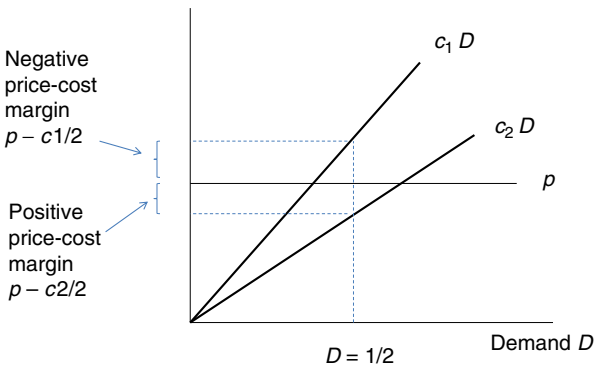


Figure 5 Price-cost margins.

unprofitable patients to the other provider. In contrast, if it is positive a reduction in waiting times attracts profitable patients.

The above analysis assumes that providers are altruistic, which is a realistic and plausible assumption for healthcare providers. Introducing altruism generates two additional effects. On one hand, altruism makes the providers even more willing to work at a negative profit margin, which in turn tends to reinforce the result that competition increases waiting times. On the other hand, altruism induces providers to compete for patients because they can do more 'good.' This effect goes in the opposite direction.

See also: Competition on the Hospital Sector. Interactions Between Public and Private Providers. Moral Hazard. Primary Care, Gatekeeping, and Incentives. Rationing of Demand. Specialists

References

Brekke, K., Siciliani, L. and Straume, O. R. (2008). Competition and waiting times in hospital markets. *Journal of Public Economics* **92**, 1607–1628.

Dixon, H. and Siciliani, L. (2009). Waiting-time targets in the healthcare sector. How long are we waiting? *Journal of Health Economics* **28**, 1081–1098.

Martin, S. and Smith, P. C. (2003). Using panel methods to model waiting times for National Health Service surgery. *Journal of the Royal Statistical Society* **166**, 1–19.

Further Reading

Barros, P. P. and Olivella, P. (2005). Waiting lists and patient selection. *Journal of Economics and Management Strategy* **14**, 623–646.

Dimakou, S., Parkin, D., Devlin, N. and Appleby, J. (2009). Identifying the impact of government targets on waiting times in the NHS. *Health Care Management Science* **12**, 1–10.

Gravelle, H., Smith, P. C. and Xavier, A. (2003). Performance signals in the public sector: the case of health care. *Oxford Economic Papers* **55**, 81–103.

Iversen, T. (1997). The effect of private sector on the waiting time in a National Health Service. *Journal of Health Economics* **16**, 381–396.

Marchand, M. and Schroyen, F. (2005). Can a mixed health care system be desirable on equity grounds? *Scandinavian Journal of Economics* **107**, 1–23.

Martin, S. and Smith, P. C. (1999). Rationing by waiting lists: An empirical investigation. *Journal of Public Economics* **71**, 141–164.

Propper, C., Burgess, S. and Gossage, D. (2008). Competition and quality: Evidence from the NHS internal market 1991–9. *The Economic Journal* **118**, 138–170.

Propper, C., Sutton, M., Whitnall, C. and Windmeijer, F. (2008). Did targets and terror reduce waiting times in England for hospital care? *The B.E. Journal of Economic Analysis & Policy* **8**(2), (Contribution, Article 5).

Sharma, A., Siciliani, L. and Harris, A. (2011). *Waiting times and socioeconomic status: Does sample selection matter?* HEDG Working paper.

Siciliani, L. and Martin, S. (2007). An empirical analysis of the impact of choice on waiting times. *Health Economics* **16**, 763–779.

Water Supply and Sanitation

J Koola, Harvard University Kennedy School of Government
AP Zwane, Bill & Melinda Gates Foundation

Glossary

Chlorine Is sodium chloride, used as a disinfectant to clean contaminated water.
Point of use Is the location or time in which the water is actually used or consumed.
Sanitation Is the practice of eliminating contact between humans and urine and feces.
Wells Are an excavation or structure built to provide access to groundwater at the surface.

Introduction

Water supply and sanitation are at their core a public health issue. Every year, 2.2 million people die from diarrheal diseases, a leading cause of which is unhygienic water and sanitation. Improving health and mitigating diarrheal morbidity and mortality is the underlying rationale for water and sanitation Millennium Development Goals, which call for reducing by one half those who lack access to safe and sustainable water and sanitation. In the US, David Cutler and Grant Miller have shown that piped water, centralized water treatment, and waterborne sanitation interventions were jointly responsible for most of the rapid decline in the child mortality rate in the early twentieth century. Tara Watson has demonstrated more recently the link between water and sanitation coverage and substantial health improvements on Native American Indian reservations.

The outstanding question is not whether water and sanitation prevent diarrheal diseases in a biomedical sense, but what interventions are appropriate and effective in settings in which piped water and sanitation are unavailable because of their expense. In much of Africa, rural residents typically live on their farms rather than being concentrated in villages. In such circumstances, policy often focuses on providing improved drinking water sources outside the home, such as communal taps, wells, and protected springs. In developing countries in rural Asia and Africa, feasible sanitation interventions are generally latrines of some sort. The health impacts and cost-effectiveness of these interventions and the urban equivalents of shared standpipes and septic tanks or shared toilet blocks are less well understood compared to our understanding of the epidemiological transition in developed countries. The challenge in identifying appropriate institutional mechanisms for service provision is also great.

This article critically reviews the evidence on the health and nonhealth benefits of water and sanitation investments that are common in developing countries. Evidence on valuation for these interventions has been explored, and the implications of the evidence on valuation for government policy to support improvements in this sector have been discussed. Some implications for a future research agenda have also been briefly highlighted.

Water Supply

Water Quantity

This article considers separately how quantity and quality of water affect health because different water supply interventions affect water quality and quantity asymmetrically. For example, adding chlorine to water affects quality but not quantity. In contrast, providing household connections to municipal water supplies to households that currently use standpipes is likely to have a bigger effect on the convenience of obtaining water, and thus on the quantity of water consumed, than on water quality. Much of the most convincing nonexperimental evidence on the health impact of water and sanitation makes it difficult to separate the impact of quantity and quality because the interventions studied both reduced the cost of collection and improved quality, making it unclear which route of disease transmission matters the most in practice.

In the 1980s and 1990s, nonrandomized studies were frequently cited as evidence that water quantity was more important for health impacts than was water quality. Some researchers argued that these results could be explained because increased availability and convenience of water facilitate frequent washing of hands, dishes, bodies, and clothes, thereby reducing disease transmission. There is indeed strong evidence that hand washing is important for health. However, it is difficult to assess the causal impact of water quantity on hand washing in the absence of randomized evaluations or other convincing identification. In the Section Water Supply, numerous randomized evaluations that have shown impacts of improved water quality on health is discussed.

Although impacts may be heterogeneous across settings, and caution is warranted in drawing general conclusions, the one available randomized evaluation finds that increasing the quantity of water while maintaining unchanged quality did not lead to significant health improvements. Researchers, Florence Devoto, Pascaline Dupas, Esther Duflo, and William Pariente examine provision of piped connections to homes in urban Morocco previously served by public taps. This increased the quantity of water used by the household, but did not improve water quality, because the alternative, chlorinated

water from communal taps was of similar quality to the water received at home.

As part of a planned piped water service extension in Tangiers, Morocco, these researchers randomly selected half the households eligible for a first connection to receive (1) information about and an offer of credit toward a new connection and (2) administrative assistance in applying for credit. Take up was 69% (compared with 10% in the control group).

These researchers compared outcomes of those who received this treatment with outcomes for households in the control group. They found that piped water provision in this urban Moroccan context had few health benefits. There was no evidence for an impact of treatment on a subjective ranking of health of the family or on diarrhea in children under 6 years of age (although baseline rates were relatively low, with the average child in the control group experiencing 0.27 days of diarrhea in the past week). Households in the treatment group reported increasing their frequency of baths and showers: The number of times that respondents in the treatment group washed themselves (through baths or showers) during the past 7 days was 25% higher than in the control group. However, hygiene practices that required less water, such as hand washing, were not affected, according to self-reports.

It is not to be concluded that increased water quantity never yields health benefits. The benefits of increased water quantity may be context specific and require further research for a complete understanding. In particular, understanding when and how increased access to water leads to more hand washing is a research priority.

In the study mentioned above, having a piped water connection had substantial private benefits, despite the lack of impact on self-reported diarrhea, consistent with the evidence that most households that received information and an offer of credit toward a new connection were willing to pay for it. In particular, piped water connection saved time, which was spent for leisure and social activities. Measures of social integration and overall welfare improved. One year into the program and it was noted that, not only did the encouragement design result in high rates of take up in the treatment group, but also for these households, their average monthly water bill more than doubled, from 73 to 192 Moroccan dirhams, or US$9–24 a month (the previous cost came from households that took water from their neighbors). Other authors also note evidence of substantial willingness to pay for water quantity in observational studies.

Water Quality

Several randomized evaluations find that improvements in water quality reduce reported diarrhea. One study by Jessica Leino, Michael Kremer, Edward Miguel, and Alix Zwane examines source water quality improvements. The researchers estimate that protecting springs reduced fecal contamination, as measured by the presence of *Escherichia coli* bacteria, by two-thirds for water at the source but by only 25% for water stored at home. This is likely due in large part to recontamination in transport and storage within the household. Despite the incomplete pass-through of the water quality improvement, mothers reported approximately 25% less child diarrhea in the treatment group. The importance of recontamination suggests either to treat water at the point of use, close to the time of use, or to treat water in a way that provides residual protection, for example, with chlorine at a sufficiently high dose to remain at levels that provide disinfection for at least 24 h.

Household water treatment at the point of use, for example, with filtration or chlorine treatment, also reduces child diarrhea. The bulk of the evidence suggests that, with take up rates on the order of 70% (achieved via frequent visits and reminders to subjects), household water treatment reduces child diarrhea by 20–40%. There are multiple comprehensive reviews of this literature. Some question the validity of this literature because the outcome measure in these studies is typically mothers' reports of child diarrhea. Studies with objective outcomes, infrequently measured, would be preferable. However, the extent of reporting bias in treatment groups would have to be very large to explain the reported reductions in diarrhea associated with cleaner water. To the extent that reporting bias lowers estimates of diarrhea in treatment and comparison groups, such bias may make it harder to statistically detect reductions in diarrhea. If the reductions in diarrhea were even a fraction as large as those estimated, water treatment would still be very cost-effective.

Because water treatment can be extremely cheap, even a 20–40% reduction in diarrhea makes water treatment extremely cost-effective. For a sense of how cheap it is to treat water, a 1.42-Ga generic bottle of bleach with approximately 6% sodium hypochlorite concentration sold at Walmart for $2.54 as of December 2009 has enough chlorine to treat 163 400 l of water. This corresponds to a price of $0.00002 per liter of water treated. Actual costs of treatment with chlorine are higher because chlorine used for treatment is normally at lower concentrations and the concentration quality has to be made more consistent. Nonetheless, under the assumptions that chlorination reduces diarrhea by 20–40% and that mortality reductions are proportional to reported morbidity reductions, the cost per disability-adjusted life year (DALY) of chlorine provision using the traditional social marketing approach is less than $40, considerably less than the benchmark of $100–150 per DALY saved that is typically used in health planning in low-income countries.

Sanitation

Health Impacts – Diarrhea

The methodology used to evaluate sanitation programs is diverse but generally weaker than that which has been used to identify the impacts of water on health. Many studies are cross-sectional, some are longitudinal, others compare differences between two 'matched' groups, but very few rely on a randomized design. Studies of sanitation often try to address omitted variable bias by controlling for variables suspected of being confounding or by 'matching' subsamples. This method can only attempt to control for observable characteristics, such as maternal education or income, but will not control for unobservable characteristics, such as health attitudes.

Moreover, controlling for all possible confounding variables which could plausibly influence an outcome of interest is an impracticable undertaking.

Two more shortcomings of impact studies in the sanitation sector is that they neither have adequate sample size nor do they account for the fact that the interventions are provided at the community, rather than the household level. A large sample size is necessary to control for many factors in a multivariate regression. When interventions are targeted at the community level, individual households cannot be considered discrete observations because households within a community are likely to resemble each other, termed intercluster correlation. This 'clustering' means that a statistical test has less power to determine the existence and size of a treatment effect. Researchers point out that many early evaluations of sanitation compare an intervention village to a 'control' village, but this is tantamount to a sample size of two individuals because households within a village are not independent units of observation. Early reviews of health impact studies of water and sanitation interventions pointed out these methodological flaws and others.

Although there is broad consensus that improving sanitation will have a positive health impact, very few studies have established this causal effect in practice. In two survey papers, several authors critically examined the evidence for the health impacts of water, sanitation, and hygiene interventions alone or in combination. Despite the limited number of studies on sanitation interventions, these reviews suggest that sanitation is effective in reducing diarrheal illness, with a pooled relative risk estimate of 0.678 (with a 95 confidence interval of 0.529–0.868). Yet, the research on which these reviews are based has significant shortcomings. Of the four studies identified as sanitation specific, three were classified by the authors as poor quality. For example, two of these did not have an adequate control group and there was not clear or convincing measurement of confounding variables in two others. The only intervention studies were not randomized and used heterogeneous interventions and methodologies that prevented pooling of results. Moreover, nearly all combined sanitation with water or hygiene interventions, making it impossible to determine the contribution of sanitation alone.

In the ensuing years since these reviews, research quality has not improved dramatically. In a recent systematic review, researchers stressed the lack of rigorous evidence on the contribution of sanitation interventions to prevent diarrhea in young children. Others come to the same conclusion after conducting a Cochrane review of excreta disposal interventions to prevent diarrhea. A cohort study on the effectiveness of a city-wide sanitation intervention in Salvador in Northeast Brazil suffers from several of the methodological flaws reviewed above. For example, there is no external control group; the children, aged 0–36 months, are recruited from households where the intervention took place. The researchers simply compare the prevalence of diarrhea among two different cohorts of similarly aged children before and after the intervention and conclude that the intervention reduced diarrheal rates in children by 21%. Another shortcoming of the study's design is that the main outcome of interest is self-reported diarrhea, which is problematic because of respondents' recall bias and the bias introduced by frequent interactions with researchers. The single exception, to our knowledge, is a randomized-controlled trial of a community-lead total sanitation campaign (TSC) in 40 villages in Orissa, India. This study finds that the campaign increased child mid-upper arm circumference (MUAC) z-scores by roughly 0.25 standard deviations. However, the study lacked sufficient power to detect effects on diarrheal disease in children.

Health Impacts Beyond Diarrhea – Nutritional Status and Parasitic Infection

Recently, nutritionists including Jean Humphries have hypothesized that reducing a child's fecal bacteria exposure during the first years of life through improved sanitation (and/or handwashing or water treatment) may improve gut function (the ability of the gastrointestinal tract to absorb nutrients) and subsequent growth. The prenatal period and the first 2 years of life are a critical window for intervention in growth and development: infection and poor nutrition during this window can negatively impact an individual's long-term cognitive development and lifetime physiologic trajectory. The new hypothesis is that nutritional supplementation seems to be necessary but not sufficient to eliminate growth shortfalls because chronic infection and colonization of the gut by fecal bacteria, spread via poor conditions, impedes nutrient absorption and creates low-level immune system stimulation, a condition called environmental (or tropical) enteropathy.

If the environmental enteropathy hypothesis were to be correct, this would significantly alter our understanding of the health benefits associated with sanitation, and increase the estimated cost-effectiveness of these interventions. The change would likely be very large, because of the lifetime gains associated with better nutrition in early childhood.

Recent rigorous research suggests that avoiding worm infections in childhood brings not only better health outcomes, because of reduced anemia, but also better nonhealth outcomes. In an article with the memorable title 'Worms at Work,' randomized control trial data show that children without worms are less fatigued who go to school more and ultimately have higher incomes as adults. Complete sanitation services certainly reduce worm loads and infection rates; however, it is uncertain whether partial sanitation coverage is sufficient to do this. To the extent that feasible, affordable sanitation interventions reduce worm infections, the benefits associated with these sanitation investments may be much greater than suggested by health benefit calculations alone because of the associated lifetime schooling and wage increases.

Nonhealth Benefits of Sanitation

Limiting the gains from sanitation to health benefits, however, ignores a growing body of evidence that indicates that people realize substantial nonhealth benefits from improved sanitation, such as convenience, safety, and dignity. Lack of facilities for defecation is a problem faced by everyone, but differing norms for behavior and modesty of women can make this especially problematic for them. Women and girls may be expected to defecate only when it is dark, which can increase urinary tract infection rates, chronic constipation, and

psychological stress. Leaving the home for secluded areas after dark also makes women and girls vulnerable to physical assault. The challenges of menstrual management and the symptoms of the postnatal period compound these problems. Measuring the quality of life benefits for women from reduced stress or shame when sanitation services are available is very hard to do. Women and children in particular are targeted as primary users of sanitation for reasons including their privacy and security.

Suggestive evidence from the economic and sociological literature demonstrates that women's participation is important for providing and managing local public goods. These frequent associations between women, sanitation, and public goods imply a need for a greater understanding and further research on both women's and men's roles in sanitation provision and behavior change.

Elastic Demand and the Subsidy Debate

This section considers the case for subsidies in the water supply and sanitation sector, with a particular focus on sanitation. Willingness to pay for water quantity has been well established, including in the setting in Morocco discussed above. Evidence on valuation of water quality has been summarized elsewhere. Subsidies for sanitation are currently at the center of a policy debate because of a new push to meet the MDGs and rethink old ways of doing business in that sector.

Supply-side interventions along with hardware subsidies have been the typical approach to rural sanitation. Traditional public finance reasoning supports such an approach on both efficiency and distributional grounds. Subsidizing sanitation hardware has been justified on the grounds that the public benefit far outweighs the private one. In addition, the high upfront costs of sanitation infrastructure necessitates subsidies for the credit-constrained poor. Even with hardware subsidies, the extra transportation and construction costs associated with latrine building may still be prohibitive for the extremely poor. One study in Mozambique found that the additional transportation and construction costs for a latrine made the already subsidized cost (only 5% of the household's average monthly income) into a 'medium' cost which was too burdensome majority of the poor population.

Recent funding trends have begun to disavow subsidies in the sanitation sector. Community-led total sanitation, for example, is a new participatory methodology for sanitation behavior change which discourages the use of subsidies and emphasizes cost sharing by households instead. There are several nonrandomized, cross-sectional studies that have been frequently cited in the gray literature as evidence to support the case for cost recovery, rather than subsidies. One example of this comes from a report conducted by the Water and Sanitation Program in South Asia that compiled lessons learned from eight case studies of rural sanitation programs in Bangladesh, India, and Pakistan. Six of the eight case studies examined the TSC in India. The report compares programs offering full hardware subsidies (e.g., Andhra Pradesh TSC) to programs offering only partial subsidies (e.g., West Bengal TSC) and programs with full cost recovery (e.g., Plan Bangladesh). The analysis relied on data from interviews with key stakeholders and observations at selected villages that were included in the program and villages that were not included in the program. As emphasized above, such a methodology does not create a credible counterfactual. The study concludes that there is an association between high subsidies and poor program performance: the two worst performing programs (Andhra Pradesh TSC and Pakistan's Lodhran Pilot Project) had the highest hardware subsidies. The report goes on to say that high hardware subsidies 'reduce the sense of ownership by those that receive the heavily subsidized facilities' and that there is 'increasing evidence that (subsidies) tends to result in low toilet usage and wasted investments.' Others also argue charging a fee for sanitation infrastructure will induce adoption, ownership, and more use.

There are three assumptions underlying the argument that cost recovery increases usage. First, higher prices may act as a 'screening' mechanism for those who most need the product because households will self-select into purchase. Second, positive prices are interpreted as a signal of quality. And third, purchase will increase use because households rationalize purchases *ex post* because of a 'sunk-cost' effect. The sunk-cost effect implies that the household pays a psychological cost whenever it buys a product that it subsequently does not use. Several researchers use similar methodologies to exploit variation in offer prices and purchase prices for water treatment products and bed nets, respectively, to test for evidence of the screening and the sunk-cost effects. Both studies agree that cost sharing drastically reduced demand and does not target those most in need. For bed nets, cost sharing did not reduce wastage. Higher willingness to pay is associated with a greater propensity to use the water treatment product. These two studies add to a growing body of empirical evidence which refutes the simple claim that a positive price effectively targets households who are more in need of a produce or makes use of the product more likely.

Whether and how much to subsidize sanitation falls into a much broader debate about cost sharing for health products in general. The empirical evidence points to an income elasticity of demand for health that is positive and above one in both developed and developing countries, which implies that even a small increase in income would result in a very large increase in the demand for the good.

Another consistent finding in the experimental literature is that there is a very steep demand curve for preventative health products, such as bed nets, water treatment, or deworming pills. The price elasticity of demand around zero is very large for people in lower income groups, meaning that even a small change in price can create a very large effect on demand for the good. For example, it was found that demand for deworming drugs dropped from 79% to 19% when the price was raised from 0 to 30 cents. Similarly, others found that raising the price of bed nets from 0 to 60 cents lowered demand by 60% points in Kenya and that a 13 cent increase in the price of water treatment lowered demand for the product by 30% points in Zambia. These dramatic findings would seem to defend full subsidies of health products with positive externalities for poor households.

The theoretical literature on externalities and the empirical evidence discussed above would seem to suggest that sanitation should continue to be subsidized. Subsidies may also

have potential downsides, and these drawbacks in the sanitation sector need to be rigorously tested. Subsidies may lead to inefficient allocation in two ways. First, households may delay their decision to purchase and install sanitation infrastructure because of the expectation of receiving a future subsidy. And second, subsidies that follow purchase or construction may lead to overconsumption of the good to benefit from the subsidy. For example, a review of a sanitation progress in Andhra Pradesh, India, found that households had built expensive toilets to receive the subsidy, but abandoned them soon after. In addition, subsidization of sanitation infrastructure has been blamed for distorting the private supply market and artificially inflating hardware prices. And finally, in practice, elite capture of construction subsidies has meant that the transfer is not always received by the intended beneficiary.

Conclusion

Further research on the health externalities of improved sanitation and water quality would provide policymakers with the evidence they need to decide on whether and how much to subsidize these interventions. How subsidies might be best provided is another area for exploration, so that the pitfalls identified above can be minimized or avoided. An investigation of households' valuation for improved sanitation, like the research that has been done on water quality and bed nets, would better inform policymakers in this debate.

Understanding how water and sanitation service bundle public and private benefits is also important in determining how to provide these goods. The majority of nonhealth benefits of sanitation, including reduced shame and stress, are likely what economists call 'private benefits.' Water quantity provides many private benefits, in terms of quality of life, whereas water quality may have a more public health character. Sanitation likely combines both attributes. Individuals are typically considered to be best placed to allocate limited resources between competing goals, to the extent that they are the sole beneficiaries of those goods, and thus it is often argued that donor, or public funds should be reserved for funding those investments that provide public benefits, not private ones, no matter how attractive the attributes of the private goods may be.

However, the case for investment in private goods can be stronger if a donor values the welfare of women and girls more than decision-making processes within households might. Despite the fact that women benefit disproportionately from sanitation and water services, if they have little ability to control how resources are spent, unmet demand may persist even if nominally affordable or seemingly attractive options are available. A donor may wish to support providing services that women want as a way of redistributing resources toward women. Untangling these issues remains a fruitful area of research.

Disclaimer

Views presented in this article should not be construed to be those of the Bill & Melinda Gates Foundation or its leadership.

See also: Infectious Disease Externalities. Peer Effects in Health Behaviors. Pricing and User Fees. Public Health in Resource Poor Settings. Survey Sampling and Weighting. Value of Information Methods to Prioritize Research

Further Reading

Agarwal, B. (2000). Conceptualizing environmental collective action: Why gender matters. *Cambridge Journal of Economics* 283–310.

Ahuja, A., Kremer, M. and Zwane, A. P. (2010). Providing safe water: Evidence from randomized evaluations. *Annual Review of Resource Economics* **2**.

Ashraf, N., Berry, J. and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review* **100**(5), 2383–2413.

Baird, S., Hicks, J. H., Kremer, M. and Miguel, E. (2011) Worms at work: Long run impacts of child health gains. *Working paper.*

Cairncross, S. (2004). *The case for marketing sanitation*. Nairobi, Kenya: Water and Sanitation Program, World Bank.

Clasen, T. F., Roberts, I. G., Rabie, T., Schmidt, W.-P. and Cairncross, S. (2006). Interventions to improve water quality for preventing diarrhoea. *Cochrane Database of Systematic Reviews* (3), Artic. No. CD004794.

Curtis, V. and Cairncross, S. (2003). Effect of washing hands with soap on diarrhea risk: A systematic review. *Lancet Infectious Diseases* **35**, 275–281.

Cutler, D. and Miller, G. (2005). The role of public health improvements in health advances: The 20th century United States. *Demography* **421**, 1–22.

Devoto, F., Duflo, E., Dupas, P., Parienté, W. and Pons, V. (2009). Happiness on tap: The demand for and impact of piped water in urban Morocco. *Working Paper.* Cambridge, MA: Massachusetts Institute of Technology.

Galiani, S., Gertler, P. and Schargrodsky, E. (2005). Water for life: The impact of privatization of water services on child mortality. *Journal of Political Economy* **1131**, 83–119.

Gamper-Rabindran, S., Khan, S. and Timmens, C. (2010). The impact of piped water provision on infant mortality in Brazil: A quantile panel data approach. *Journal of Development Economics* **92**, 188–200.

Genser, B., Strina, A., Santos, L. A., et al. (2008). Impact of a city-wide sanitation intervention in a large urban centre on social, environmental and behavioural determinants of childhood diarrhoea: Analysis of two cohort studies. *International Journal of Epidemiology* 831–840.

Guerrant, R. L., Oria, R. B., Oria, M. O. B. and Moore, S. R. (2008). Lima AAM. Malnutrition as an enteric infectious disease with long-term effects on child development. *Nutrition Reviews* **669**, 487–505.

Holla, A. and Kremer, M. (2008). Pricing and access: Lessons from randomized evaluations in education and health. *Center for Global Development Working Paper No. 158.*

Humphrey, J. H. (2009). Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet* **374**(9694), 1032–1035.

Kremer, M., Leino, J., Miguel, E. and Zwane, A. P. (2011). Spring cleaning: Rural water impacts,valuation and property rights institutions. *Quarterly Journal of Economics* **126**(1), 145–205.

Kremer, M., Miguel, E., Meeks, R., Null, C. and Zwane, A. (2009b). Willingness to pay for cleaner water in less developed countries: Rigorous evidence and directions for future research. *Working Paper, International Initiative for Impact Evaluation 3IE.*

Kremer, M., Miguel, E., Mullainathan, S., Null, C. and Zwane, A. (2009). Making water safe: Price, persuasion, peers, promoters, or product design.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **721**, 159–217.

O'Reilly, K. (2010). Combining sanitation and women's participation in water supply: An example from Rajasthan. *Development in Practice.*

Schmidt, W. P. and Cairncross, S. (2009). Household water treatment in poor populations: Is there enough evidence for scaling up now? *Environmental Science and Technology* **434**, 5542–5544.

Waddington, H. and Snilstveit, B. (2009). Effectiveness and sustainability of water, sanitation, and hygiene interventions in combating diarrhoea. *Journal of Development Economics* **13**, 295–335.

Watson, T. (2006). Public health investments and the infant mortality gap: Evidence from federal sanitation interventions and hospitals on U.S. Indian reservations. *Journal of Public Economics* **908–909**, 1537–1560.

Wright, J., Gundry, S. and Conroy, R. (2004). Household drinking water in developing countries: A systematic review of microbiological contamination between source and point-of-use. *Tropical Medicine and International Health* **91**, 106–117.

Zwane, A. and Kremer, M. (2007). What works in fighting diarrheal diseases in developing countries? A critical review. *World Bank Research Observer*

Zwarteveen, M. and Meizen-Dick, R. (2001). Gender and property-rights in the commons: Examples of water rights in South Asia. *Agriculture and Human Values* 11–25.

Welfarism and Extra-Welfarism

J Hurley, McMaster University, Hamilton, ON, Canada

Glossary

Capabilities The set of all possible physical and social functionings for a person.
Consequentialism The normative tenet that any policy or action be judged solely in terms of the resulting, or consequent, effects.
Extra-welfarism A normative framework of economics which holds the evaluation of a policy or resource allocation that should be based on a larger set of information than the utilities attained by members of society solely.
Functioning A person's ability to engage in private and social activities, frequently also taken to include emotional and psychological states of health.
Individual sovereignty A tenet of welfare economics that holds individuals are the best judges of their own welfare.
Pareto criterion An efficiency criterion that holds a resource allocation is efficient if it is not possible to reallocate resources so as to increase one person's utility without decreasing another person's utility.
Potential Pareto criterion An efficiency criterion in economics that holds an allocation is preferred to another if the gains to the winners are sufficiently large to compensate the losers and still leave the winners better off.
Social welfare function A function that maps from the levels of utility attained by members of society to the overall level of welfare for society.
Utility It is variously defined in the history of economics. Two dominant interpretations are hedonistic utility, which equates utility with pleasure, desire-fulfilment, or satisfaction; and preference-based utility, which defines utility as a real-valued function that represents a person's preference ordering.
Welfarism A tenet of welfare economics that holds the evaluation of a policy or resource allocation should be based solely on the utilities attained by members of society.

Introduction

Welfarism and extra-welfarism are alternative normative economic frameworks for ranking resource allocations. By normative, we mean economic analysis intended to answer questions such as, What 'ought' we to do? or Which resource allocation is the best? or Is policy A preferred to policy B? Normative analysis unavoidably rests on value judgments regarding, for example, as to what constitutes a benefit. Normative economic analysis contrasts with positive economic analysis in that, by answering questions such as, What will be the effect of policy A on the allocation of resources among the members of the society?, it attempts to describe what will happen without making any judgment as to the goodness or desirability of the predicted effects.

Welfarism is one element of the welfare-economic framework that dominates normative analysis in economics. It dictates that the only information relevant for ranking alternative allocations of resources is the utilities attained by the individuals in a society.

Extra-welfarism, in contrast, argues that normative economic analysis should be based on a larger set of information than simply the utilities attained by individuals in the society. Different variants of extra-welfarism emphasize different types of information to either supplant or supplement utility information.

As noted, normative economic analysis is conducted to rank-order policy options. Each policy generates a particular allocation of goods and services within the society and an associated distribution of well-being among the members of the society. The policies considered can be quite narrowly defined and generate relatively circumscribed effects, such as would be the case when comparing alternative dosages of a drug used to treat a relatively rare, minor ailment, or they can be broadly defined with wide-ranging, profound effects on resource allocation, such as would be the case when comparing alternative systems of health-care finance, which may affect labor markets, income distribution, health-care consumption, and other economic activities. In either case, the term 'resource allocation' refers simply to who gets what in society, including health-care goods and services, other government goods and services, private consumption goods and services, and so forth. Hereafter, we use the terms 'policy' and the associated resource allocation created by a policy interchangeably.

Welfarism and the Welfare-Economic Framework

The mainstream welfare-economic framework is built on four central tenets: utility-maximization, individual sovereignty, consequentialism, and welfarism.

Utility-maximization refers to the behavioral assumption that individuals choose rationally: given a set of choice options, an individual chooses the most preferred option among them according to the defined notions of consistency. Utility has been interpreted variously in the history of economics and continues to have multiple interpretations. The two dominant interpretations are hedonistic utility and preference-based utility. Hedonistic utility, which derives from classical utilitarianism, equates utility with the pleasure, happiness, or satisfaction that an individual derives from a good or an activity. It is a psychological construct and an individual is assumed to choose and act so as to maximize his utility.

In contrast, the preference-based definition of utility eschews a psychological interpretation and defines utility as a function that represents a preference ordering: those goods or activities that an individual chooses are assigned a higher utility value than those that are not chosen. Utility therefore simply represents preferences: it makes no assumptions as to the reasons why one thing is preferred to another. Modern microeconomic theory adopts this preference-based interpretation of utility. At times, some have argued that utility is defined only over goods and services, but in modern economics, utility can be defined over goods, services, activities, and virtually any phenomenon on which an individual can express a preference.

Individual sovereignty (sometimes referred to as 'consumer sovereignty') is the maxim that individuals are the best judges of their own welfare. Any judgment regarding an individual's welfare should be based on the person's own assessment. Individual sovereignty explicitly rejects paternalism – the notion that a third party may know better than the individual what is best for himself or herself. It dictates that individual preferences be respected in the process of evaluation.

Consequentialism holds that any policy must be judged exclusively in terms of the resulting, or consequent, effects. The motivation for or intention behind the policy does not matter; ethical imperatives such as duties, rights, and obligations do not matter. All that matters is the effects that flow from the policy.

As already noted, welfarism holds that the goodness of any resource allocation is to be judged solely on the basis of the utilities attained by the affected individuals. No other aspect of the situation matters. Together, these four tenets require that any policy be judged solely in terms of the resulting utilities achieved by individuals, as assessed by the individuals themselves.

The final ranking also depends on the criterion used to rank alternative policies based on this utility information. Early Neoclassical Welfare Economic Theory assumed (like classical utilitarianism) that utility was cardinally measurable and interpersonally comparable, and defined the best allocation as the one that maximized total utility (i.e., the sum of individual utilities) for the population. Modern welfare economics assumes that utility is ordinally measurable and noninterpersonally comparable, and it replaces the utility-maximizing criterion with the Pareto criteria: an allocation of resources is judged to be Pareto optimal if it is not possible to reallocate resources so as to increase one person's utility without decreasing another person's utility. Although many economists prefer the Pareto criterion to the sum-maximizing criterion because it makes less restrictive assumptions regarding utility and it makes weaker ethical assumptions, the Pareto criterion suffers some important limitations for applied welfare analysis. For a given set of resources, each of the many possible allocations can be Pareto optimal: the Pareto criterion does not necessarily lead to full ranking that identifies a single allocation as best. Besides, as nearly all real-world policy changes hurt at least one individual, organization or group in a society, strict application of the Pareto criterion leads to policy paralysis in which no policy can be judged better than the *status quo*.

These limitations of the Pareto criterion led to the development of the potential Pareto criterion (also call the hypothetical compensation test), which states that one allocation is preferred to another if the gains to the winners are sufficiently large to enable them to (hypothetically) compensate the losers while are still leaving the winners better off. Crucially, however, compensation does not have to be paid to the losers, so the losers are in fact still worse off. The potential Pareto criterion is the basis for much normative economic policy analysis.

Neither the sum-maximizing criteria of classical utilitarianism nor the Pareto criterion are sensitive to how utility is distributed among the members of a society. Welfare economics has tried to incorporate distributional concerns through the concept of a social welfare function. A social welfare function maintains the welfarist assumption that rankings depend solely on the utilities attained by the members of the society, but includes a measure of the society's aversion to inequality. Hence, even if the total utility is the same under both the policies, and if the society is averse to inequality the policy under which utility is distributed more equally among individuals would be preferred.

Welfare-economic analysis has been importantly shaped by the two fundamental theorems of welfare economics. The first theorem states that a well-functioning market (where this has a specific meaning) leads to a Pareto optimal allocation of resources. The second theorem states that, given the right initial distribution of income in society, any Pareto optimal allocation (including those judged to be equitable) can be achieved through market allocation. When combined with the claim that questions of distribution are fundamentally political rather than economic, the second theorem has led economists to focus on issues of efficiency, effectively separating the analysis of efficiency from the analysis of distribution. In addition, given that any allocation can, in principle, be achieved through a system of well-functioning markets, market-based allocation serves as the reference standard for judging efficiency. Hence, within the welfare-economic framework government intervention in a market can be justified by equity concerns (e.g., income inequality) or by market failure, a situation in which deficiencies in the market cause market allocation to be inefficient. The latter, however, has dominated economic analyses on the role of government policy. As a corollary, within this framework the objective of any corrective public policy is to achieve the allocation that would have resulted from a well-functioning market.

Empirical Welfare Analysis

Using this welfare-economic framework as a guide to empirical normative analysis presents a number of challenges. The methodology of applied welfare analysis is called cost–benefit analysis. The goal of cost–benefit analysis is to determine whether the adoption of a policy will be more efficient than a specified alternative policy, where efficiency is defined by the potential Pareto criterion. Within cost–benefit analysis, utility is measured using a money metric. Benefit to a member of a society is defined as the amount of money a person is willing to pay for the effect achieved by a policy (e.g., improved health). As both benefits and costs are measured in monetary units, a policy is deemed efficient (relative to the alternative

against which it is being compared) if the net benefit (incremental benefits–incremental costs) is positive. The goal of cost–benefit analysis is to mimic for government policies the allocation of resources that would have resulted from a well-functioning market.

Extra-Welfarism

At its most general sense, extra-welfarists argue that normative assessment should be based on a wider set of information than solely the utilities attained by individuals. Although utility information may be relevant, it is insufficient and additional, or extra, information should be incorporated into the analysis. Extra-welfarism is not strongly prescriptive as to what this additional information is; indeed, many extra-welfarists argue that the relevant information depends on the context. As extra-welfarism is, in part, a reaction against the dominant welfarist paradigm in economics, it tends to be more eclectic as different extra-welfarist writers have reacted against different elements of welfarism and proposed different ways of addressing welfarism's limitations. As a result, it is perhaps easier to articulate what extra-welfarism is not than what it is. Still, although extra-welfarism continues to develop within heath economics, it embodies a sufficient set of foundational principles and ideas as to define a coherent normative framework distinct from welfarism (Hurley, 2000; Brouwer *et al.*, 2008).

Extra-welfarist ideas have existed within health economics since long before the term 'extra-welfarism' was coined. Much of the work of extra-welfarists in health economics has been to integrate these ideas into a framework with deeper conceptual foundations. The work of Culyer (1990) – who coined the term 'extra-welfarism' – has been particularly important in this respect. These extra-welfarist tendencies in health economics arose for a number of reasons:

- Health economists have traditionally worked closely with noneconomist health professionals, including clinicians, epidemiologists, bioethicists, decision scientists, and others not schooled in welfare economics and who found some of its elements unpalatable. In particular, a primary outcome of health-care interventions is improved health, and notably, 'lives saved.' Many noneconomists objected to assigning a monetary value to lives saved. This gave rise to the use of cost-effectiveness analysis rather than cost–benefit analysis in the evaluation of health-care interventions. Cost-effectiveness measures outcomes in natural units (e.g., life-years gained and cases detected) and refrains from assigning a monetary value (or indeed any explicit social value) to the health gains produced by a health intervention.
- Government health policy commonly rejects the welfare-economic view that the desired allocation of health-care resources is the one that would follow from market forces based on people's willingness to pay. The explicitly stated objective of many governments is to improve population health based on allocation of health-care resources according to 'need', regardless of a person's ability to pay or willingness to pay. This requires that access to health care be independent of a person's income or wealth. Even if the society could somehow get the distribution of income 'correct' (as envisioned in most of the welfare-economic analysis), it would be very difficult to achieve this objective through market-based allocation of health care using prices. Instead, the objective requires allocation based on nonmarket principles.
- Health economics has been heavily influenced by the 'decision-maker' approach to cost–benefit analysis, which rejects the individualistic, welfarist foundations of traditional cost–benefit analysis. The decision-maker approach instead emphasizes the objectives (and weights) of high-level policy decision makers. This approach has strong affinities with the view within decision sciences that the role of the analyst is to assist a decision-maker in achieving the decision-makers stated goals. Combined with the point above regarding the stated objectives of most of the health policies, this perspective led health economists to develop methods of evaluation in which health – rather than utility – is the primary measure of benefit.
- The demand for health care derives from the demand for health. This derived demand for health care means that health care is an input into the production of health, which permits normative assessments of health-care consumption using supply-side notions of efficiency, unlike other sectors of the economy in which individuals demand goods for their direct effects on utility. This underlying production relationship enables a third-party analyst to use evidence regarding the impact of a health-care service on health (generated, for instance, by clinical studies) to assess the efficiency of health-care consumption: it cannot be efficient to consume a service known to be ineffective in producing health. This perspective on normative assessment in the health sector reinforces a focus on health as the main outcome of interest.

Extra-welfarism in health has been shaped importantly by the work of the Nobel-Prize winning economist Amartya Sen. Sen broke sharply with welfarism and has developed an alternative framework based on the concepts of human functionings and capabilities (e.g., Sen, 1999). Sen argued that welfarism was an insufficient basis for normative economic analysis because utility focuses too narrowly on people's mental and emotional reactions to circumstances and not enough on what they can achieve with their material and other resources. Welfarism, for instance, suffers from the problem of adaptation. A person born into poverty, who adjusts his/her life expectations to conform to his/her limited life possibilities may, as measured by utility, be better off than a well-off person given every advantage and opportunity in life but whose expectations exceeded what his/her was able to realize and who therefore ends up disappointed. Welfarism also either ignores nonutility aspects of a situation, such as whether basic human rights are being violated, or, to the extent that they are captured by the analysis, they enter only through the metric of utility. In the end, Sen argues that neither utility – whether defined in traditional hedonistic terms as desire-fulfillment or pleasure or in modern terms as preference satisfaction – nor a person's material possessions can serve as a proper basis for assessing alternative social policies.

Sen argued that, instead, evaluation should focus on functionings and capabilities. The central ideas of his approach can be briefly summarized as follows. He begins with commodities – goods and services – that can be under a person's command and which have characteristics: a bicycle is a commodity whose salient characteristic is that it can provide transportation; health care is a commodity whose salient characteristic is the potential to improve a person's health. Commodities and their characteristics do not depend on features of the person who possess them. A functioning is what a person succeeds in doing with the commodities at his or her command; it is what the person manages to do or to be. Bicycling is a functioning. The functionings possible for a person depend on the commodities at the command of a person and features of the person and their environment. A quadrapalegic with a standard bicycle cannot achieve the functioning of bicycling. Functionings range from the trivial (enjoy ice cream) to the profound (form meaningful relationships with others). The set of all possible functionings for a person, which Sen calls their capabilities, depends on their feasible set of commodities available and, given their personal characteristics and social institutions, the feasible set of means available to them to transform commodities into functionings. Alternative states of the world should be evaluated based on the people's capabilities in that world; that is, the functionings open to them. Sen emphasizes a nonconsequentialist perspective that the evaluation be based not on the functionings a person actually achieves, as these may depend on personal choices not of concern to normative assessment, but rather on what was possible for the person – the life prospects that were available to them given the social arrangements and their place in society. Sen's framework is commonly referred to as the 'capabilities' approach within extra-welfarism.

As noted, Culyer was one of the earliest health economists to recognize the potential for adapting Sen's ideas to the health sector. The concept of need, for example, is widely used by health professionals, health scientists, and, historically by health economists, though it does not fit easily into the individualistic, welfarist framework of traditional welfare economics in which people are characterized only by utility; and mainstream economics has in fact been hostile to the concept, seeing it largely as an attempt by some to gain privileged status for their preferences. Yet, the concept of need for health care fits easily into the capabilities approach wherein health care is a commodity that is essential for (and needed by) an ill-person to become healthy and thereby be able to realize many different types of functionings.

Within a broad welfarist framework, functionings and capabilities could, in principle, be valued using a metric of utility. Indeed, among welfarists who believe that preferences extend over any aspect of life, there is no concern of extra-welfarism that could not, in principle, be captured through preferences. The point of extra-welfarism, however, is that preferences are the wrong metric by which to value such outcomes. The extra-welfarist criticism of utility is not that it can only be applied to a limited range of outcomes; rather the criticism is that it provides a limited valuation of all those outcomes to which it is applied.

Though Sen's capabilities approach is a particularly influential source of extra-welfarist thought in health economics, it is only one variant of extra-welfarism. Brouwer *et al.* (2008) represents perhaps the most comprehensive attempt to articulate the scope and definition of extra-welfarism in the health sector, and in particular ways in which extra-welfarism differs from the welfarist normative framework. They emphasize four ways in which extra-welfarism differs from the welfarist framework: (1) extra-welfarism permits the use of outcomes other than utility; (2) extra-welfarism permits the use of sources of valuation other than the affected individuals; (3) extra-welfarism permits the weighting of outcomes (whether utility or other) according to the principles that need not be preference-based; and (4) extra-welfarism permits interpersonal comparisons of well-being in a variety of dimensions, thus enabling movement beyond Paretian economics. Below we summarize some of the main elements of extra-welfarism in the health sector, discussing these and other points related to extra-welfarism.

Extra-welfarism emphasizes the use of nonutility information in normative economic analysis. Conceptually, extra-welfarism does not reject a possible role for utility information – it simply argues that utility information alone is insufficient. The predominant nonutility outcome of interest within health economics is health status. The focus on health derives from a number of factors – the influence of the decision-maker approach and the fact that health policy-makers emphasize health as the outcome of interest; the fact that health is observable and, within certain bounds, interpersonally comparable; and the fact that health can be integrated into Sen's capability approach. Extra-welfarism can also accommodate other types of nonutility information such as rights, dignity, and other ethical concerns. Indeed, it places no restriction on the type of nonutility information that can, in principle, be included in an analysis. Some have argued, however, that a shortcoming of how extra-welfarism in the health sector has developed in practice is a near-exclusive focus on health as the outcome of interest in evaluation. To the extent that this is true, extra-welfarism has simply substituted one restrictive definition of the outcome space – utility – with another – health.

Health status can be measured in numerous ways depending on the specific question under consideration, and the emphasis on health as the outcome of concern has spurred economists to develop health measures with desirable properties. The dominant health-related outcome measure for the evaluation of health programs, services, and technologies is the quality-adjusted life-year (QALY). The QALY illustrates well that extra-welfarists do not completely reject the use of utility information and utility-based methods, but that they often use utility information in ways not consistent with welfare economics. The QALY is designed to capture the effect of a health intervention on both the quantity and quality of life-years. It is defined simply as the quality-weighted sum of life-years: $\sum_i Q_i Y_i$, where Y_i is the number of years spent in health state i and Q_i is a quality weight for health state i that takes on a value between 0 and 1, where the weight for full health is 1.0 and the weight for being dead is 0. The quality weights used to construct a QALY measure are often constructed using methods derived from utility theory; that is, the quality weights underlying the QALY are elicited using the methods drawn from utility theory.

This fact has led to considerable confusion and debate regarding the interpretation of a QALY, and this debate reflects

important issues in the debate on welfarism and extra-welfarism. The confusion stems from the fact that a QALY can be interpreted differently depending on the auxiliary assumptions one is willing to make. Under certain assumptions regarding the nature of people's utility functions, a QALY can be interpreted as a measure of individual utility or preferences. These assumptions are restrictive, however, and do not accord well with evidence on the actual structure of people's utility functions. Hence, welfarists – for whom utility is the outcome of interest – have criticized the QALY because it does not represent well people's preferences over health states. Extra-welfarists, however, have argued this criticism is misguided: although preference information is used in the construction of a QALY, the QALY is not intended as a measure of preferences but instead it is a measure of health, or perhaps as more commonly stated, health-related quality of life. Hence, the fact that QALYs do not accurately map preferences over health states is not a weakness. Finally, some extra-welfarists have explored the potential for interpreting the QALY as a measure of the value of an individual's capability set. Doing so requires making the rather strong assumption that the QALY can represent all of an individual's well-being rather than making the more traditional assumption that QALY represent only one component of an individual's well-being (the health-related component). Regardless of the specific interpretation one may adopt, the point is that the QALY illustrates one way extra-welfarists integrate utility information in unconventional ways, and, in addition, that one cannot understand certain debates in health economics without being attuned to subtle distinctions between welfarist and extra-welfarist approaches.

Unlike welfare economics, extra-welfarism permits sources of valuation other than the individuals affected by a policy. Extra-welfarism is comfortable with certain types of paternalism. Acting as agents for the members of the society, decision-makers are not assumed to act in the way they think individuals would act, but rather as they think individuals ought to act. The potential sources of value are many. The valuation could be based on empirical information drawn from the general public, but it may include information other than preferences, it may even include preferences but aggregated in ways inconsistent with individualistic welfare-economic analysis, it may gather information in ways inconsistent with preference measurement such as deliberative process and consultations of various types, or it could be responses to questions quite different from preference information. Valuation could be reasoned argument that has no basis in empirical measurement.

Extra-welfarism and welfarism differ importantly in a number of ways pertaining to the aggregation of costs and benefits across individuals. Extra-welfarism is more accommodating of interpersonal comparisons of well-being. Sen has argued in particular for frameworks that admit more possibilities than the two extremes of complete rejection of interpersonal comparability (as in modern welfare economics) and the assumption of full interpersonal comparability (as in classical utilitarianism). He emphasizes the scope for limited degrees of comparability that allow for at least partial orderings of the different possible states of the world. Such partial orderings are all that the analysis requires in many situations. The shift to nonutility outcomes such as health also provides greater scope for assuming degrees of interpersonal comparability. Again, even if one recognizes that it is not possible to conclude which of the two people have greater health in all situations, it often is possible to make relatively uncontested judgments regarding relative levels of health across individuals. The same is true for pivotal concepts such as the need for health care.

Extra-welfarism provides greater scope than welfarism for integrating equity concern. Both the Pareto criterion of analytical welfare economics and the net-benefit criterion on cost–benefit analysis are completely insensitive to distributional concerns. Extra-welfarism, however, allows for the integration of nonunitary equity weights that reflect the fact that society may value health gains among certain groups in society more than others. The equity weights constitute a type of nonutility information that may be derived from a number of sources. Some examples include differential weighting by age (greater weight to the young), by one's role in society (e.g., greater weight to caregivers, especially those caring for children), by the extent to which one is responsible for one's ill-health (lesser weight to those who are at least in part responsible for their ill-health), or by baseline health status (greater weight to those with more severe health conditions). The use of such weights also implies breaking with the welfare-economic assumption of anonymity, which holds that no characteristic of an individual matters for evaluation except their value with respect to the outcome of interest.

Extra-welfarism is also more accommodating of equity concerns beyond distributional equity, such as procedural equity. Procedural equity emphasizes fairness in the process by which resources are allocated, respect for the rights of individuals, and related matters that may not bear on the actual final distribution of resources. Although welfarist approaches have developed the idea of 'process utility' to capture some of these types of concerns, such an approach values them only to the extent that they affect utility, whereas many argue that aspects of procedural equity lie outside the logic of a consequential calculus.

Empirical evidence consistently demonstrates that among the general public concerns regarding equity weigh heavily in judgments on the allocation of health-care resources. People are willing to reduce the total amount of health produced in order to achieve a more equal distribution of health among the members of the society. Moreover, although extra-welfarism can readily accommodate such attitudes, some analysts have been critical of the fact that, as a pragmatic matter, a good deal of extra-welfarist evaluation in health has adopted a health-maximization criterion: that is, the best option is that which maximizes the amount of health produced, in a manner analogous to the classical utilitarian approach within welfarist economics. As with the concern regarding a near-exclusive focus on health in the outcomes space, the issue is not what extra-welfarism embodies in principle, but rather what happens on the ground in actual practice.

Extra-Welfarism in Empirical Normative Analysis

There is no single 'extra-welfarist' approach to empirical normative analysis. In the area of health technology assessment, extra-welfarist ideas are incorporated into cost-effectiveness and cost–utility analyses. In the normative analysis of broader

health system policies extra-welfarist ideas manifest themselves in a focus on health as the outcome of interest rather than willingness to pay as measured by the demand curve for health-care services. Two brief examples below illustrate how selected aspects of extra-welfarism translate into empirical normative analysis, and especially how these aspects differ from the welfare-economic approach. The first illustrates, in the context of health technology assessment, how extra-welfarism draws on sources of valuation distinct from the individualistic approach of welfarism. The second illustrates how a focus on health rather than willingness to pay can change the normative conclusions regarding the optimal policy to combat insurance-induced moral hazard in the health services market.

Sources of valuation in extra-welfarism

Consider the economic evaluation of a new health-care technology the only effect of which is to reduce the probability of dying within a given period. Under the welfare-economic framework, these additional life-years would be valued according to the amount each affected individual was willing to pay to reduce the chance of death, so some (e.g., the wealthy) would be willing to pay more than others (e.g., the poor). The new technology would be deemed efficient if the net benefit was positive. If, instead, a cost-effectiveness analysis (CEA) is conducted, the analyst refrains from placing a value on the additional life-years produced by the new technology. The efficiency of the intervention is expressed using the incremental cost-effectiveness ratio (ICER), which indicates the additional cost incurred per life-year gained. Although CEA allows the analyst to avoid placing a monetary value on a life-year gained, if the results of the CEA are to be used as the basis for an adoption decision, the decision-maker must decide whether the additional cost per life-year is above or below the amount the *society* is willing to pay-that is, are the extra benefits worth the extra costs. The extra-welfarist CEA bases the decision on the social value of the health benefit (e.g., life-year gained), which may be set by the decision-maker, by community consultation, or by some other process. Although, as noted above in the discussion of equity weights, the social value of the health gain may differ across the groups of individuals in a society, the social value does not depend on the individuals' own valuation of the health gain.

The evaluation of policies to combat moral hazard

The analytic importance of using health rather than willingness to pay as a measure of social value can be illustrated with an example adapted from Reinhardt (1998). Consider two families with identical preferences and full insurance for physician visits, the Chens, who are wealthy, and the Smiths, who are poor. Each family has just had a baby. Baby Chen is healthy but Baby Smith is sickly. With full insurance, the Smiths demand nine physician visits during the period and the Chens demand six. To combat moral hazard, the insurer seeks to reduce total visits by five. To do so, it imposes a copayment equal to $15 per visit. Under cost-sharing, both families demand five visits per period, for a reduction of five visits overall (from 15 to 10). Within the standard welfare-economic analysis, the reductions of four and one, respectively, for the Smiths and the Chens are the optimal way to reduce total visits by five because it imposes the lowest welfare loss as measured by willingness to pay. However, because Baby Smith is sickly while Baby Chen is healthy, the marginal health gain of a physician visit is always greater for the Smiths than for the Chens. If we measure benefit by the health effects rather than willingness to pay, then to reduce overall visits by five in a way that minimizes health effects, Baby Chen's visits should be reduced from six to two whereas the Baby Smith's are reduced from nine to eight. Such a reduction cannot be achieved by a single user-charge policy. Consequently, this extra-welfarist analysis indicates that moral hazard should be combated by an alternative policy that can selectively reduce the visits that have the least health gain.

Current Issues in Extra-Welfarism

Extra-welfarism continues to develop in health economics, and we highlight just a few areas of the ongoing development. One active area of work seeks to translate Sen's capability approach into a practical, empirical method. Capability sets, as noted above, include not only the functionings a person actually achieves, but all the possible functionings a person could have achieved. Hence, a major challenge is how to measure these potential functionings that were not achieved by an individual, and in particular to distinguish those that were available but not chosen from those that were not available due to some social or economic barrier in society. A second area focuses on the relationship between extra-welfarism and welfarism. Brouwer *et al.* (2008) argue that welfarism is simply a special case of extra-welfarism. Much of the literature has tended to focus on the two polar cases – welfarism which limits the outcome space to utility – and extra-welfarist approaches that ignores utility in the outcome space (even if it has used elements of utility in constructing indices such as the QALY). Less well developed is systematic thinking regarding the rich middle ground that admits both utility and nonutility outcomes, and in particular, principles that can guide the role of each in an analysis. The nature of the outcome information included, for instance, might depend on the level of aggregation at which a decision is being made. At high levels of aggregation, such as a central ministry allocating resources to regional authorities responsible for their respective populations, the concepts of need and health likely dominate considerations of possible differences in preferences. However, in the design of programs at the local level, the role for preferences may be larger as the program has more direct dealings with those receiving services. Similarly, it is natural to ascribe a larger role for preferences (especially over process aspects of care delivery) in those contexts in which differences in health and other outcomes are small. Finally, there is ongoing tension between the emphasis in extra-welfarism on diversity, flexibility, and adaptability of methods and standardization of methods both to increase comparability across studies and to improve quality in those aspects for which there may be generally recognized 'better' methods (e.g., of certain aspects of measurement). This tension is an inherent feature of extra-welfarism. Only by attending carefully to its demands can extra-welfarism avoid the pitfalls of, on one hand, lacking sufficient shared, core

principles required for a coherent, distinct vision and, on the other, a type of standardized rigidity against which extra-welfarism is in part a reaction.

Acknowledgment

The author would like to thank Junying Zhao for many stimulating discussions regarding extra-welfarism, and especially on the ideas of A. Sen.

See also: Cost-Effectiveness Modeling Using Health State Utility Values. Cost–Value Analysis. Disability-Adjusted Life Years. Efficiency and Equity in Health: Philosophical Considerations. Efficiency in Health Care, Concepts of. Equality of Opportunity in Health. Ethics and Social Value Judgments in Public Health. Health and Health Care, Need for. Health and Its Value: Overview. Incorporating Health Inequality Impacts into Cost-Effectiveness Analysis. Measuring Equality and Equity in Health and Health Care. Measuring Health Inequalities Using the Concentration Index Approach. Measuring Vertical Inequity in the Delivery of Healthcare. Multiattribute Utility Instruments: Condition-Specific Versions. Problem Structuring for Health Economic Model Development. Quality-Adjusted Life-Years. Time Preference and Discounting. Unfair Health Inequality. Utilities for Health States: Whom to Ask. Willingness to Pay for Health

References

Brouwer, B. F., Culyer, A. J., van Exel, N. J. and Rutten, F. (2008). Welfarism vs. extra-welfarism. *Journal of Health Economics* **27**(2), 325–338.

Culyer, A. J. (1990). Commodities, characteristics of commodities, characteristics of people, utilities, and the quality of life. In Baldwin, S., Godfrey, C. and Propper, C. (eds.) *Quality of life: Perspectives and policies*, pp. 9–27. London: Routledge.

Hurley, J. (2000). An overview of the normative economics of the health sector. In Culyer, A. J. and Newhouse, J. P. (eds.) *Handbook of health economics*, pp. 55–118. Amsterdam: Elsevier Science B.V.

Reinhardt, U. (1998). Abstracting from distributional effects, this policy is efficient. In Barer, M., Getzen, T. and Stoddart, G. (eds.) *Health, health care and health economics: Perspectives on distribution*, pp. 1–53. Toronto: John Wiley and Sons.

Sen, A. (1999). *Commodities and capabilities*. Oxford: Oxford University Press.

Further Reading

Anand, P. and Hees, M. (2006). Capabilities and achievements: An empirical study. *Journal of Socio-Economics* **35**, 268–284.

Boadway, R. and Bruce, N. (1984). *Welfare economics*. Oxford: Basil Blackwell.

Coast, J., Smith, R. and Lorgelly, P. (2008). Welfarism, extra-welfarism and capability: The spread of ideas in health economics. *Social Science and Medicine* **67**, 1190–1198.

Cookson, R. (2005). QALYs and the capability approach. *Health Economics* **14**, 817–829.

Culyer, A. J. (1989). The normative economics of health care finance and provision. *Oxford Review of Economic Policy* **5**(1), 34–58.

Hurley, J. (1998). Welfarism, extra-welfarism and evaluative economic analysis in the health sector. In Barer, M., Getzen, T. and Stoddart, G. (eds.) *Health, health care and health economics: Perspectives on distribution*, pp. 373–396. Toronto: John Wiley and Sons.

Sen, A. (1979). Personal utilities and public judgments: Or what's wrong with welfare economics. *The Economic Journal* **89**(355), 537–558.

What Is the Impact of Health on Economic Growth – and of Growth on Health?

M Lewis, Georgetown University, Washington, DC, USA

Introduction

Health status improvements over the past 400 years have been steady, with a surge over the last century that is nothing short of spectacular, and a period during which economic progress has soared. Vaccines, antibiotics, and other medical advances have contributed to reductions in illness (morbidity) and mortality. Economic growth has shown equally impressive gains catapulting much of the world into higher income status with all of the associated consumption and health benefits that higher incomes allow. Richer people are better educated and nourished, and can afford investments that improve population health (3.9 Cutler and Lleras-Muney).

With economic advances and health status moving in tandem, the question is: Do health improvements spur growth, or do income increases determine progress in population health, or both? The evidence is mixed and hampered both by the lack of an acceptable and universal measure of health status and by the empirical difficulties inherent in controlling for reverse causality as well as nonlinearities.

Measurement of both growth and health pose challenges for both empirical estimation and for policy implications. For growth, the measurement is either of a static level of income, itself difficult to measure, or the dynamic of economic growth. Both may drive improvements in health, or serve as proxies for other achievements of higher-income countries such as education, effective public health interventions, or strong institutions, all of which correlate with improved health status.

The existing health metrics are inadequate to the task of effectively measuring the links between growth and health. No single measure exists to capture mortality and morbidity. Mortality only occurs once and is therefore a rare event. Common reliance on infant mortality reflects the high concentrations of deaths in the first year of life, and the availability of accessible and comparable data across countries. As infant mortality effectively captures differences across population health it is the measure of choice (1.1 Murray). Life expectancy serves a similar purpose but fails to capture health status as effectively, relegating its use to comparing broad national trends. More recent measures include life satisfaction and emotional well-being, malnutrition, and stunting (a measure of long-term malnutrition).

Accumulated evidence on the relationship between health and growth indicates that both directions of causality are plausible but neither is definitive. Existing evidence from cross-country macroeconomic studies analyze country-level relationships. More heterogeneous, microeconomic investigations provide analytic underpinnings for interpreting the findings of macroeconomic empirics and offer insights into some of the causal pathways of the relationships between health and growth and growth and health. These are discussed here.

How Have We Become so Healthy?

Mankind has become increasingly healthier over the past four centuries, and significantly so since 1900. Life expectancy at birth in the USA rose from 46 years and 48 years for men and women, respectively, in 1900, to 75 years and 81 years in 2010; a shift mirrored elsewhere in the industrialized countries, and observed in many developing countries as well. The major factors driving progress include improved nutrition, rising education, and advances in public health (1.7 Sahn). In contrast, health care investments have had far more modest effects on population health.

On the basis of his Nobel winning historical research on food production, malnutrition, and productivity Fogel (2004) estimates that improvements in the quantity and quality of food contributed to 40% of observed mortality declines since 1700, much of it during the twentieth century. Better nutrition raised agricultural output stemming famines and undernutrition. Parallel public health investments combined to steadily raise health status. Together labor productivity rose spurring economic growth.

Hygiene has also played a major role. In examining the explanation for mortality declines in England and Wales during the nineteenth and twentieth centuries historical research finds that immunizations, expanded access to piped water and sanitation, separation of water and sewerage, and better nutrition enhanced citizens' abilities to ward off infections, thereby reducing morbidity and mortality levels. Scientific knowledge of disease transmission contributed importantly to public health, and led to reductions in transmission of water and foodborne diseases that further improved health status (4.1 Cookson and Suhrcke, 4.3 Mills). Again, the evidence shows no measurable effect of medical interventions on survival.

Over the centuries both health and economic growth have improved. The question is the possible causal relationship between them to inform policy on raising both health status and incomes. One important conclusion is that health investments have valuable social and household benefits; whether these benefits translate into faster economic growth is simply a search for further benefits.

Correlates of Population Health and Income

The first empirical estimation of the relationship of income and health status was offered by Preston (1975) using life expectancy as a proxy for health status. More recent data confirm his finding of a concave relationship between gross domestic product (GDP) and life expectancy (**Figure 1**) and show the same relationship as his original results, with the relationship becoming stronger over time as poor countries catch up to the wealthier world in life expectancy.

 doi:10.1016/B978-0-12-375678-7.00602-7

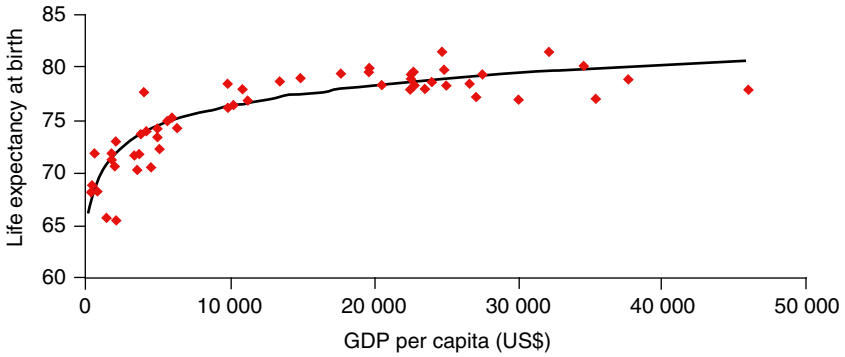


Figure 1 The Preston curve, 2001. Reproduced with permission from World Bank, World Development Indicators.

For the lowest income countries, small shifts in income lead to disproportionately higher life expectancies, often reflecting the first moves toward the demographic transition, as well as progress in public health. Preston attributed declines in mortality less to income and more to public health interventions and has suggested the combined importance of adequate calories, income, and education, echoing Fogel and others' historical conclusions.

Extending Preston's cross-sectional evidence with longitudinal data Deaton (2006) shows divergence among GDP, life expectancy, and infant mortality suggesting that the relationship between growth and health does not hold within countries over time. Each of the curves in Figure 2 is the standard deviation of the variable relative to its value in 1960. Thus, although country-level health indicators have converged per capita incomes have progressively diverged. If that is indeed the case then changes in income appear to be unrelated to health status measures.

Other researchers have argued for a strong and consistent relationship between growth in per capita incomes and reductions in infant and child mortality, and life expectancy based on cross-country, time-series data, and they point to poor economic performance as a cause of child mortality. They conclude that causality moves from income to health. Research examining the impact of income on emotional well-being in the US drawing on the Gallup-Healthways Well-Being Index shows that emotional well-being rises with log income but only to an annual income of US$75 000 after which there is no association. So income buys increasing well-being but at the lower end of the income scale.

Macroeconomic studies under the World Health Organization's (WHO) 2001 Commission on Macroeconomics and Health initiative produced evidence of the link between health and economic growth concluding that health investments will make countries richer. The evidence affirms correlations, but causality remains elusive due to the inability to adjust for reverse causality, and the problem of omitted variables. There are a number of variables that affect both health and growth such as climate and disease prevalence complicating efforts to measure the effects of one on the other.

Subsequent research on the relationship between health and changes in income explore the dynamics of health and income shifts. In a highly controversial study geography is used as an instrumental variable for health status to address the endogeneity problem. The study finds a high correlation between population health and economic growth. Subsequent research reviewed 13 studies of cross-country regressions and all repeat the same strong results. Considerable debate and challenge has ensued on the appropriateness of geography (distance from the equator) and its importance to growth. Subsequent empirical studies from multiple researchers show that once the effect of geography on a country's choice of institutions is controlled for, geography has little independent effect on growth.

Creative efforts were made to address the endogeneity problem including examination of twentieth-century breakthroughs in science such as drug therapies and insecticides and new global institutions such as WHO, but none find any independent effect on income growth though these new interventions did contribute to increases in population. Recent efforts to include microeconomic measures on the impact of health on productivity in a macroeconomic accounting framework find some modest gains in GDP. Incorporating general equilibrium effects to account for the diminished returns to labor with rapidly rising population have little effect as improvements in life expectancy do not lead to increases in per capita income or worker productivity.

Thus, despite creative approaches the controversy on whether health spurs growth remains a conundrum.

Individual Health and Productivity

Examining the same sets of relationships between income and health at the household level allows insights into how the factors interrelate to produce better health or higher incomes, and results in more robust findings. The downside of microeconomic studies is their limited generalizability, given the importance of different contexts in explaining effectiveness of interventions.

Morbidity and Income

Illness undermines productivity, in patterns similar to those Fogel identified between malnutrition and productivity. This cost-of-illness approach examines the short and longer term impacts of illness on education, labor productivity, employment, and economic activity.

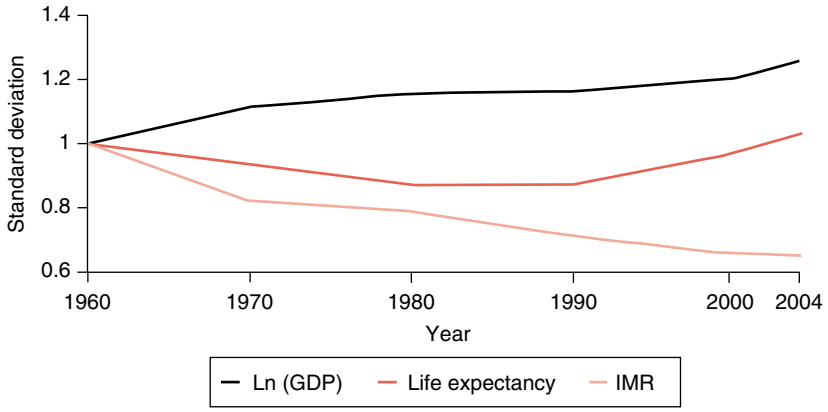


Figure 2 Normalized cross-country standard deviations of health and income, 1960–2004. The infant mortality rate (IMR) measures the number of children born who die before their first birthday per 1000 births. The standard deviation of the under-five mortality rate shows a similar evolution and is not presented in this figure. The IMR is a significant factor in life expectancy calculations, because, particularly in countries with high death rates, a significant portion of a country's deaths occur in the first year of life. Reproduced with permission as indicated in the NBER paper, Deaton, A. (2006). Global patterns of income and health: Facts, interpretations, and policies. *NBER Working Paper No. W12735.* Cambridge, MA: National Bureau of Economic Research.

A study of hookworm eradication in the American South tracking the impact of local infection control on education enrollment, attendance, and literacy suggests the importance of vector control in both education and incomes over lifetimes. The phasing in of hookworm control allowed an assessment of the patterns and extent of responses to eradication. Areas with higher pre existing infection rates saw bigger improvements after eradication. A parallel set of studies for malaria in the US c. 1930 and in Brazil, Colombia, and Mexico in the 1950s traced the effects of eliminating childhood exposure to malaria on adult literacy and earnings. The study concludes that "persistent childhood malaria infection reduces adult income by 40 to 60 percent." Similar but less dramatic effects emerged from malaria eradication across Indian states in the 1950s, where literacy and primary completion rates rose by 10% in malaria-free areas (1.12 Cohen).

More recent experiments suggest the potential value of targeted treatment measures on schooling and earnings. In a study of randomly assigned deworming treatment for school children in western Kenya a 25% increase in student attendance was found among those receiving medication, although there was no effect on school performance, possibly due to the lack of any change in other inputs at the school.

Expanded access to antiretroviral treatment in many countries of sub-Saharan Africa has offered a laboratory to test the impact of interventions. In western Kenya those under treatment are 20% more likely to join the labor market and increase their weekly hours worked by 35%. In another study in the tea-growing region of western Kenya human immunodeficiency virus (HIV)-positive tea pickers with 6 months on antiretroviral treatment increased their number of days worked and therefore their wages. Wage earnings rose from 75% to 89% of the wages of non-HIV-positive workers, thereby almost regaining the lost earning levels.

The impact of potential and actual parental death from HIV has shown some unexpected effects. In areas of high HIV prevalence across southern Africa children are less likely to go to school, take longer to go through school, and are less likely to graduate from primary school. More than half of the explanation is attributed to the expectations of a shorter life of parents, which will affect life chances of children rendering schooling of marginal benefit to future income. This complements evidence on the life chances of orphans emanating from an analysis of 10 Demographic and Health Surveys. Orphans in sub-Saharan Africa are significantly less likely to be in school as compared with their peers. Similar results emerge for Indonesia. However, country-level evidence is inconsistent in Africa. In Tanzania no impacts on schooling of HIV deaths of parents were found, perhaps because extended family members take on parenting roles for orphans (1.13 de Walque; 1.27 Thirumurthy).

A related and dramatic achievement in reducing the cost-of-illness is the multidonor program in the Niger Delta in West Africa where since the late 1990s pesticide spraying has controlled the black flies that had rendered the area uninhabitable. A total of 25 million hectares of rich agricultural land has been recultivated in reclaimed areas demonstrating yet again the value of public health interventions (1.16 Grépin).

All of these studies document augmented productivity and impacts on education or output from improvements in health status, or measure the costs of poor health on these same indicators. Interventions at the household level as well as targeted regional investments in public health activities can have important effects on education and productivity (3.9 Cutler and Lleres-Muney).

Limited evidence exists on the link between income and health. Analysis of a unique data set for South Africa permits quantification of the impact of old-age pensions in South Africa on health status. Where households pool income, including pensions, the overall health status of the household improves through positive effects on nutrition, living conditions, and stress levels of adults. Surveys typically fail to capture income pooling, and this study allows good insights into how pooled incomes can influence investments that enhance health status.

Investing in Early Childhood Development and Adult Performance

Targeted investments and the impact on economic well-being have received increasing attention with particular focus on children. Mounting multisectoral evidence from economics, psychology, and neuroscience makes the case that investments in disadvantaged young children have profound effects on learning, earnings, and adult health. Recent overviews point to factors such as maternal undernutrition, poverty, poor health, and unstimulating home environments as strongly associated with adult cancer incidence, mental illness, lower incomes, and lower birthweight of offspring. A review of microeconomic studies concludes that nutrition and possibly other dimensions of health compromise productivity. Heckman (2007) emphasizes the economic importance of noncognitive skills and their importance in future economic and social behavior, which in turn influence productivity and earnings. The US research suggests that the economic rate of return to preschool attendance among disadvantaged children dwarfs returns to other, later academic investments (3.1 Royer and Bauer; 4.13 Karoly).

Longitudinal studies provide strong evidence of the value of early childhood interventions. Low birthweight negatively affects long-run adult outcomes such as height, intelligence quotient (IQ), educational attainment, and earnings. Bolstering early nutrition interventions translate into greater cognitive development, physical stature and strength, greater learning, higher adult productivity, and healthier offspring. A 35-year longitudinal study in Guatemala traced adult participants who had received a randomly distributed nutrition supplements as school children. Women had 1.17 more years of schooling, their children were 179 g heavier at birth, and their children were 30% taller as adults compared with the children of those who had not received the supplement. Men in the treatment group earned an average wage of 46% above those who only received the calorie-based supplement (1.3 Soares; 1.7 Sahn).

Consistent, robust evidence on returns to early childhood investments confirms the importance of targeted interventions for disadvantaged children. Although often difficult because many young children remain at home, targeted early childhood programs have a clear payoff in higher productivity and earnings over a lifetime and better health in adulthood. They also offer the possibility of breaking the intergenerational cycle of poverty.

Health-Related Interventions and Health: Implications for Policy

If economic progress simply solved the problems of mortality and morbidity, then public policy would no longer be concerned with public health and health care needs would be met. Similarly, if investments in health provided the needed impetus for growth resource allocation decisions would be straightforward. However, circumstances are more complex and targeted policy decisions influence economic growth and population health.

Part of the interest in the growth to health link is to determine whether economic progress means more investments in health. A panel study of all countries in the 2010 WHO database examines the elasticity of public health spending with respect to national income controlling for demographics, source of financing, and other characteristics. A dynamic model using a lagged dependent variable adjusted for endogeneity shows results consistent with earlier studies in that GDP and public health expenditures move in tandem, but contrary to other studies finds significant elasticities below 1.0 for all but the lowest income countries, and lower values for the dynamic model results. So under this specification, income growth does not translate into commensurate increases in health spending.

If greater health investments could assure both better health status and rapid economic growth, then spending on health would need to be a priority given the large social and economic benefits. However, ample evidence suggests that spending alone falls short of even achieving health goals. The link from health spending to health outcomes is weak, and cross-country evidence shows minimal correlation between spending and health outcomes. As global access to important preventive and treatment technologies does not differ dramatically and funding is on the rise, it suggests that institutions and other factors such as education underpin the divergence between spending and outcomes. Chronic absenteeism, inadequate budget execution, illegal payments, and poor management combined with a severe lack of accountability translate into absence of the very technologies that can save lives at the point of service, such as drugs, supplies, and equipment, which in turn means low returns on investment. Poor governance in service delivery suggests government failure, effectively government interventions that have gone wrong. Without sound institutions, public health investments will not improve health, let alone economic growth (1.22 Government Regulation and Corruption – no author).

So despite advances in technology, countries experience marginal gains where countries lack institutions that can ensure effective delivery and financing of health care services. However, the discrepancy extends to the Organization for Economic Cooperation and Development (OECD) countries, where studies of survival rates for specific diseases do not correlate with either total or public spending.

Debate on the path to better health and economic growth will continue given the inconclusive nature of the evidence. Better research and measurement will help to hone the findings and provide stronger policy guidance; but stronger institutions to ensure effective health care delivery will require equal attention.

See also: Education and Health. Health Status in the Developing World, Determinants of. HIV/AIDS: Transmission, Treatment, and Prevention, Economics of. Intergenerational Effects on Health – *In Utero* and Early Life. Nutrition, Health, and Economic Performance. Preschool Education Programs. Public Health in Resource Poor Settings. Public Health: Overview

References

Deaton, A. (2006). Global patterns of income and health: Facts, interpretations, and policies. *NBER Working Paper No. W12735*. Cambridge, MA: National Bureau of Economic Research.

Fogel, R. (2004). *The escape from hunger and premature death, 1700–2100: Europe, America and the Third World*. Cambridge, UK: Cambridge University Press.
Heckman, J. (2007). Investing in disadvantaged young children is good economics and good public policy. Testimony before the Joint Economic Committee, Washington, DC, June 27.
Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies* **29**(2), 231–248.

Further Reading

Bleakley, H. (2007). Disease and development: Evidence from hookworm eradication in the American South. *Quarterly Journal of Economics* **122**(1), 73–117.
Bloom, D., Canning, D. and Sevilla, J. (2004). The effect of health on economic growth: A production function approach. *World Development* **XXXII**, 1–13.
Commission on Macroeconomics and Health (2001). *Macroeconomics and health: Investing in health for economic development*. Geneva: World Health Organization.
Cutler, D. and Miller, G. (2005). The role of public health improvements in health advances: The twentieth century United States. *Demography* **42**(1), 1–22.
Grantham-McGregor, M., et al. (2007). Development potential in the first 5 years for children in developing countries. *Lancet* **369**, 60–70.
Jack, W. and Lewis, M. (2009). Health investments and economic growth: Macroeconomic evidence and microeconomic foundations. In Spence, M. and Lewis, M. (eds.) *Health and growth*, pp. 1–39. Washington, DC: Growth Commission.
Lopez-Casasnovas, G., Rivera, G. B. and Currais, L. (eds.) (2005). *Health and economic growth: Findings and policy implications*. Cambridge, MA: MIT Press.
Spence, M. and Lewis, M. (eds.) (2009). *Health and growth*. Washington, DC: Growth Commission.

Willingness to Pay for Health

R Baker, C Donaldson, and H Mason, Glasgow Caledonian University, Glasgow, UK
M Jones-Lee, Newcastle University Business School, Newcastle upon Tyne, UK

Introduction

The 'willingness to pay' (WTP) method was first applied in the health area in the famous study of WTP to avoid heart attacks, by Acton (1973). WTP for health is an issue in individual (personal) and societal (public) decision making about health care. The term usually refers to individuals' willingness to spend money personally, i.e., 'out of pocket,' to obtain health gains for themselves or to avoid health losses or reduce health risks for themselves. WTP in this sense can be observed both in actual behavior and in responses to hypothetical questions. In either case, WTP is interpreted as an indicator of how much personal satisfaction or well-being (often called 'utility') individuals derive from (or believe they derive from) different health outcomes. It is thought that aggregate WTP out of pocket in a group of people benefiting from a program may be used as an indicator of how much money society should be willing to spend on the program, although this depends on the funding arrangements (whether health care is publicly or privately financed) and who is asked (patients or the public).

The satisfaction and well-being individuals derive from health outcomes may be measured in different ways. Most of these, including quality-adjusted life-years (QALYs) and disability adjusted life years (DALYs), aim at measuring the value of different health outcomes relative to each other with a view to aiding priority setting within given health budgets. The special property of measurements in terms of WTP is that they purport to establish the value of health outcomes relative not only to each other but also to other goods and services. In principle, WTP data may thus inform decisions not only about priority setting within given health budgets but also as to how much money should be allocated to health services versus other goods and services.

In the following, the authors' focus is first on the meaning and measurement of individual WTP and the aggregation of this across individuals. Actual WTP for different technologies and programs in a public health service can be different from the simple individual aggregate, most notably because of various concerns for equal access and equity. The authors return to this important point in the Issues section at the end.

Welfare Theoretical Background and Application (in Principle) to Health Technology Assessment

What is 'WTP'?

In standard welfare economics, maximum WTP represents the theoretically correct measure of 'strength of preference' for, or value of, a commodity. In areas of public-sector activity, such as health care, in which conventional markets do not exist, decisions still have to be made about how best to use limited resources. This requires valuation of both resource costs of interventions and their benefits (the benefits being health gain and other sources of well-being). This can be elicited either in surveys by use of hypothetical WTP questions – essentially, the contingent valuation approach in which respondents state values for the good in question – or revealed as a result of observing real market-based choices. Owing to the rarity of the latter in health care, the contingent valuation method tends to be relied on more than revealed preference.

WTP focuses on the valuation of benefits, whereby a health care option may be described to a respondent and the person is asked what is his/her maximum WTP for it. In principle, with this type of information, the combination of interventions in cash-limited publicly funded systems could be chosen, which maximizes the value of benefits (possibly distributionally weighted) to the community. In privately funded systems, WTP information would more likely be used to assess whether aggregate WTP for a particular health care intervention exceeds costs; if it does, the implication is that this intervention can be added to the benefits package. In each of these cases, WTP is necessarily treated as a cardinal measure of value, i.e., a measure which captures strength as well as direction of preference.

It is important to distinguish WTP, as a measure of benefit, from the price of a good. In many cases, people would be willing to pay more than the market-clearing price of a good. For any individual, the difference between benefit, as represented by his/her maximum WTP for the good, and the price paid by him/her for the good represents a gain in well-being from having the good provided at the market price.

WTP and Levels of Decision Making

WTP methods have been used at three main levels of decision making in health care. The first level is that of patients within a clinical area being asked to choose between close substitutes of the sort that might be evaluated within a randomized trial. It could be argued that when making such choices between therapies, the people best placed to judge are patients who will receive the therapies. The second level is that where geographically defined health authorities have to prioritize across such clinical areas in a publicly funded system (e.g., as in one study, by Olsen and Donaldson (1998), comparing helicopters, hearts, and hips). Here, it could be argued that those best placed to decide are taxpayers (or those who paid taxes in the past). In more insurance-based systems, where the decision might be about whether to add a program to the benefits package, those same people might be asked about their WTP for that program alone, the decision being based on whether the aggregate WTP is greater than the costs of said program. The different scenarios that emanate from these contextual backgrounds, and the ways that WTP can be used within them, have been laid out by O'Brien and Gafni (1996) and Shackley and Donaldson (2000).

Internationally, the advent of Health Technology Assessment (HTA) agencies has now taken WTP methodology to the national level. Many of the recommendations made by HTA organizations involve considerations of the costs and benefits of single interventions, where there is no obvious basis for a comparison between alternatives, apart from with the costs and effects of treatments that constitute current practice or the status quo. In such cases, the decision about whether to provide the intervention is not obvious and hence gives rise to the question of what monetary value to place on a gain of one QALY.

All of the above scenarios also give rise to the question of whether WTP can be used to value wider benefits of health care (e.g., location of care and process utility) over which people have preferences but which might not be captured easily by narrower, health-focused methods, such as QALYs.

Measurement of Individual WTP and Estimation of Aggregate Value

In 1963, the first empirical application of WTP, published in a journal, was in the area of environmental policy evaluation. During the 1970s, the method was further developed in studies of the valuation of saving lives, as applied to safety and transport policies, largely through the work of Jones-Lee (1974) and Jones-Lee *et al.* (1985). This latter body of work is of great relevance to the issues faced currently in health as the development of methods to elicit a value of saving a life has strong parallels with the challenge of valuing a QALY (or 'healthy year'). Therefore, in this section, the authors commence with an introduction to measurement methods by focusing on lessons learned in area of safety (especially the issue of how to estimate the value of a prevented fatality, or VPF), before moving to how these lessons have been built on in deriving estimates of the value of a QALY.

Revealed Preference versus Contingent Valuation in Safety

Basically, the revealed preference approach involves the identification of situations in which people actually do trade-off income or wealth against physical risk – for example, in labor markets where riskier jobs can be expected to command clearly identifiable wage *premia*. By contrast, the contingent valuation approach applied to safety typically involves asking a representative sample of people more or less directly about their individual WTP for improved safety (or, sometimes, their willingness to accept compensation for increased risk).

The difficulty with the revealed preference approach when applied to labor market data is that it depends on being able to disentangle risk-related wage differentials from the many other factors that enter into the determination of wage rates. The approach also presupposes that workers are well informed about the risks that they actually face in the workplace. In addition, those whose jobs do carry clearly identifiable wage *premia* for risk may not be representative of the work force as a whole, in that such people almost certainly have a below-average degree of risk-aversion.

The great advantage of the contingent valuation approach is that it allows the researcher to go directly and unambiguously to the relevant wealth/risk trade-off – at least, in principle. However, the contingent valuation approach has the disadvantage of relying on the assumption that people are able to give considered, accurate, and unbiased answers to hypothetical questions about typically small changes in already very small risks.

Aggregation

So, under what has naturally come to be known as the 'WTP' approach to the valuation of safety, one first seeks to establish the maximum amounts that those affected would individually be willing to pay for (typically small) improvements in their own and others' safety. These amounts are then simply aggregated across all individuals to arrive at an overall value for the safety improvement concerned (see example in following paragraph). The resultant figure is thus a clear reflection of what the safety improvement is 'worth' to the affected group, relative to the alternative ways in which each individual might have spent his or her limited income. Furthermore, defining values of safety in this way effectively 'mimics' the operation of market forces – in circumstances in which markets typically do not exist – insofar as such forces can be seen as vehicles for allowing individual preferences to interact with relative scarcities and production possibilities to determine the allocation of a society's scarce resources.

To standardize values of safety that are derived from the WTP approach and render them comparable with values obtained under other approaches (such as gross output – see Jones-Lee (1994)), the concept of the prevention of a 'statistical' fatality or injury is applied. To illustrate this concept, suppose that a group of 100 000 people enjoy a safety improvement that reduces the probability of premature death during a forthcoming period by, on average, 1 in 100 000 for each and every member of the group. The expected number of fatalities within the group during the forthcoming period will thus be reduced by precisely one, and the safety improvement is therefore described as involving the prevention of one statistical fatality. Now suppose that individuals within this group are, on average, each willing to pay £w for the 1 in 100 000 reduction in the probability of death afforded by the safety improvement. Aggregate WTP will then be given by £$w \times 100\,000$. This figure is naturally referred to as the WTP-based value of preventing one statistical fatality (VPF) or alternatively as the value of statistical life (VOSL). So, for instance, if w=£10, VOSL=£1 000 000 in our example.

Clearly, in the above example, average individual WTP, £w, for the average individual risk reduction of 1 in 100 000 is a reflection of the rate at which people in the group are willing to trade-off wealth against risk 'at the margin,' in the sense that the trade-offs typically involve small variations in wealth and small variations in risk. Empirical work on the valuation of safety thus tends to focus on these individual marginal wealth/risk trade-off rates.

On a somewhat more cautionary note, it is extremely important to appreciate that, defined in this way, the VPF is not a 'value (or price) of life' in the sense of a sum that any given individual would accept in compensation for the certainty of his or her own death – for most of us, no finite sum

would suffice for this purpose so that in this sense life is literally priceless. Rather, the VPF is an aggregate WTP for typically very small reductions in individual risk of death (which, realistically, is what most safety improvements really offer at the individual level).

In valuing safety it is also useful to have a valuation of an averted statistical injury (VSI). The VSI is generally pegged against the VOSL rather than estimating WTP for an injury. To do this, a standard gamble exercise can be used to calculate the marginal rate of substitution (MRS) of an injury against death. The value of preventing this injury is then the fraction of the VOSL that is given by the MRS, i.e., if the MRS is 0.09, the value of preventing this injury will be 9% of the VOSL.

Recent Developments in Valuing Life

Viewed from an historical perspective, both quantitative and qualitative research have tended to cast doubt on the reliability and validity of WTP values for safety derived through the direct contingent valuation method. As well as sequencing and framing effects, a prominent issue has been the lack of ability of the method to account for embedding and scope. That is, respondents tend to view safety improvements as a 'good thing' and, therefore, will often state much the same WTP for different sizes of risk reduction, whether for fatal or nonfatal injuries. It may be unreasonable to expect respondents to give accurate answers to hypothetical questions, which involve direct trade-offs between wealth and small reductions in risk.

In view of these difficulties with the direct contingent valuation approach, Carthy *et al.* (1999) suggested a less-direct, 'chained' approach, which breaks down the valuation process into a series of more manageable steps. More specifically, respondents are first presented with a question asking them about their WTP for the certainty of a complete cure for a given nonfatal road injury and their willingness to accept compensation (WTA) for the certainty of remaining in the impaired health state. On the basis of reasonable assumptions concerning underlying preferences, it is then possible to obtain an estimate of each respondent's marginal rate of substitution of wealth for the risk of the nonfatal injury as a weighted average of his/her WTP and WTA responses. Respondents are then presented with a 'standard gamble' (SG) question aimed at determining the ratio of individual marginal rates of substitution of wealth for risk of death relative to the corresponding marginal rate of substitution for risk of the nonfatal injury. The monetary value from the first stage can then be combined with the ratio from the second stage to obtain a WTP for reduced risk of death.

The approach is, perhaps, more realistic in that most people can relate to giving a monetary value for avoiding a nonfatal injury of the sort they are likely to have experienced, and people are not asked directly to place a monetary value on a small risk reduction. The method has shown promise in terms of being subject to less marked embedding effects and other biases than earlier approaches.

Linking Safety and Health: Modeling the Value of a QALY

As noted, the most common measure of benefits applied in health economic evaluation is the QALY. Although the cost per QALY gained for technology 'A' can be compared with the cost per QALY gained of technology 'B' (and hence the *relative* efficiency of A vs. B can be assessed), weighing up costs (measured in monetary terms) and health benefits (measured in QALYs) for a single technology is problematic. As a result, health care decision makers have tended to invoke decision rules (such as cost per QALY limits above which technologies are unlikely to be recommended for adoption). To generate research evidence to underpin such decision rules, there has been international interest in the estimation of the monetary value of a QALY. If such a value could be estimated, then costs and benefits could be compared using the same metric (i.e., money), and one could more readily judge whether benefits offset costs. Existing evidence from the economics of safety literature is a natural starting point for such a project.

A straightforward way to combine work on health and safety valuation is to take the well-established VOSL based on research into road safety and, from it, attempt to model the monetary value of an year of life. This can reasonably be interpreted as the value of a healthy year, which is the same as a QALY. Indeed, this is what was attempted in the UK Social Value of a QALY (SVQ) project and more recently in the European Value of a QALY project (EuroVaQ). A simplified version of the method of transforming the VPF into a value of a QALY is presented in **Box 1**. However, the value resulting from this would reflect a particular QALY type. By QALY type, the authors mean that QALYs can be generated in at least two ways, these being by adding years to life expectancy or by

Box 1 Modeling the value of a QALY

A straightforward way to compute the value of a QALY is to start with the well-established roads VPF for the UK. For example, if a representative death is taken avoided as being that of a person aged 35 years, assume that the VPF is £1.4 m (or $£1.4 \times 10^6$) and that the person concerned would have lived for another 40 years, a rough calculation of the value of a life year gained by that person would be as follows:

$$V = \frac{£1.4 \times 10^6}{40}$$
$$= £35\,000$$

However, if one were to assume that not all of the 40 years gained would be spent in full health (especially later years) and a discount rate applied, the denominator would fall, thus raising the value of a QALY above £35 000. For example, if the discount rate was taken to be 3.5%, then the annualized sum that would have a discounted present value of £1.4 m over 40 years would be £77 300.

In SVQ, similar approaches were used to model the value of QALYs resulting from extending life and from quality-of-life enhancement only. For example, the latter is based on UK values for four different scenarios of serious injury. Each health state was broken down into three phases; in hospital effect (valued at 0.69 or 0.16 on the EQ-5D tariff, depending on severity of injury and generally modeled as lasting for 1 month), initial after effects (generally for 2 months and valued at 0.76) and longer term effects (for remaining life and valued at either 0.76 or 0.3, again depending on severity). Assuming that any given injury would occur at the mean age of the UK population, with 26 expected remaining QALYS, an overall total QALY loss for each scenario has been calculated. The VSI of £150 000 then has been divided by the total QALY loss for each scenario and computed a weighted average based on probability of each scenario occurring.

enhancing the quality of remaining life years without extending life. The former can be further subdivided into avoiding immediate threats to life – hereafter called 'life saving,' or adding years to the end of one's life – hereafter called 'life extension.' The procedure outlined in Box 1 reflects the first of these, although a more-sophisticated approach, still using the VOSL as a basis, was also employed to model the value of a QALY arising from life extension as opposed to life saving.

Valuing Improvements in Quality of Life

WTP procedures have been used directly by the UK Department for Transport to derive the value of preventing a serious injury (VSI). From such data, the monetary value of health gains (as measured by instruments in the QALY field) can be estimated. But one can also combine a monetary value for a health gain with health state utility scores to achieve values for different improvements in quality of life. The simplest way to do this is to use established utility tariffs in instruments such as the EQ-5D, the Health Utilities Index, the 15-D, etc. For instance, if a survey respondent was willing to pay a maximum of £3000 WTP for a move from a more-impaired health state for 1 year to a less-impaired state for the same period, where the tariff difference between the states is 0.1, then the WTP for a QALY implied by this would be £30 000 (i.e., £3000/0.1) for that individual. If repeated over several respondents and scenarios, a representative WTP for a QALY could be derived (Gyrd-Hansen, 2003).

Others have combined WTP questions and utility scores in the same survey, for example, in the social value of a QALY studies in the UK and Spain, the feasibility of doing this was assessed by presenting members of the public with appropriately framed valuation questions in a survey – see Box 2 for example health states, and Box 3 for example question to illustrate how changes in quality of life and WTP were estimated and combined (each from the UK study).

From Box 3, it can be seen that any individual respondent would be faced with a set of WTP and standard gamble questions, the two sets then being combined in different ways to arrive at values of a QALY. The reason for having respondents undertake a health state utility assessment exercise, rather than combine WTP values with a preexisting tariff (such as that which exists for the EQ-5D quality-of-life system), was that the researchers wanted each respondent's WTP value to be combined with their own personal health state utility value for purposes of internal consistency.

Box 2 Stomach and head health states

Stomach: 3 months

Initially you will have severe stomach pains, diarrhea, vomiting, and fever for 7 days, severe enough to interfere with most of your usual activities.

Things then improve, but for up to 1 year from initial onset you will suffer an episode of stomach discomfort and sickness every couple of weeks, with each episode lasting for 2–3 days. These episodes are not so severe but may interfere with some of your usual activities.

(*Half of the respondents were given stomach health state descriptions of 3 months, 12 months, and lifetime durations.*)

Head: 3 months

One will have episodes of throbbing pain across the front of your head and will feel sick and may occasionally be sick. One will feel like he/she wants to lie still in a darkened room.

During the next 3 months one will suffer an episode of head pain and sickness every couple of weeks, with each episode lasting between 8 h and 2 days. These episodes will interfere with many of your usual activities. After 3 months he/she returns to the current health with no further effects from this illness.

(*The other half of the respondents were given head health state descriptions of 3 months, 12 months, and lifetime durations.*)

Examples of Applications in Safety and Health

Values of Life

Turning to the question of the figures that are actually applied in practice, WTP-based values of safety are currently used in road project appraisal in the UK, USA, Canada, Sweden, and New Zealand, with several other countries employing values that have been substantially influenced by the results of WTP studies. More specifically, in the UK, the Department for Transport currently employs a figure of £1.64 million in June 2007 prices for the prevention of a statistical fatality in its roads project appraisal. In turn, the Department's values for the prevention of serious and slight, nonfatal injuries are £185 220 and £14 280, respectively, again in 2007 prices.

In the USA, the US Department of Transportation currently values the prevention of a statistical road fatality at US$6 m (approximately £3.96 m). In turn, Transport Canada applies a WTP-based value for the prevention of a statistical fatality of Cdn$ 4 m (approximately £3.22 m) based on a survey of the literature. The WTP values used in Sweden and New Zealand were derived under the contingent valuation approach and in 2010 prices are SEK 18.5 million (approximately £1.8 m) and NZ$ 3.56 million (approximately £1.7 m).

WTP for a QALY: Estimates from Modeling Studies

Table 1 gives a typical set of values of a QALY that have arisen from the UK-based modeling study described above (see Box 1). It would seem that different 'QALY types' would imply different values. Based on WTP for life saving (to reduce the risks of life-threatening events), values close to £70 000 per QALY were produced, as compared to values approximately £35 000 for a life-extending QALY (at the end of life). Estimating gains from improvements in quality of life, with no increase in number of remaining years, produced a lower value of approximately £10 000 per QALY. These differences are striking, and further studies are needed to clarify them.

Box 3 Valuing a QALY via surveying the general public

The value of a QALY is derived via a 'chaining' procedure. In the initial part of the chain, the respondent is asked about whether she/he would be prepared to pay anything to avoid being in this state, and, if so, what is the maximum amount she/he is willing to pay.

In the second part of the chain, the respondent would be asked a 'standard gamble' question involving a choice between two options. In the standard way of deriving a QALY index, one option would leave the respondent in the stomach/head condition for certain for the remainder of his/her life, whereas the other option would involve a gamble with varying probabilities of a better or worse outcome. 'Better' usually means a return to full health for the rest of one's life, whereas worse is usually characterized as immediate death. Visual procedures are used to guide the respondent through the process, and the index is derived from the point at which the respondent feels it is difficult to choose between the outcome for certain and the gamble.

Let us assume that, for one respondent, the probability at which she/he finds it difficult to choose between the head condition for certain and taking the gamble is 0.95, and that his/her WTP to avoid a year in the stomach condition was £1000. Dividing £1000 by 0.05 (which comes from subtracting 0.95 from 1) would give a value of a QALY for that person of £20 000. This can be done across several individuals to arrive at an average value of a QALY for a population.

For either head or stomach conditions, each respondent was asked two WTP questions (to avoid the 3-month state and the 12-month state) and three standard gamble questions (3 months for certain vs a gamble with outcomes of return to current health or 12 months in the state; 12 months for certain vs a gamble with outcome of return to current health or rest of life in the state; and rest if life for certain vs gamble with outcomes of current health or immediate death). In fact, slightly more WTP and standard gamble questions were asked of each respondent, but these are not relevant to this paper.

Table 1 Values of a QALY via alternative calculations from modeling based on VPF and VSI

Basic modeling approach	*Value of a QALY (£)*
Life saving	70 000
Life extending	35 000
Quality-of-life enhancing	10 000

Source: Reproduced from Donaldson, C., Baker, R., Mason, H., et al. (2011). The social value of a QALY: Raising the bar or barring the raise? *BMC Health Services Research* **11**, 8. doi:10.1186/1472-6963-11-8.

The issue of WTP for QALY types has been explored recently in subsequent surveys on the 'European value of a QALY' (see the EuroVaQ website at http://research.ncl.ac.uk/eurovaq/). It is also worth noting that the VPF itself is just over nine times the value of preventing a statistical injury. That there is no single value of a QALY is in line with other published views, the lowest value also being reflective of earlier published studies, which looked at the value of QALY gains arising from quality-of-life enhancement only.

A small number of other studies have modeled the value of a QALY from a Value of a Statistical Life. Using Australian data Abelson (2003) reported a value of AUS$108 000 (approximately £59 000). In Sweden, a value of US$90 000 (approximately £64 000) was reported by Johannesson and Meltzer (1998), whereas Hirth *et al.* (2000) calculated a value of US$161 305 (approximately £115 000).

Values of a QALY from Survey Research

Because the surveys referred to above from the UK and Spain (see Baker *et al.* and Pinto Prades *et al.* in further reading list) were intended only as feasibility studies, they were based on small samples not necessarily representative of the population as a whole. Nevertheless, the work suggests that it is feasible to conduct a survey to elicit monetary values for a QALY from a representative sample of the public so long as the procedure is broken down into manageable steps. However, it also became apparent that the mean estimates produced by such questions are particularly prone to the influence of 'outlier responses,' and that great care is therefore required in the selection of central-tendency measures. The most common example of an outlier was that many people were willing to take only very small risks of a more adverse outcome to avoid the stomach and head health states in the standard gamble questions or were even not willing to gamble at all. As well as such 'floor' effects respondents may also have a WTP ceiling (or budget constraint), an amount they express whether for a small or large perceived gain. Thus, when WTP values and health state utilities are combined in such circumstances, the implied WTP per QALY for such individuals can be so high as to lead to an implausible population average WTP per QALY across the whole sample. This was indeed the case in these studies, with the value running into several millions of pounds/euros! Other ways of managing the data can be used, however. For example, if mean WTP and mean QALY loss are combined (a 'ratio of means' as opposed to 'mean of ratios' approach), much more conservative values of a QALY emerge.

Issues

WTP and Ability to Pay

On the face of it, it would seem that it is problematic to use WTP measures to inform decisions about the allocation of resources for commodities, such as health care, for which such allocation is supposed to be on the basis of (some notion of) need. This is because WTP is obviously associated with ability to pay. However, Government departments in the UK and in other countries employ WTP-based values for the prevention of a statistical premature fatality (VPF) that are based on central-tendency measures (typically arithmetic means) of the population distribution of individual WTP for risk reduction. The fact that these VPFs are then applied uniformly to all groups in society whatever the income levels of members of the group clearly entails the implicit use of inverse-income distributional weights and further overcomes the challenge of ability to pay influencing WTP.

It would also appear that there is already a precedent in public-sector decision making for (at least implicitly) applying distributional weights to individual WTP for reductions in risks to life in order to arrive at an overall value in the form of a population mean (or possibly median) of individual values. This represents the state of the art of dealing with any such biases.

Validity and Scope

A major focus in the environmental WTP literature with respect to validity has been on scope effects: especially whether respondents are willing to pay more for greater amounts of the good being valued than for smaller ones, as one would expect. Carson (1997) showed that most studies (31 of 35 reviewed) reveal sensitivity to scope. Despite such positive results, doubts still remain about WTP values elicited through hypothetical surveys, and scope tests have taken on the status of being the 'acid test' for any particular study. As in other areas of application, results in health have been mixed (Smith, 2001; Olsen *et al.*, 2004). Also, this has led to more qualitative methods being used to examine the thought processes underlying respondents' stated values, a trend which will likely (and justifiably) continue.

Valuing Others

So far only passing reference has been made to people's concern – and hence WTP – for others', as well as their own health and safety. Insofar as people do display such 'altruistic' concern, then one would naturally expect that it would be appropriate to augment the WTP-based VPF to reflect the amounts that people would be willing to pay for an improvement in others' safety. However, it turns out that under plausible assumptions about the nature of people's altruistic concern for others' safety on the one hand and their material well-being on the other (the latter being reflected by their wealth or consumption), augmenting the VPF to reflect WTP for others' safety would involve a form of double counting and would therefore ultimately be unjustified. For example, suppose that individual A is concerned not only about individual B's safety but also about the latter's wealth or consumption. Furthermore, suppose that individual A's altruistic concern for B is 'pure,' in the sense that it respects B's preferences. Although A will then regard a reduction in B's risk of premature death as a 'good thing,' he/she will also regard the increase in B's taxation (or other expenditure) required to finance the risk reduction as an exactly offsetting 'bad thing.' Taking account only of A's WTP for B's safety improvement would therefore quite literally involve double counting. Thus, the issue of whether and how people's concern for others' safety ought to be taken into account under the WTP approach hinges on the essentially empirical question of the relationship between such concern and concern for others' wealth or consumption. Detailed discussions of the issue of altruism and safety exist in the literature, and it would seem that similar arguments apply to societal WTP values elicited in the health field.

Actual WTP in a National Health Service with Concerns for Equity

As noted above in the Section on Valuing Improvements in Quality of Life, WTP for improvements in quality of life can be estimated by combining WTP with individual utility scores for health states (measured on the 0–1 scale used in QALY calculations). For instance, if a national health service's WTP for a gained healthy year is £30 000, the WTP for improving a person's health (utility) by 0.1 for 1 year may be estimated to be $0.1 \times 30\,000 = £3000$. But this presupposes that individual utilities for health states correctly reflect societal preferences for priority setting – and hence correctly indicate societal WTP. There is much evidence that this assumption does not hold in societal decisions about resource allocation among groups of patients with different degrees of severity of disease and among groups with equal interests in treatment but different capacities to benefit from treatment. Most notably, in several jurisdictions the general public and societal decision makers have been shown to value utility gains more the more severe the initial condition is (Nord, 1999; Shah, 2009). To use conventional utilities from the QALY field to estimate what would be reasonable for a national health service to be willing to pay for technologies and programs that improve quality of life is thus problematic.

To capture concerns for equity, Nord (1999) suggested a type of priority weights for life years with a structure that is different from that of utilities normally used in QALY calculations. In principle, such weights can be used instead of conventional utilities when combining with a public health service's WTP for gaining a healthy year to estimate WTP for improvements in quality of life. This is an area of continuing research. There is at present no universal agreement as to what would be the right alternative weights to use for this purpose.

A similar point about equity can be made regarding WTP for gained life years. Even if a national health service indicates a WTP of, for instance, £30 000 for one gained healthy year (one QALY), it does not follow that N gained years should give a WTP of $N \times 30\,000$. A national health service may for instance consider that this would lead to unjustified age discrimination in priority setting in surgery and other one-time treatments of life-threatening conditions. It does not follow that a gained life year in a state of reduced health, say at utility level 0.8, should give a WTP of only $0.8 \times 30\,000 = £24\,000$, as this by many would be regarded as unfair discrimination against people with chronic disease or disability.

Conclusions

Measurements of individual WTP for health benefits are potentially useful to examine whether the individual welfare gains from health programs outweigh their opportunity cost, i.e., outweigh the welfare gains that could have been obtained by spending the resources on other goods and services instead. In this, the WTP approach has an advantage over approaches that use QALYs or DALYs. As with other approaches to valuation of health care, there are various methodological challenges associated both with measurement and aggregation of WTP. There is also an important distinction between estimating aggregate individual WTP and determining what should be the WTP in a national health service that strives for both efficiency and equity in health care. Further research on WTP measurements will help to clarify the potential and the role of the WTP approach in health care resource allocation.

See also: Cost–Value Analysis. Multiattribute Utility Instruments and Their Use. Quality-Adjusted Life-Years

References

Abelson, P. (2003). The value of life and health for public policy. *The Economic Record* **79**, S2–S13.

Acton, J. P. (1973) *Evaluating Public Programs to Save Lives: The Case of Heart Attacks*. Report No. R950RC, Santa Monica: RAND Corporation.

Carson, R. T. (1997). Contingent valuation surveys and tests of insensitivity to scope. In Kopp, R. J., Pemmerhene, W. and Schwartz, N. (eds.) *Determining the value of non-marketed goods: Economic, psychological and policy relevant aspects of contingent valuation methods*. Boston: Kluwer.

Carthy, T., Chilton, S. M., Covey, J., et al. (1999). On the contingent valuation of safety and the safety of contingent valuation: Part 2 – the CV/SG 'chained' approach. *Journal of Risk and Uncertainty* **17**, 187–213.

Gyrd-Hansen, D. (2003). Willingness to pay for a QALY. *Health Economics* **12**, 1049–1060.

Hirth, R. A., Chernew, M. E., Miller, E., Fendrick, M. and Weissert, W. G. (2000). Willingness to pay for a quality-adjusted life year: in search of a standard. *Medical Decision Making* **20**, 332–342.

Jones-Lee, M. W. (1974). The value of changes in the probability of death or injury. *Journal of Political Economy* **82**, 835–849.

Jones-Lee, M. W. (1994). Safety and the saving of life: The economics of safety and physical risk. In Layard, R. and Glaister, S. (eds.) *Cost benefit analysis*. Cambridge: Cambridge University Press.

Jones-Lee, M. W., Hammerton, M. and Philips, P. R. (1985). The value of safety: Results of a national sample survey. *Economic Journal* **95**, 49–72.

Nord, E. (1999). *Cost-value analysis in health care*. Cambridge: Cambridge University Press.

O'Brien, B. J. and Gafni, A. (1996). When do the 'dollars' make sense. Towards a conceptual framework for contingent valuation studies in health care. *Medical Decision Making* **16**, 288–299.

Olsen, J. A. and Donaldson, C. (1998). Helicopters, hearts and hips: Using willingness to pay to set priorities for public sector health care programmes. *Social Science and Medicine* **46**, 1–12.

Olsen, J. A., Donaldson, C. and Periera, J. (2004). The insensitivity of 'willingness to pay' to the size of the good: New evidence for health care. *Journal of Economic Psychology* **25**, 445–460.

Shackley, P. and Donaldson, C. (2000). Willingness to pay for publicly financed health care: How should we use the numbers? *Applied Economics* **32**, 2015–2021.

Shah, K. K. (2009). Severity of illness and priority setting in healthcare: A review of the literature. *Health Policy* **93**, 77–84.

Smith, R. (2001). The relative sensitivity of willingness-to-pay and time trade-off to changes in health status: An empirical investigation. *Health Economics* **10**, 487–497.

Further Reading

Baker, R., Bateman, I., Donaldson, C., et al. (2010). Weighting and valuing quality-adjusted life-years using stated preference methods: Preliminary results from the social value of a QALY project. *Health Technology Assessment* **14**, 27.

Donaldson, C. (1999). Valuing the benefits of publicly-provided health care: Does 'ability to pay' preclude the use of 'willingness to pay'? *Social Science and Medicine* **49**, 551–563.

Donaldson, C., Birch, S. and Gafni, A. (2002). The pervasiveness of the 'distribution problem' in economic evaluation in health care. *Health Economics* **11**, 55–70.

Johannesson, M. (1995). The relationship between cost effectiveness analysis and cost benefit analysis. *Social Science and Medicine* **41**, 483–489.

Mason, H., Jones-Lee, M. W. and Donaldson, C. (2009). Modelling the monetary value of a QALY: A new approach based on UK data. *Health Economics* **18**, 933–950.

Pinto Prades, J. L., Loomes, G. and Brey, R. (2009). Trying to estimate a monetary value of a QALY. *Journal of Health Economics* **28**, 553–562.

Viscusi, W. K. (1978). Labor market valuations of life and limb: Empirical evidence and policy implications. *Public Policy* **26**, 359–386.

Index

This index is in letter-by-letter order, whereby hyphens and spaces within index headings are ignored in the alphabetization, and it is arranged in set-out style, with a maximum of four levels of heading. Location references refer to the volume number, in bold, followed by the page number. Page numbers suffixed by *F* and *T* refers to figures and tables respectively. vs. indicates a comparison.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

Q

R

S

T

U

V

W

Y

Z